INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Effective Hybrid Outlier Selection Method for Breast Cancer Classification Using Machine Learning Algorithms

P B Tintu[1]*, S Veni[2], S Manju Priya[2]

1 Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
2 Professor, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India

OPEN ACCESS

Received: 27-06-2024

Accepted: 03-11-2024

Published: 26-11-2024

Citation: Tintu PB, Veni S, Priya SM (2024) An Effective Hybrid Outlier Selection Method for Breast Cancer Classification Using Machine Learning Algorithms. Indian Journal of Science and Technology 17(43): 4494-4501. https://doi.org/10.17485/IJST/v17i43.2109

* **Corresponding author**.

tintupadikkal@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop a model for the prediction of Breast cancer. Cancer is one of the deadliest diseases and it is regarded as the second leading cause of death in women throughout the sphere. Former detection of cancer can save the patient's life. Outliers can have an impact on the model's performance. For this reason, eliminating outliers is the first factor to be considered. **Methods:** In this study, the Wisconsin Diagnostic Breast Cancer dataset was used. It consists of 569 instances of which 357 instances are benign and 212 are malignant cases. It has 32 attributes including two class attribute labels (diagnosis: B= benign, M= malignant), ID number, and 30 real value attributes. These attributes are computed from a digitized image of a Fine Needle Aspiration (FNA) procedure of a breast mass and are used to describe the characteristics of the cell nuclei present in the image. The HOTSM outlier detection approach, which handles anomalies in two stages, was proposed in the current study. First, the Inter Quartile Range (IQR) was employed to diminish the influence of outliers. After the analysis had been finished, the non-outlier data was transmitted to an isolation forest, wherein the absolute mean error was calculated. Pearson's Correlation was employed to minimize the dimensionality. **Findings:** For the performance evaluation, two datasets are generated; one using isolation forest and the other using HOTSM. The performance of both datasets is tested using SVM, Decision Tree, and Random Forest classifiers, highest accuracies are obtained as 97.80 %,96.80%, and 98.4% respectively. It was found that the dataset generated using the proposed method performed well. The proposed model is capable of identifying Breast cancer, more accurately. **Novelty:** The Interquartile Range has been utilized for altering the traditional isolation forest algorithm, enhancing performance metrics. The thorough removal of anomalies reduces the likelihood of misdiagnosis, yet they cannot exclude all outliers.

Keywords: Outliers; Breast Cancer; Accuracy; Machine Learning; Hybrid

1 Introduction

Cancer invades the human body if the cell grows abnormally and can filter through or spread throughout the body[1]. Breast cancer has a high mortality rate and is regarded as the second-leading factor in female fatalities. According to the WHO, breast cancer[2] affects people more frequently than any other type of cancer, with more than 2.3 million cases reported each year. In 95% of the countries in the globe, breast cancer is the leading or secondary cause of cancer-related deaths for women. Globally, it is projected that 10 million people will die from cancer and 20 million new cases of the disease have been reported. The disease is characterized by abnormal cell amplification in the breast, Breast cancer can be classified into two types either malignant or benign.

Machine learning has been the focus of numerous studies for the identification of breast cancer[3]. Regardless of the dataset's characteristics; the study has consistently focused on improving prediction accuracy to enable accurate diagnosis. Nevertheless, unless enhanced with specific data mining approaches, ML algorithm modalities that have been demonstrated on various breast cancer datasets still cannot deliver accurate and consistent results in diagnosis[4].

Preprocessing was applied to the WDBC dataset, which was acquired from the Kaggle repository. It is an open source. This dataset is the Wisconsin Diagnostic Breast Cancer (WDBC) dataset which was donated on November 1st, 1995. Muhammet F A[4] 2020 proposed a model to increase the classification accuracy of breast cancer disease. The gain ratio was utilized to determine the features and model the features using the 10-fold cross-validation method with six algorithms. Similarly, N. F. Idris N.F et al. in 2021[5] concentrated on combining a Machine Learning algorithm with several attribute selection techniques were evaluated and the results determined the most effective strategy. Correlation-based feature selection, recursive feature elimination, linear discriminate analysis, and PCA were the features that were chosen by Harikumar R et al[6] in 2021. The comparison study performed by Kumari et al. in 2020 developed a model for the, the investigation was done using different cross-validation levels and split training dataset percentages in order to find how to enhance the classification algorithm on the WDBC dataset[7]. Improvement was seen when the training set was 85.5%, which produced 99% accuracy. It is acceptable to claim that the training set's overfitting is what caused this improvement. Similar research was done on the WBDC, WDBC, and Coimbra datasets by Raj et al. in 2021[8], and that paper suggested a fuzzy strategy for enhancing ML algorithms. Karczmarek et al. in 2021 used the Naive Bayes algorithm[9]to examine the Light gradient boosting Model, AdaBoost, and Extreme gradient boosting feature selection approaches. Khan MAH et al. in 2020[10] once again tested the Correlation-based Feature Selection, Univariate Selection, and Recursive Feature Elimination feature selection approaches. These feature selection techniques were subjected to the Random Forest, which was then tested on the WDBC dataset. Recent research on the WDBC dataset documented the study and analysis of machine learning algorithms as well as the methods employed to enhance the performance of machine learning algorithms Islam et al. in 2020[11]. Prior to applying it to any ML techniques, Chen et al. in 2021[12] presented a method using the WDBC dataset's grouping and noise removal. According to a review of the literature on the cutting-edge techniques used, as far as we are aware, no research has been done to examine the presence of outliers on the WDBC dataset; hence the issue of multi-linearity in the WDBC dataset still exists.

2 Methodology

Figure 1 depicts the framework for determining the likelihood of breast cancer disease. The strategy entails obtaining a dataset on the disease of breast cancer and preprocessing it to get rid of outliers and missing information. Additionally, a strongly correlated feature discovery algorithm was performed on the pre-processed dataset, and the outcomes were worn in machine learning techniques to determine whether a patient's tumors were benign or malignant. In time, a performance score based on the confusion matrix was used to compare the results.

2.1 Description of Data

The Kaggle repository provided the information needed for this investigation. There are 569 cases that are benign or cancerous in this dataset, also known as the WDBC dataset. When this occurs, 212 cases (37.26%) are malignant, while 357 cases (62.74%) are benign. The dataset has 32 attributes, including id, 31 real value attributes, and class attribute labels (diagnosis: B= Benign, M= Malignant). These characteristics are used to describe the characteristics of the cell nuclei in a picture of a breast mass biopsy that was captured digitally. The Radius, Texture, Smoothness of the Perimeter Area, Compactness, Concavity, Concave spots, and Symmetry Fractal dimension are among the eleven real-valued properties of cell nuclei that are computed in the WDBC dataset. A total of 30 traits were created by estimating the means, standard errors, and worst values for each of these characteristics.

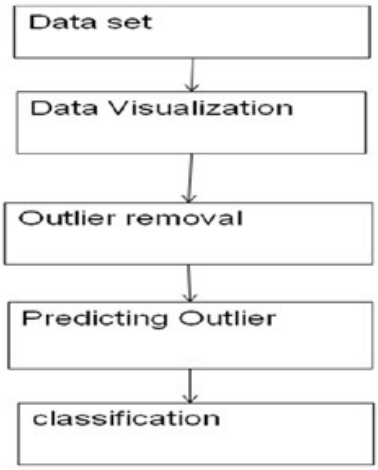

Fig 1. Represents the Proposed Framework in which the dataset is collected, visualized, outliers are removed and predicting of outliers is done and accuracy is calculated using classification

2.2 Preprocessing

Before using machine learning algorithms on the data set, it must be adjusted. The quality and pre-processing of the data collection enhance the model's functionality and precision. These are the preprocessing stages that are employed:

2.2.1 Missing Values Checking

There are 569 instances in the data set which consists of 32 records. It was noted that the variable id, which only maintains a serial record of the instances, has no effect on disease prediction or the dataset's description. The id feature was consequently eliminated. Unnamed: 32, the final feature, always had the value null. The feature was also eliminated from the dataset as a result of what might have been an error in the data collection procedure.

2.2.2 Checking for Outliers

An observation or statistic that deviates from the typical pattern of a distribution is called an outlier. Outliers[13] are a small number of data that are notably different or do not fit the overall trend. The distribution's mean and standard deviation are impacted by the skewness that comes from this. This analysis finds the presence of outliers in the dataset, as seen in Figure 2. Outliers were consequently located and removed from the relevant features.

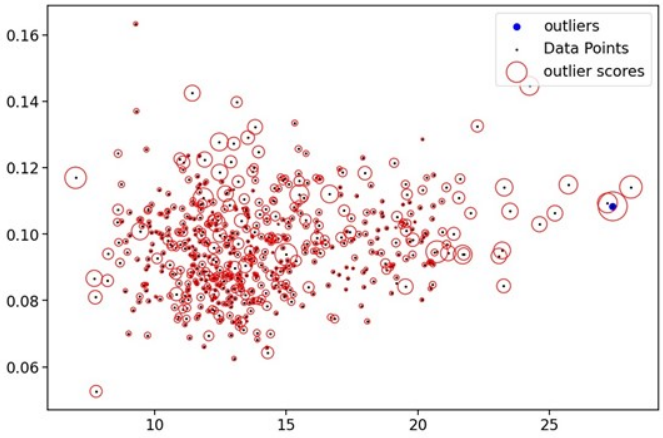

Fig 2. The identified outliers are marked as Data points which plots the display, marking the existence of outliers present in the data set

2.2.3 Detection of Outlier

The Decision Tree algorithm [14] is considered as the base of Isolation Forest. It separates the anomalies by selecting a feature arbitrarily from the available features and then applying a split value across the highest and lowest values of that feature. Instead of profiling typical data snippets, the Isolation Forest isolates abnormalities. Instead of profiling regular instances as in IQR, the goal here is to identify instances that do not fit the normal profile as anomalies.

2.2.4 Hybrid Outlier Selection Method

For Outlier removal from the WDBC dataset, the feature scaling was done using the HOTSM method (Hybrid Outlier Selection Method). In this method, robust scaling is useful when features have marginal outliers. The formulae can be written as

New Xi= (Xi-Xmd) / IQR

Xmd is the median of Xi. IQR means Inter Quartile Range. After performing analysis, the outlier data was passed to the Isolation Forest and the mean absolute error was calculated. The dimensionality reduction was performed using Pearson's Correlation.

Take into account a dataset D with a feature set F.

F = {x1, x2, x3. . . , xn}

Figure 3 displays the WDBC data's independent features' and the dependent class variable's correlation coefficient values.

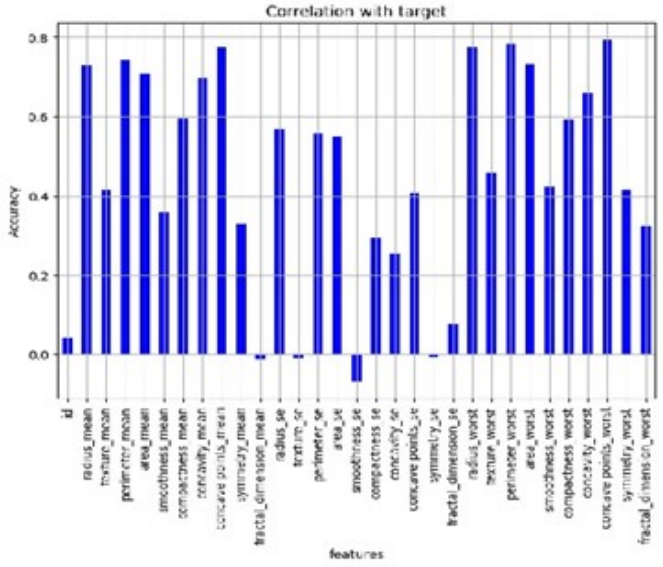

Fig 3. Shows the relationship coefficient among features that are independent and dependent variables related to class

The best predictive characteristics were found using a filter technique. This page calculates and displays the correlation between each attribute. Features in the training dataset that have a correlation of less than 0.7 are eliminated according to the correlation criteria, which are set at 0.7. While other attributes are chosen that have a greater threshold. Based on nine features that were highly linked with the predicted attribute diagnosis, the following Figure 4 was created.

2.2.5 Splitting of Data

To prevent overfitting the model, the data are divided. By removing unneeded variation from the data, dimension reduction is the transfer of data into decreased dimensionality space, leading to the determination of a subspace where the data are located. Feature extraction and feature selection are examples of dimensional reduction approaches. Finding features is the process of feature extraction and eliminating unnecessary, less important, or duplicated dimensions' information from a dataset. It is possible to find and remove as much unnecessary and redundant information using feature selection. As a result, feature selection reduces processing and computing costs while also improving the model created from the chosen data.

In a number of earlier studies, the feature selection method has been applied to healthcare data. Although certain prior studies that deal with the datasets used in this investigation are relevant, in most cases, the performance of such systems did not match predictions. The failure of some systems to recognize the most crucial and highly linked features is one of the causes

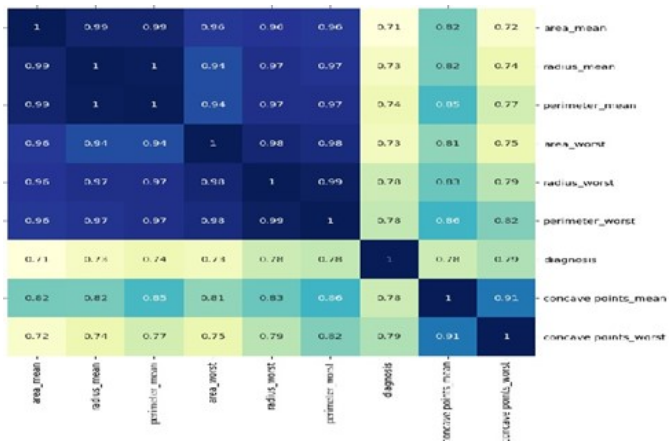

Fig 4. After identification of the outlier best features are identified using correlation and Highly Correlated Features plotting with the threshold set as 0.70

of their poor performance. Training data made up 80% of the dataset for this study, and test data made up 20%.SVM, DT, and RF were selected as the three classifiers.

2.3 Support Vector Machine Algorithm

N-dimensional vector space projection of the data provided as input points and selection of the appropriate hyper-plane to maximize the variance amongst the two classes are both steps in the classification method known as SVM[15,16]. The performance of the SVM is significantly influenced by the choices made for parameters like kernel, C, and gamma. A function called a kernel transforms a low-dimensional space into a multidimensional one that makes categorization easier. The kernel regulates the nonlinearity. In this experiment, the kernel coefficients were modified.

2.4 Decision Tree Algorithm

A powerful predictive learning technique called a decision tree is utilized to address classification and regression issues. It employs a top-down, tree-based advancement technique.[17] In order to ensure that the data in each group is similar, it breaks the data into several different groups at each level employing a staged splitting technique. Every leaf node of the decision tree reflects a separate class; each branch indicates a test at the end, while every inner node represents a test characteristic. The tree develops from the root node by selecting the ”best characteristic” or ”the most effective attribute” among the set of characteristics prior to performing ”splitting.” This is done using the information and entropy gain procedures. The ”best attribute” provides the most beneficial data. The rate of change in the entropy of qualities is known as information gain.

2.5 Random Forest

A group of various independently functioning randomized decision trees is known as a Random Forest. The trees are produced by bootstrapping the data. The class with the greatest number of scores in the random forest model produces the model's forecast for classifying an input vector that is used. Every single tree is known as a random forest and sends a single score according to the total number of prediction values provided[18]. Random Forest is produced by repeatedly splitting a binary tree into comparable nodes. The parent node influences the child node's similarity through inheritance.

2.6 Metrics for evaluating performance

As shown in Table 1, the metric accuracy and prediction before outlier analysis have been evaluated by applying the 2 by 2 confusion matrix. Accuracy indicates a percentage or an opportunity of a total number of accurately foretold events.

As shown in Table 2, the metric accuracy and prediction model accuracy after outlier analysis have been evaluated by applying the 2 by 2 confusion matrix. Accuracy indicates a percentage or an opportunity of a total number of accurately foretold events.

Table 1. Performance measures of different classification algorithms with Performance measures is indicated in the table

Classification Algorithms	Performance Measures	Before Outlier analysis
SVM	Accuracy	0.96
	Recall	0.96
	Precision	0.96
	F1 measure	0.96
Decision Tree Classifier	Accuracy	0.9
	Recall	0.9
	Precision	0.89
	F1 measure	0.9
Random Forest Classifier	Accuracy	0.97
	Recall	0.97
	Precision	0.97
	F1 measure	0.97

Table 2. Performance measure of different classification algorithms with Performance measures is indicated in the table

Classification Algorithms	Performance Measures	IF	HOTSM
SVM	Accuracy	0.973	0.978
	Recall	0.97	0.98
	Precision	0.97	0.98
	F1 measure	0.97	0.98
Decision Tree Classifier	Accuracy	0.9	0.968
	Recall	0.91	0.96
	Precision	0.89	0.96
	F1 measure	0.9	0.96
Random Forest Classifier	Accuracy	0.97	0.984
	Recall	0.97	0.98
	Precision	0.97	0.98
	F1 measure	0.97	0.98

3 Result and Discussion

The suggested methods were tested using various machine-learning methods. The confusion matrix was 2 x 2 made In order to evaluate each algorithm and provide the performance statistic. The performance indicator ”accuracy” was used to assess the proposed models. The traditional strategy performs the worst of all the options, as shown in Figure 5 since outliers increase data unpredictability and decrease statistical power. Classifiers like SVM, Random Forest, and Decision Tree were used out of these SVM and Random Forest have similar accuracy of 98%. The work was done using Google Colaboratory.

3.1 Performance Evaluation

Performance of the proposed model compared with previous work on the wdbc dataset. The name of the authors, classification method and approach used, and accuracy obtained are shown in Table 3.

Table 3. Performance Evaluation of previous work and proposed work

Reference Citation	Classification Algorithms	Accuracy (in %)
(4)	KNN SVM Naïve Bayes Decision tree algorithm Random forest	96.4 96.4 94.7 95.6 95.6
(5)	FID3 algorithm	94.5
(16)	PCB-iForest	97

Continued on next page

Table 3 continued		
Proposed Hostm model	SVM+HOTSM Decision Tree + HOTSM Random Forest classifiers+ HOTSM	97.8 96.8 98.4

3.1.1 FID3 algorithm

In one previous work F. A. Muhammet 2020 performed a comparative analysis of breast cancer using KNN, SVM, Naïve Bayes Decision tree algorithm Random Forest algorithm to test the model it provides the highest accuracy of 96.40% in KNN. N. F. Idris & M. A. Ismail in 2021 for Breast cancer disease classification used the Fuzzy-ID3 algorithm with the FUZZYDBD method. The algorithm gave an accuracy of 94.50%. Heigl M, Anand KA, Urmann A, Fiala D, Schramm M, and Hable 2021 used PCB-iForest as the Improvement of the Isolation Forest Algorithm for outlier detection. It provides the highest accuracy of 97.0%. In this work, the collected dataset was made free from outliers by using Isolation Forest first and then noted classification accuracies. The proposed model HOSTM was applied to the collected dataset and performed classification. HOTSM algorithm works in the concept that after performing Outlier analysis, the Interquartile range is calculated and they are passed to PCB for dimensionality reduction. Best features are identified and on applying classification algorithms like Random Forest, SVM, and Decision Tree, the Random Forest classifier with the proposed approach provides the highest accuracy of 98.40%, which is the highest as compared to all the previous works.

4 Conclusion

In order to enhance the precision of predicting the prognosis of a breast cancer disease, the research primarily focuses on enhancing machine learning models. In this study, a coherent model for outlier detection is proposed. In data mining it is mandatory to remove the outliers from the data sets otherwise it may affect the performance of the model. Here, the proposed model improved the performance of the prevalent isolation forest and created a new data set that is free from outliers. Handling outliers in two levels removes the biases that exist while using a single model. The new data set generated using the proposed model did the breast cancer prediction more accurately Random Forest, SVM, and Decision Tree algorithms were used to generate the model and obtained accuracies of 98.4%, 97.8%, and 96.8%, respectively. This model can also be applied to other datasets for getting better performance measures. Most of the previous work in breast cancer prediction didn't perform the outlier removal. However, it is necessary because the surveyed datasets have a high probability of outliers. Presently, the proposed model is tested with only binary classification problems; in the future. It can be tested with multi-classification problems. The findings demonstrate that HOTSM approaches for outlier detection in combination with different classification algorithms may offer practical tools for inference in this context. The performance of classification systems on various feature selection methodologies needs to be improved so that they can forecast more variables.

References

1) What Is Breast Cancer. 2021. Available from: http://www.cdc.gov/breast-cancer/index.html.
2) Khamparia S, Bharati P, Podder D, Gupta A, Khanna TK, & D NHP, et al. Diagnosis Of Breast Cancer Based On Modern Mammography Using Hybrid Transfer Learning. *Multidimensional Systems And Signal Processing*. 2021;32. Available from: https://doi.org/10.1007/s11045-020-00756-7.
3) Derangula S, & P K Edara, Karri. Feature Selection Of Breast Cancer Data Using Gradient Boosting Techniques Of Machine Learning. *Clinical Medicine*. 2020. Available from: https://www.academia.edu/66487804.
4) Muhammet FA. A Comparative Analysis Of Breast Cancer Detection And Diagnosis Using Data Visualization And Machine Learning Applications, . *Healthcare*. 2020. Available from: https://doi.org/10.3390/healthcare8020111.
5) Idris NFA, Ismail MA. Breast Cancer Disease Classification Using Fuzzy-ID3 Algorithm With FUZZYDBD Method: Automatic Fuzzy Database Definition. *PeerJ Computer Science*. 2021. Available from: https://doi.org/10.7717/peerj-cs.427.
6) Harikumar R, Sannasi C. Effective Classification Framework For Breast Tumors Using Optimized Multi-Kernel SVM With Controlled Skewness. *International Journal of Aquatic Science*. 2021. Available from: https://www.researchgate.net/publication/344027095.
7) Kumari M, Singh V, Ahlawat P. Automated Decision Support System For Breast Cancer Prediction. *International Journal On EmergingTechnologies*. 2020;11(4):193–201. Available from: https://www.researchtrend.net/ijet/pdf/Automated%20Decision%20Support%20System%20for%20Breast%20Cancer%20Prediction%20Madhu%20Kumari%2027999.pdf.
8) Raj S, Singh S, Kumar A, Sarkar S, Pradhan C. Feature Selection And Random Forest Classification For Breast Cancer Disease. In: Data Analytics in Bioinformatics: A Machine Learning Perspective. 2021. Available from: https://doi.org/10.1002/9781119785620.ch8.
9) Karczmarekp K, Dariuszczerwinski P. Fuzzy C Means-based isolation Forest. . 2021. Available from: https//doi.org/10.1016/j.asoc.2021.107354.
10) Khan MAH, Thomson B, Debnath R, Motayed A, Rao MV. Nano wire Based Sensor Array for Detection of Cross Sensitive Gases Using PCA and Machine Learning Algorithms. *IEEE Sensors*. 2020. Available from: https://doi.org/10.1109/JSEN.2020.2972542.
11) Islam MM, Md R, Haque H, Iqbal, Md M, Hasan M, et al. Breast cancer prediction: A comparative study using machine learning techniques. 2020. Available from: https://doi.org/10.1007/s42979-020-00305-w.
12) Chen, Hua, Wang, Nan, Du, Xueping, et al. Classification Prediction of Breast Cancer Based on Machine Learning. *Computational Intelligence and Neuroscience*. 2023;6530719:9–9. Available from: :https://doi.orgl/10.1155/2023/6530719.

13) Zhou S, Hu C, Wei S, Yan X. Breast Cancer Prediction Based on Multiple Machine Learning Algorithms. *Technology in Cancer Research & Treatment*. 2024;23. Available from: https://doi.org/10.1177/15330338241234791.
14) Manikandan P, Durga U, Ponnuraja. An integrative machine learning framework For classifying seer Breast cancer. *Sci Rep*. 2023;13. Available from: https://doi.org/10.1038/s41598-023-32029-1.
15) Karczmarekp P, Ale. K-Means-Based isolation forest. 2020. Available from: https://doi.org/10.1016/j.knosys.2020.105659.
16) Heigl M, Anand KA, Urmann A, Fiala D, Schramm M, Hable R. On The Improvement Of The Isolation Forest Algorithm For Outlier Detection With streaming data. *Electronics*. 2021. Available from: https://doi.org/10.3390/electronics10131534.
17) Wang H, Jiang W, Deng X, Geng J. A New Method For Fault Detection Of Aeroengine Based on Isolation forest. *Measurement*. 2021;185. Available from: https://doi.org/10.1016/j.measurement.2021.110064.
18) Loo NL, Chiew YS, Tan CP, Mat-Nor MB, Ralib AM. A Machine Learning Approach To Assess Magnitude Of Asynchrony Breathing. *Biomedical Signal Processing and Control*. 2021;66. Available from: https://doi.org/10.1016/j.bspc.2021.102505.