INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Efficient Outlier Detection Using Isolation Forest Based on Robust Scaling and Principal Component Analysis for the Prediction of Anxiety Disorder

T M Prajesha[1]*, **S Veni**[2]

1 Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore, 641 021, Tamil Nadu, India
2 Professor, Department of Computer Science, Karpagam Academy of Higher education, Coimbatore, 641 021, Tamil Nadu, India

OPEN ACCESS

Received: 19-03-2023
Accepted: 03-07-2023
Published: 05-08-2023

Citation: Prajesha TM, Veni S (2023) An Efficient Outlier Detection Using Isolation Forest Based on Robust Scaling and Principal Component Analysis for the Prediction of Anxiety Disorder. Indian Journal of Science and Technology 16(29): 2244-2251. https://doi.org/10.17485/IJST/v16i29.638

* **Corresponding author**.

tmprajesha@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 anxiety disorder. Presence of outliers may affect the model performance. Hence, here the main consideration is the removal of outliers. **Methods:** This study proposed an outlier detection method IF-RSPCA in which outliers are handled in two phases. In the first phase, reduce the impact of outliers by using Inter Quartile Range (IQR), then dimensionality reduction is done using principal component analysis. Training time can be reduced by Principal Component Analysis. In the second phase of IF-RSPCA, check the outliers for still exist and removed. **Findings:** For the performance evaluation two datasets are generated; one using conventional isolation forest and the other using IF-RSPCA. Performance of both datasets are tested using K Neighbors, Decision Tree and Naïve Bayes classifiers, highest accuracies are obtained as 96.10 %,93.80% and 90.2%, respectively. It is found that the dataset generated using proposed method performed well. As compared to previous works on survey datasets, the proposed model is capable of identifying the anxiety, more accurately. **Novelty:** Conventional isolation forest algorithm is modified using Inter Quartile Range there by improve the performance measures. It helps to removes all the outliers completely. The previous models for the predictions of anxiety disorder are developed without performing the outlier detection; hence chances of misclassification exist. In several previous works, isolation forest combined with clustering, but it is not capable of removing all the outliers. In these cases, outlier handling is performed in a single phase only.

Keywords: Principal Component Analysis; Isolation Forest; Robust Scaling; K Nearest Neighbors; Decision tree

1 Introduction

According to the reports of World Health Organization nearly 10% of population in the world suffering some kind of mental disorders. The persons suffering mental disorders as increasing as time by time. But there exists a treatment gap this is due to the lack of specialist in this field [(1)]. So, it is necessary to build a model for the prediction of mental disorder. Anxiety disorder is the main consideration in this paper. A person suffering anxiety disorder having the feeling of fear and worry that is not focuses on a specific situation or object. If the person not getting treatment in the initial stage it may lead to severe stage, which make suicidal thoughts [(2)], depression etc. So, it is necessary to diagnose the mental disorder at the initial stage.

In all the previous works for the anxiety prediction surveyed data set was used for analysis. Commonly people hesitate to participate such surveys. So, we cannot ensure the information obtained are genuine or correct. Chances of wrong data is very high in surveyed data set. Hence it is essential to remove such data for the better performance of the model. But all the previous works the dataset used straightly for developing the machine learning model.

Outlier detection is a vital part of machine learning process. Effectiveness of the machine learning model is highly based on accuracy of the data. The statistical methods used for analysis highly sensitive to outliers, hence results may change. So, it is necessary to remove outliers in the preprocessing stage. AJ Rosellini at el. developed a model for the prediction of the internalizing disorders. This disorder may start from the childhood. The dataset questionnaires are filled by the parents are used for the analysis. Nine classifiers including SVM, Logistic Regression, K Nearest neighbors etc. used for the analysis.0.76 to 0.83 AUC was obtained with these algorithms [(3)].

Neesha Jothi at el. in 2020 developed a model for the prediction of anxiety disorder. They used Sharply feature selection methods directly to the surveyed dataset without removing the outliers. The model created and tested with several classifiers like Naïve Bayes, Random Forest, J48 etc. [(4)]. A machine learning pipeline model is used for the prediction of Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD) [(5)]. Here 4184 undergraduate students participated in the survey. Machine learning pipeline includes the XGBoost, Support Vector Machine, K Nearest Neighbors, Random Forest, Logistic Regression and Neural network. The final outcome is predicted with another XGBoost Classifier. The model accuracy for the prediction of GAD is 73%.

Arkaprabha Sau, Ishita Bakta in 2019 suggested an anxiety prediction model that used 10-fold cross validation and 5-fold cross validation along with CatBoost classifiers [(6)]. CatBoost is an ensemble learning algorithm, developed by several classification algorithms, which process sequentially. In 2020 Burke et al. developed a model to classify the suicide attempt history [(7)]. It was implemented using Random Forest and Decision Tree algorithms. Dataset contain 53 features based on Behavioral Health Screen, which is used to test the person have any sign of mental illness. The Random Forest classifier provide 96% accuracy.

k-Means based Isolation Forest is a better model for identifying anomalies [(8)]. Here classical isolation forest augmented in an innovative way. This method helps to generate a search tree having many branches. The number of divisions in each decision tree node is predicted by K-Means algorithm. This method performed better in spatiotemporal data and geographical data. The anomaly score of each record in the dataset is determined by this method.

For the safe and reliable operation, it is very essential to find the faults in aero-engine data. In a dynamic method used for fault detection of aero-engine is proposed by Hongfei and co-authors [(9)]. It is done using isolation forest classifier. An adaptive dynamic threshold is set by sliding window approach. This model was verified using the residual data of the turbofan engine gas path system that belongs to three different fault states. It provides better accuracy and also provides less running time. A set of methods for enhancing the Isolation Forest was proposed in by Pawel at el [(10)]. It was based on Fuzzy C-Means (FCM). It provides the membership grades of each node form isolation tree and also make a cluster. The data points with similar characteristics come under a cluster. This information is also used for calculating the anomaly score. Here author performed the experiment with one cluster, two cluster and with two clusters and distance influence. Experiments are carried out using 27 datasets. But FCM Based Isolation Forest some provide poor performance in some datasets which is below 80% accuracy. That means the chances of outliers still exists.

2 Methodology

In this paper a new model, Isolation Forest based on Robust Scaling and PCA(IF-RSPCA) for the outlier detection is proposed. Here outliers are managed in two phases. In the first phase robust scaling is performed in the dataset. As compared to other scaling method robust scaling reduces the influence of outliers. Still now there exist some observations significantly different from others. Such outliers are then finding out by isolation forest algorithm and then removed. Dimensionality reduction using PCA (Principal Component Analysis) is also applied in this proposed method. It helps to reduce the training time.

The previous models for anxiety prediction are trained using the survey dataset. The possibility of the outliers is more in survey dataset. It is necessary to remove outliers for developing better model. Some outlier detection models are also reviewed [8–10], but it handles the outliers in a single phase. Hence chances of bias exist, removal of outliers depends on a single algorithm. By using hybrid models the removal of all most all the outliers can be ensured.

2.1 Data Collection

Data collection performed in an online survey. Collected details of persons in different ages, starting from 18 and above. Anxiety is the consistent feeling of worry and fear that is not due to a particular situation or object. There are several psychometric tools are available for the analysis [8–10]. Here the data collection mainly based on patient health questionnaire (PHQ) and hospital anxiety and depression scale (HADS). Dataset contain 1000 observations and 23 attributes. The attributes collected are given in following tables. They are categorized as general, questions based on HADS and questions based of PHQ.

Table 1. Attributes in Dataset

No	Attributes	Data Type
1	Sex	Nominal: Male (0) or female (1)
2	Age group	Nominal
3	Education	Ordinal
4	Financial Situation	Ordinal
5	Health Status	Ordinal
6	Unemployed	Nominal: Yes or No
7	Student	Nominal: Yes or No
8	Feel tense or 'wound up'	0,1,2,3
9	Feels like something awful to be happen	0,1,2,3
10	Have a worrying thought	0,1,2,3
11	Can feel relax	0,1,2,3
12	Have a frightened feeling	0,1,2,3
13	Have feeling of restless	0,1,2,3
14	Suddenly feel as panic	0,1,2,3
15	Feel little interest or pleasure for doing things	0,1,2,3
16	Feels like hopeless and depressed	0,1,2,3
17	Facing trouble to asleep or sleeping more	0,1,2,3
18	Feels little energy and tired	0,1,2,3
19	Eat too much and poor appetite	0,1,2,3
20	Feeling bad about yourself	0,1,2,3
21	Difficulty in concentrating	0,1,2,3
22	Slowly moving and speaking	0,1,2,3
23	Suicide thoughts or self-hurting	0,1,2,3

2.2 Robust Scaling

Data scaling is one of the preprocessing procedures in which data converted into some other range that is suitable for the machine learning modelling. Models that developed using scaled data provide better performance as compared to the model developed with unscaled data. Here scaling is performed based on inter quartile range (IQR). Hence it is robust to outliers. Robust scaling does not completely avoid outliers, it only reduces the impact of outliers. This method is so much useful when the data set consist of marginal outliers.

To find IQR the dataset divided into quartiles. First find the median(q_2) of the entire data. Based on that the data set divided into upper and lower halves. If q_3 and q_1 are the median of upper half and lower half of the data. Then IQR can be calculated as q_3 - q_1 [11].

Suppose X_i is a value in a distribution and X_{md} is median of entire distribution then Robust scaled value of X_{new} can be calculated as

$$X_{\text{new}} = \frac{X_i - X_{md}}{IQR} \tag{1}$$

2.3 Principal Component Analysis

PCA is a famous standard linear method based on the second order statistics of the data. It is a widely used preprocessing method for decorrelating and compressing data before classification and regression.

Multidimensional data set can be reduced to lower dimensions by using PCA. Data analysis is easier and faster when dimensionality reduction is performed. Here original data converted into new coordinate system. First principal component generated from the projection of data that makes greatest variance. The second greatest variance gives second principal component and so on[12]. To find out the principal components the covariance matrix is created from the variables. Then compute eigen values and eigen vectors of the covariance matrix to identify the principal components[13]. The highest eigen value generate the first principal component[14]. Here presence of outliers may affect the results. So again, needed some methods to identify outliers in the observations.

2.4 Isolation Forest

The data points that have different characteristics from normal instances are called anomalies. So, it is easy to isolate from all other data points[15,16].

In this algorithm outliers are isolated with the help of a binary tree called iTree.

Procedure of the iForest algorithm is given below

- In dataset Di set of all feature variables are Q
- Select a random feature Fi from Q
- A value p is selected from a feature Fi, which is lies between maximum and minimum values of Fi and considered as root in iTree
- The data in Di which is less than p is written in the left side of the iTree and the data which is greater than p written in the right side of the iTree
- Repeat the above process until the tree reaches its maximum depth or data set have only one more piece of dataCombine the results of iTrees made from all Fi in Q. Calculate the anomaly score of every data point[17].

$$S(x, \varphi) = 2^{\frac{-E(h(x))}{C(\Psi)}} \tag{2}$$

Ψ is the count of samples used to train iTree

$C(\Psi)$ is a normalization factor

2.5 Proposed Algorithm

- Q is the set of all features in Dataset Di
- Select a random feature Fi from Q and find the median q2
- Divide the data in Fi into two halves based on value of q_2. All values < q_2 to lower half (LH) and all the values > q_2 reaches to the upper half (UH).
- Find the median of LH(q_1) and median of UH(q_3)
- Calculate interquartile range, IQR= (q_3 - q_1)
- Convert the samples as

$$x_{\text{new}} = \frac{x_i - q_2}{IQR} \tag{3}$$

- Repeat steps from (b) to (f) until all features are selected
- Perform dimensionality reduction
- Sample p selected from new Fi which lies between the maximum and minimum values of Fi considered as root of iTree
- If the data point in Fi <p added to the left side of the iTree and if iTree
- Repeat the process until the tree reaches the maximum depth or take all the in Fi
- Combine the results of iTrees made from all Fi in Q.Calculate the anomaly score of the every data point

$$S(x,\varphi) = 2^{\frac{-E(h(x))}{C(\Psi)}}$$

3 Results and Discussion

Presence of the outlier is identified using Isolation Forest. PyCaret library is used for the implementation of isolation forest. Abnormal observation has anomaly score nearly 1. Out of 1000 attributes 50 attributes are identified as outliers. The outliers are removed and the performance of the dataset evaluated using classification algorithms such as K Neighbors, Decision Tree and Naïve Bayes. Performance measures like Accuracy, Recall, Precision and F1 measures are evaluated for each classifier.

In proposed method two level treatment for the anomaly is performed. Before using isolation forest, robust scaling performed in the dataset. Robust scaling reduces the influence of outliers because scaling performed based on IQR. Next dimensionality reduction performed using PCA. It helps to reduces the training time. Then isolation forest is applied to find the outliers. Then another 50 attributes are identified as outliers.

Table 2. Anomalies detected in IF and IF-RSPCA methods

Algorithm	Anomalies	Number of Positive and negative samples
IF	11, 37, 89, 127, 134, 157, 225, 254, 288, 294, 312, 327, 357, 389, 403, 450, 489, 508, 529, 536, 538, 568, 599, 615, 630, 719, 741, 745, 746, 758, 770, 787, 816, 821, 830, 831, 846, 887, 889, 897, 900, 944, 946, 948, 954, 964, 977, 978, 981, 999	0-769 ,1-181
IF-RSPCA	11, 22, 33, 40, 89, 134, 205, 241, 254, 260, 288, 304, 312, 322, 327, 396, 400, 403, 415, 422, 450, 472, 508, 536, 546, 559, 568, 599, 630, 652, 664, 682, 702, 745, 770, 772, 787, 821, 846, 860, 861, 887, 897, 909, 948, 954, 970, 974, 977, 999	0-759, 1-191

Following table shows performance evaluation of IF and IF-RSPCA. Three classification algorithms used for the study. Here proposed model IF-RSPCA give better results.

Table 3. Comparison of performance measures

Performance Measures	Classification Algorithms					
	K Neighbors Classifier		Decision Tree Classifier		Naïve Bayes Classifier	
	IF	IF-RSPCA	IF	IF-RSPCA	IF	IF-RSPCA
Accuracy	0.946	0.961	0.928	0.938	0.889	0.902
Recall	0.827	0.872	0.837	0.842	0.985	0.976
Precision	0.892	0.923	0.825	0.851	0.655	0.678
F1 measure	0.854	0.893	0.821	0.837	0.784	0.795

Accuracy represents the number of observations correctly predicted among the total number of observations. Figure 1 shows the accuracy comparison of IF and IF-RSPCA. All the three classifiers give better accuracy value for IF-RSPCA. Here K Neighbors classifiers is providing the highest accuracy value of 0.961.

Precision is used to measure correct positive prediction. A model with high precision means it have less false positive predictions. Here the precision of datasets evaluated using K-Neighbors, Decision tree and Naïve bayes classifiers. It is found that the dataset obtained after IF-RSPCA provides better precisions. Highest precision obtained in K Neighbors classifier which is 0.923.

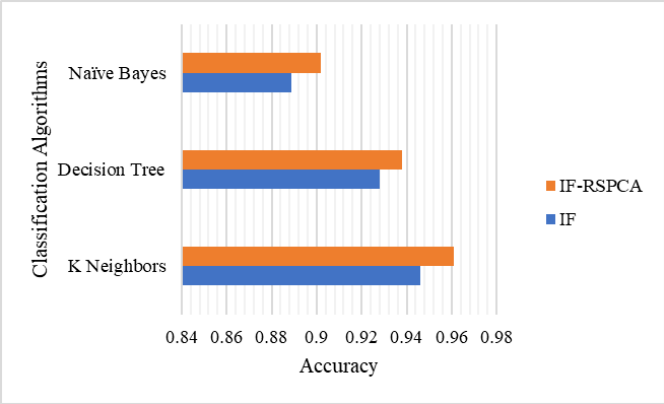

Fig 1. Comparison of accuracy of IF and IF-RSPCA

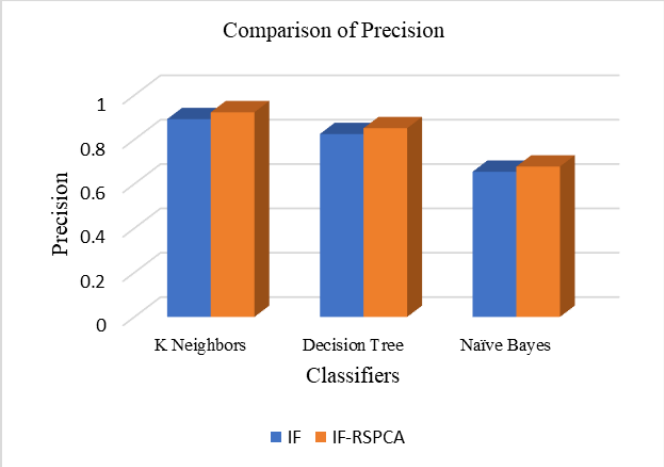

Fig 2. Comparison of precision of IF and IF-RSPCA

Ratio between the correctly predicted positive samples and total positive samples are called recall. In Figure 3 the recall of the dataset obtained using IF and IF-RSPCA are compared using three classifiers. In most of the cases proposed model gave the highest recall.

In Figure 4, graph compare four performance measures such as accuracy, recall, precision and F1 measure of three classifiers in two different datasets generated by IF and IF-RSPCA. Graph shows that proposed model provides the best dataset by avoiding outliers that influence analysis results. In the performance analysis of proposed method K Neighbors provides highest accuracy, precision and F1 measure of 0.961,0.923,0.893 respectively. Naïve Bayes gives highest recall of 0.976.

3.1 Performance Evaluation

Performance of the proposed model compared with previous work on anxiety disorder dataset collected using different online surveys. Name of the authors, classification method and approach used and accuracy obtained is showing in Table 4.

In one previous work authors used Sharply value for the feature selection. After feature selection model created and tested with three classifiers Naïve Bayes, Random Forest and J48. It provides highest accuracy of 95.70% in J48. Arkaprabha Sau and Ishita Bakta in 2019 uses 10-fold cross validation and 5-fold cross validation with Catboost. Here 5-fold cross validation provided the better accuracy. Mathew D and coauthors developed an ensemble model with SHAP feature selection. That provide an accuracy of 73.00% for the prediction of anxiety disorder. In this work the collected dataset made free from outliers by using Isolation Forest first and then noted classification accuracies. The proposed model IF-RSPCA applied in the collected dataset and performed classification. From the analysis it is observed that KNN classifier with the proposed approach provide highest accuracy of 96.10%, which is the highest as compared to all the previous works.

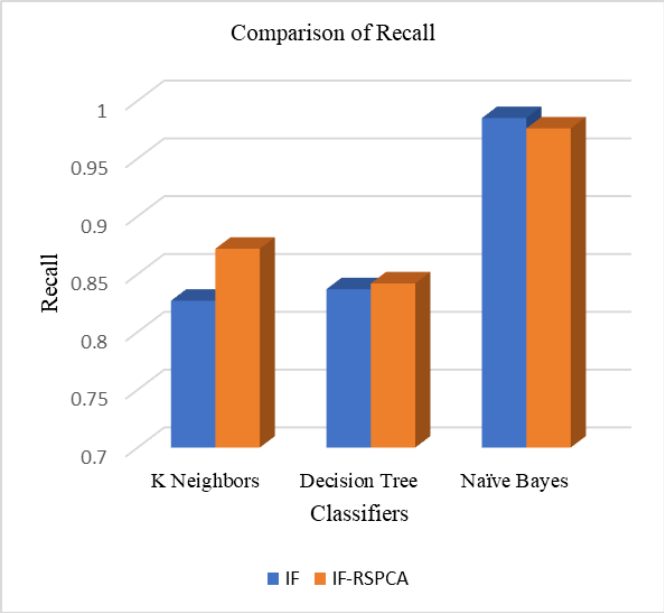

Fig 3. Performance comparison based on Recall

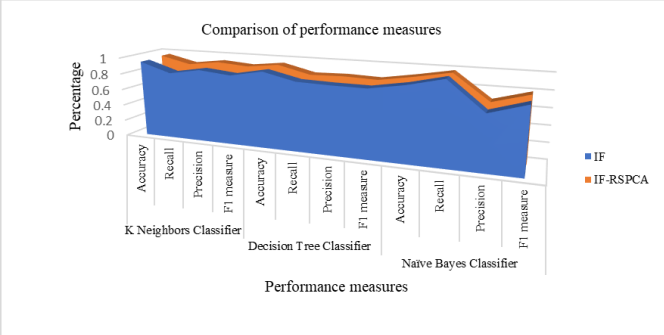

Fig 4. Comparison of performance measures

Table 4. Performance Evaluation

Authors	Classification and approach	Accuracy (in %)
Neesha Jothi,Wahidah Husain,Nur'Aini Abdul Rashid in 2020	Naïve Bayes +Sharply value	80.70
	Random Forest+Sharply value	90.50
	J48 +Sharply value	95.70
Arkaprabha Sau,Ishita Bakta in 2019	10-fold cross validation +catboost	82.60
	5-fold cross validation+catboost	89.30
Matthew D. Nemesure at el. in 2020	Stacking classifier with SHAP feature selection	73.00
In this work	KNN+IF	94.60
	Decision Tree+IF	92.80
	Naïve Bayes+IF	88.90
	KNN+IF-RSPCA	**96.10**
	Decision Tree+IF-RSPCA	93.80
	Naïve Bayes+IF-RSPCA	90.20

4 Conclusion

In this study, an efficient model for outlier detection is proposed. In data mining it is necessary 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 conventional isolation forest and creates new data set which is free from outliers. Handling outliers in two levels removes the biases that exist while using single model. New data set generated using proposed model did the anxiety prediction more accurately. K Nearest Neighbors, Decision Tree and Naïve Bayes algorithms are used to generate the model and obtained accuracies 96.10 %,93.80% and 90.2% , respectively. This model can also apply to other datasets for getting better performance measures. Most of the previous work in anxiety prediction didn't perform the outlier removal. But it is necessary because the surveyed datasets have high probability of outliers. Presently, the proposed model tested with only binary classification problem; in future. it can be tested with multi classification problems.

References

1) Arif M, Basri A, Melibari G, Sindi T, Alghamdi N, Altalhi N, et al. Classification of Anxiety Disorders using Machine Learning Methods: A Literature Review Insights of. *Biomedical Research*;2020(1). Available from: https://doi.org/10.36959/584/455.
2) Lento RM, Boland. Clinical Handbook of Anxiety Disorders. Springer. 2019;p. 203–220. Available from: https//doi.org/10.1007/978-3-030-30687-8_11.
3) Anthonyj, Rosellini S, Liu, Gracen, Anderson S, Sbi, et al. Developing algorithms to predict adult onset internalizing disorders: An ensemble learning approach. *Journal of Psychiatric Research*. 2020;121:189–196. Available from: https://doi.org/10.1016/j.jpsychires.2019.12.006.
4) Jothi N, Husain W, Rashid NA. Predicting generalized anxiety disorder among women using Shapley value. *Journal of Infection and Public Health*. 2021;14(1):103–108. Available from: https://doi.org/10.1016/j.jiph.2020.02.042.
5) Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports*. 1980;11(1). Available from: https://doi.org/10.1038/s41598-021-81368-4.
6) Sau A, Bhakta I. Screening of anxiety and depression among the seafarers using machine learning technology. *Informatics in Medicine Unlocked*. 2019;16:100149. Available from: https://doi.org/10.1016/j.imu.2018.12.004.
7) Burke TA, Jacobucci R, Ammerman BA, Alloy LB, Diamond G. Using machine learning to classify suicide attempt history among youth in medical care settings. *Journal of Affective Disorders*. 2020;268:206–214. Available from: https://doi.org/10.1016/j.jad.2020.02.048.
8) Karczmarek P, Pedrycz AKW, Al E. K-Means-based isolation forest . 2020. Available from: https://doi.org/10.1016/j.knosys.2020.105659.
9) Wang H, Jiang W, Deng X, Geng J. A new method for fault detection of aero-engine based on isolation forest. *Measurement*. 2021;185:110064. Available from: https://doi.org/10.1016/j.measurement.2021.110064.
10) Karczmarek P, Kiersztyn A, Pedrycz W. Dariusz Czerwinski, Fuzzy C-Means-based Isolation Forest. *Applied Soft Computing*. 2021;106. Available from: https//doi.org/10.1016/j.asoc.2021.107354.
11) 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;66:102505. Available from: https://doi.org/10.1016/j.bspc.2021.102505.
12) Jo HS, Park C, Lee E, Choi HK, Park J. Path Loss Prediction Based on Machine Learning Techniques: Principal Component Analysis, Artificial Neural Network, and Gaussian Process. *Sensors*;20(7):1927. Available from: https://doi.org/10.3390/s20071927.
13) Khan MAH, Thomson B, Debnath R, Motayed A, Rao MV. Nanowire-Based Sensor Array for Detection of Cross-Sensitive Gases Using PCA and Machine Learning Algorithms. *IEEE Sensors Journal*. 2020;(11). Available from: https://doi.org/10.1109/JSEN.2020.2972542.
14) Huang Y, Jin W, Yu Z, Li B. A robust anomaly detection algorithm based on principal component analysis. *Intelligent Data Analysis*. 2021;25(2):249–263. Available from: https://doi.org/10.3233/IDA-195054.
15) Salman AW, Farizi. Indriana Hidayah; Muhammad Nur Rizal, Isolation Forest Based Anomaly Detection: A Systematic Literature Review. *8th International Conference on Information Technology*. 2021. Available from: https://doi.org/10.1109/ICITACEE53184.2021.9617498.
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;10(13):1534. Available from: https://doi.org/10.3390/electronics10131534.
17) Luan S, Gu Z, Freidovich LB, Jiang L, Zhao Q. Out-of-Distribution Detection for Deep Neural Networks With Isolation Forest and Local Outlier Factor. *IEEE Access*. 2021;9:132980–132989. Available from: https://doi.org/10.1109/ACCESS.2021.3108451.