Retraction

Retraction: Prediction of Presence of Breast Cancer Disease in the Patient using Machine Learning Algorithms and SFS (IOP Conf. Ser.: Mater. Sci. Eng. 1099 012003)

Published 12 May 2023

This article has been retracted by IOP Publishing following an allegation that this article may contain tortured phrases [1].

IOP Publishing has investigated and agrees the article contains a number of nonsensical phrases that feature throughout the paper [2], to the extent that the article makes very little sense. This casts serious doubt over the legitimacy of the article.

IOP Publishing wishes to credit PubPeer commenters [3] for bringing the issue to our attention.

The authors neither agree nor disagree to this retraction.

[1] Cabanac G, Labbe C, Magazinov A, 2021, Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals, arXiv:2107.06751v1

[2] V Chaurasia et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1099 012003

[3] https://pubpeer.com/publications/078B0582B40667762452422CFD5A00

Retraction published: 12 May 2023

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Published under licence by IOP Publishing Ltd
Prediction of Presence of Breast Cancer Disease in the Patient using Machine Learning Algorithms and SFS

V Chaurasia¹, MK Pandey¹ and S Pal²
¹Research Scholar, Department of Computer Applications, VBS Purvanchal University, Jaunpur
²Head, Dept. of Computer Applications, VBS Purvanchal University, Jaunpur

E-mail: drsaurabhpal@yahoo.co.in

Abstract. Breast cancer disease is recognized as the common extensive malignant tumor in between women. Identification of the initial stage of malignant growth may treatment of this disease. Early treatment helps to alleviate the disease and helps anticipate its recurrence in women. Experts have used some fact checks and different medical methods or equipment to improve the accuracy of conclusions in clinical medical service management. In this article, it extensively discussed the implementation of data mining strategies to detection as well as prediction of breast malignant tumors, including random forest (RF), support vector classifier (SVC), k-nearest neighbors (KNN), linear discriminant analysis (LDA), Gradient Boosting Classifier (GBC), Decision Tree (DT) In addition, principal component analysis (PCA) to underline changes and show strong patterns in the informational index. The connection framework is likewise used to show the level of close relationship between attributes. The sequential feature selection (SFS) method is used for comparing the accuracy of a data set with all features and the accuracy of a classifier with selected features. The results show that RF_sfs, KNN_sfs, SVC_rbf and SVC_sfs have the highest and equal accuracy, which is 97.66%. They perform well and can predict the growth of harmful malignant tumors.

Keywords. Breast cancer, Machine learning, Clinical, Medical, SFS, PCA, Correlation matrix

1. Introduction
Breast malignant growth may be the deadliest disease among women. When cells in bosom tissue develop abnormally, they structure a mass of tissue called a tumor. The basic characteristics of these tumors are malignant or benign, the malignant tumors that are harmful and the benign tumors are non-cancerous. Tumors that can fuse with benign breasts are not dangerous and will not increase to all parts of the body. Harmful tumors are life-threatening, they can extend to different body parts, and after removal, and we often see dangerous tumors recurring anyway. Disease cells continue to divide from breast tumors and can go from side to side lymphatic veins to arrive at different body parts. It might associate with various tissues in different body parts and form into another tumor with a structure that may impede by and large limit. Some tests were performed to analyze the patients, and careful biopsies were performed where the experience activities were required. Today, the application of information mining expanded in the clinical sector [1]. The two perspectives that uphold the utilization of information mining in the welfare section that promote to avoid malignant growth, such
as early recognition, early avoidance, and indication-based drugs to correct clinical data errors. Many AI calculations are used to better treat patients. The classification in supervised method and clustering and regression in unsupervised methods are useful for identify the patients. There are many procedures for predicting and characterizing the growth design of a patient with malignant tumors [2]. Computational knowledge assumes a significant function in diagnosing and picking insightful decisions. There are an enormous number of clinical applications and analytic methods that can be grouped utilizing clever computational arrangement capacities. Data mining innovation, furnishes clients with approaches to discover new and hidden examples from huge scope data. In the healthcare field, the information discovered can be used to improve the accuracy of analysis by medical service executives and clinicians, thereby increasing and sinking the stage of caution. The concealment of data, the extraction of vague and imaginable valuable data” is called knowledge disclosure in data [3]. The purpose behind conjectures in information mining is to facilitate discover designs in information to improve their prosperity [4]. Data mining assumes a significant function to anticipate ailments in the clinical field. The prescience of the prediction is a data mining model. In this paper, the aftereffects of strategies were examined and decisions about future examination were given. The forecast of breast cancer made by a wellbeing specialist isn’t 100% precise [5]. This paper looks at execution of six algorithms and their blend utilizing outfit approach that is reasonable for direct interpretability of their outcomes. The aim of this method is to plan an important model that will be able to identify the difference between malignant tumors and benign tumors, and the model will also compare the accuracy of a classifier with all features and a model with sequential feature selection (SFS) by making a comparison [6].

The second part is a literature review, which provides a comprehensive literature review by focusing on research gaps, existing problems and challenges to existing research. The third part contains the discussion of tools and techniques for each model/tool/technique/algorithm, which are very important to the development of the proposed method. The fourth part contains the experimental method, which describes each step/stage of the method in detail. The fifth part contains the experimental setup, discusses the simulation environment and the required parameters as well as the description of the data set. The results and discussion of the sixth part describe the effectiveness of the model through the analysis of the results. The seventh part contains conclusions, describing the overall work carried out and the future scope of the research. The eighth part contains references.

2. Literature Review
Numerous works in the writing identified with breast cancer disease conclusion have prompted this work by utilizing data mining strategies. A short writing study is introduced here in Table 1.

| Author                      | Disease/Applied Area | Technique Used          | Accuracy   |
|-----------------------------|----------------------|-------------------------|------------|
| Sarvestani et. al (2010)    | Breast Cancer        | Naïve Bayes, ANN        | 84.5%, 86.5% |
| Lavanya & Rani (2012)       | Breast Cancer        | CART                    | 69.23%     |
| Chaurasia & Pal (2020)      | Breast Cancer        | LR                      | 99%        |
| Chaurasia & Pal (2014)      | Breast Cancer        | RepTree                 | 71.32%     |
| Reference | Type | Disease | Classifier 1 | Accuracy 1 | Classifier 2 | Accuracy 2 |
|-----------|------|---------|--------------|------------|--------------|------------|
| [10]      | Survivability | Malignant/Cancer of breast | RBF Network | 73.77% | Simple Logistic | 74.47% |
| Agarap (2018) [11] | | | MLP | 99.04% |
| Bazazeh&Shubair (2016) [12] | | Malignant/Cancer of breast | RF | 99.9% |
| Bataineh (2019) [13] | | Malignant/Cancer of breast | MLP, KNN, CART, NB, SVM | 96.70%, 96.27%, 91.00%, 93.62%, 96.42% |
| Bhardwaj & Tiwari (2015) [14] | | Malignant/Cancer of breast | GONN algorithm | 98.24% |
| Nguyen et.al (2013) [15] | | Malignant/Cancer of breast | Random Forest classifier and Feature selection technique | 99.8% |
| Aruna, et.al (2011) [16] | | Malignant/Cancer of breast | Naïve Bayes, SVM, RBF, J48, CART | 96.50%, 96.84%, 96.66%, 94.59%, 94.27% |
| Parekh & Jacobs (2017) [17] | Framework for Malignant/Cancer of breast and Tumor Biology | Radiomic feature mapping framework (AUC) | 91% |
| Aggrawal & Pal (2020) [18] | Prediction of death events and diagnosis of heart disease | Random Forest with SFS | 86.67% |

3. Tools and Techniques
In this planned work, six crucial classifiers such as RF, GBC, SVC, LDA, KNN and DT are used. Each classifier has its own burden, which will help the next classifier. Other than these basic classifiers we have used SFS technique to extract the classifiers accuracy.

3.1 Random Forest (RF)
RF is an all-round learning method for classification, repeated execution and different tasks. It generates a working tree by developing a large number of decision trees and generating classification or regression classes within the training time [19][20]. DT tends to over-adapt to the irregular forest privileges of its training set [21].

3.2 Support Vector Classifier (SVC)
SVM is a supervised learning model with related learning innovation, which is utilized to read information for classification and regression examination [22]. The goal of SVM development is to find a hyperplane in the N-dimensional space that collects data. To disconnect the two kinds of information central focuses, numerous conceivable hyperplanes can be chosen.

3.3 K-Nearest Neighbors (KNN)
The K-NN is a fundamental methodology that stores each accessible case and gatherings new cases reliant on the closeness metric. KNN has been used as a non-parametric system for quantifiable evaluation and example acknowledgment. K-NN is such a model based learning or lethargic realizing,
where the breaking point is simply approximated locally, and all calculations will be held until work evaluation is performed. Since this system relies upon portrayal for request, normalizing the preparation data can improve its precision [23][24].

3.4 Linear Discriminant Analysis (LDA)
LDA is an approach utilized for amount, model acknowledgment and AI to find a straight blend of highlights portraying or separating at any rate two classifications or occasions. Subsequent combinations can be utilized as immediate classifiers, or all the more by and large, for dimensionality decrease before ensuing arrangements. LDA has been firmly identified through analysis of variance (ANOVA) and regression analysis. In addition, efforts are made to convey a variable as a linear combination of different features or metrics [25][26].

3.5 Gradient Boosting Classifier (GBC)
GBC is a machine learning method for classification and regression problems. It gives a prescient model as a weak desire model (usually a decision tree). Like other upgrade methodologies, it fabricates models in an arranged plan and sums up them by permitting enhancements to any differentiable misfortune work. The possibility of gradient enhancement begins with the belief that enhancement can be interpreted as an optimization method with an appropriate cost function [27].

3.6 Decision Tree (DT)
DT taking in employments the decision tree from the view of things to the decision of the venture's objective worth. The tree model in which the objective variable can be organized in discrete qualities is known as a characterization tree. In these tree structures, leaves speak to class labels, and branches speak to a combination of segments that lead to these class names. Considering its understandability and straightforwardness, choice trees are perhaps the most prestigious AI advancements [28].

3.7 Principal Component Analysis (PCA)
PCA is a method for ascertaining principal components with utilizing them to perform information-based changes, sometimes using just the primary principal components and ignoring the respite. By anticipating information into a low-dimensional subspace, information can be extricated from a high-dimensional (include) space. It tries to protect the basic parts with more data diversity and eliminate unnecessary parts with less diversity [29].

3.8 Sequential Feature Selection (SFS)
Sequential feature selection is a bunch of greedy search calculations, which are utilized to lessen the essential d-dimensional part space to a k-dimensional component subspace with k<d along these lines, the heuristic driver determination estimation will choose a subset of features that are commonly material to the issue. The objectives of features assurance are two covering: we have to improve figuring control and lessen model speculation mistakes by taking out features. At whatever point feature determination is implanted, particularly valuable, (for example, sequential feature selection) are especially important [30].

4. Experimental Methodology
In this article, we utilize six AI algorithms, for example, RF, SVM, k-NN, LDA, GBC and DT. In this article, we utilize this information of Wisconsin’s (diagnostic) breast disease threat data file. We picked these six algorithms to locate the most appropriate technique with or without SFS to anticipate the pace of the sickness. A correlation has made in the middle of them. A cross approval system is applied to produce a preparation set for learning the classifier.
The yield from base learners and yield from diminished component SFS procedures at long last utilized for making examination on the dataset. Figure 1 indicates to the building model of the proposed work.

- In this step, all data will be preprocessed, and missing values will be replaced with the patterns and methods of the training data set.
- The informational index is partitioned into training set and test set.
- The train is divided into 10 parts.
- Fit the basic model to 9 parts and make predictions for the 10th part. It applies to all part of training set.
- At that point fit the essential model to the whole train informational set.
- Using this model, predictions can be made on the test set. Repeat steps iii to v for another basic model to obtain another set of predictions for the train group and the test group.
- The figures from the train set are used as highlights for building new models.
- This model is utilized to make last forecasts on the test expectation set, along these lines expanding the exactness rate.

5. Experimental Setup
All examination was performed on a PC prepared with a 2.4 GHz processor and 4 GB RAM utilizing the Python 3.6 stage. We utilize 10-fold cross-validation to apply the five ML calculation to every informational index to assess its presentation. We utilize the default boundaries in Python for all ML calculations.

5.1 Dataset
The standard breast disease Wisconsin (diagnostic) dataset from UCI machine learning repository [31] has been utilized for training and testing reason. The information base contains 569 instances and 32 features, and there is no missing value in the table. The portrayal of each features is given in the accompanying table 2.

| Attributes Name | Attribute Information | Graphical Presentation (Diagnosis) |
|-----------------|-----------------------|------------------------------------|

Table 2. Attribute Information
| ID number | Deleted |
|-----------|---------|
| diagnosis | malignant = 212  
benign = 357 |

Ascertained ten genuine features for every cell core

| radius | The average of the distance from the center to the outer point |
| texture | Standard deviation of gray values |
| perimeter | /// |
| area | /// |
| smoothness | nearby variety in lengths |
| compactness | perimeter$^2$/area - 1 |
| concavity | seriousness of curved bits of the contour |
| concave points | number of inward segments of the shape |
| symmetry | /// |
| fractal measurement | coastline approximation - 1 |

6. Results and Discussion
In order to evaluate the effectiveness of our model, we carried out experiments on breast cancer datasets. In order to summarize the data, the correlation matrix is used as input for more advanced analysis and as a diagnosis for advanced analysis. Figure 2 shows the correlation matrix of breast cancer dataset.

The principal component analysis performs on the informational collection to distinguish less significant factors [32]. For verifying that the factors are autonomous of one another, we have utilized PCA (as appeared in Figure 3) to explain the factors.
The significance of every classifier is estimated by accuracy [33]. The exactness is resolved dependent on the amount of the classifications of rights that are gainful and risky to without a doubt the quantity of cases, and is dictated by utilizing the accompanying conditions:
\[Accuracy = \frac{True\ Positive + True\ Negative}{Total}\] (1)

In Table 3 below, the test accuracy and training accuracy of the classifier and its SFS accuracy are listed.

| Classifiers               | Test_Accuracy | Train_Accuracy |
|--------------------------|---------------|----------------|
| RandomForestClassifier_sfs| 97.66         | 100.00         |
| KNeighborsClassifier_sfs | 97.66         | 97.24          |
| SVC_rbf                  | 97.66         | 98.24          |
| SVC_sfs                  | 97.66         | 97.49          |
| LinearDiscriminantAnalysis_ | 97.08         | 96.23          |
| LinearDiscriminantAnalysis_sfs | 95.91        | 96.23          |
| RandomForestClassifier_  | 95.91         | 99.75          |
| KNeighborsClassifier_    | 95.91         | 97.49          |
| SVC_linear               | 95.91         | 98.74          |
| GradientBoostingClassifier_ | 94.74         | 100.00         |
| DecisionTreeClassifier_sfs| 94.15         | 100.00         |
| RandomForestClassifier_FS| 93.57         | 100.00         |
| DecisionTreeClassifier_  | 92.98         | 100.00         |
| GradientBoostingClassifier_sfs | 92.98    | 99.75          |
| SVC_poly                 | 88.89         | 89.70          |

The statistical representation of Table 3 is shown in Figure 4.
7. Conclusion
Some information mining techniques are utilized for arrangement of classification of benign and malignant tumors. In this article, six supervised learning classification calculations and the consecutive segment investigation are utilized to anticipate breast malignancies and dissect them on different limits. The principle center here is which classifier has better anticipated precision. Thusly, the precision of RF_sfs, KNN_sfs, SVC_rbf, and SVC_sfs is the most noteworthy and equivalent, 97.66%, and the leftover exactness is lower. From the breast disease informational index above Table 2, every one of the 257 cases were benign and 212 cases were malignant. For all exploratory outcomes dependent on precision and different boundaries, we can presume that successive component determination is the most ideal approach to anticipate breast malignant growth.

8. References
[1] Zhao, J. and Wang, T., 2010, October. A general framework for medical data mining. In 2010 International Conference on Future Information Technology and Management Engineering (Vol. 2, pp. 163-165). IEEE.
[2] Shah, C. and Jivani, A.G., 2013, July. Comparison of data mining classification algorithms for
breast cancer prediction. In *2013 Fourth international conference on computing, communications and networking technologies (ICCCNT)* (pp. 1-4). IEEE.

[3] Chaurasia, V. and Pal, S., 2013. Early prediction of heart diseases using data mining techniques. *Caribbean Journal of Science and Technology*, 1, pp.208-217.

[4] Andrews, Leo John Baptist, and Linesh Raja. "A study on m-health inline with the sensors applying for a real time environment." *Journal of Statistics and Management Systems* 20.4 (2017): 659-667.

[5] Le Duff, F., Muntean, C., Cugia, M. and Mabo, P., 2004. Predicting survival causes after out of hospital cardiac arrest using data mining method. In *Medinfo* (pp. 1256-1259).

[6] Aha, D.W. and Bankert, R.L., 1996. A comparative evaluation of sequential feature selection algorithms. In *Learning from data* (pp. 199-206). Springer, New York, NY.

[7] Sarvestani, A.S., Safavi, A.A., Parandeh, N.M. and Salehi, M., 2010. October. Predicting breast cancer survivability using data mining techniques. In *2010 2nd International Conference on Software Technology and Engineering* (Vol. 2, pp. V2-227). IEEE.

[8] Lavanya, D. and Rani, K.U., 2012. Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services*, 2(1), p.17.

[9] Chaurasia, V. and Pal, S., 2020. Applications of machine learning techniques to predict diagnostic breast cancer. *SN Computer Science*, 1(5), pp.1-11.

[10] Chaurasia, V. and Pal, S., 2014. Data mining techniques: to predict and resolve breast cancer survivability. *International Journal of Computer Science and Mobile Computing IJCSMC*, 3(1), pp.10-22.

[11] Agarap, A.F.M., 2018, February. On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset. In *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing* (pp. 5-9).

[12] Bazazeh, D. and Shubair, R., 2016, December. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)* (pp. 1-4). IEEE.

[13] Al Bataineh, A. and Kaur, D., 2019. Optimal Convolutional Neural Network Architecture Design Using Clonal Selection Algorithm. *International Journal of Machine Learning and Computing*, 9(6).

[14] Bhardwaj, A. and Tiwari, A., 2013. Breast cancer diagnosis using genetically optimized neural network model. *Expert Systems with Applications*, 42(10), pp.4611-4620.

[15] Nguyen, C., Wang, Y. and Nguyen, H.N., 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic.

[16] Aruna, S., Rajagopalan, S.P. and Nandakishore, L.V., 2011. Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology*, 2(2011), pp.37-45.

[17] Parekh, V.S. and Jacobs, M.A., 2017. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ breast cancer*, 3(1), pp.1-9.

[18] Aggrawal, R. and Pal, S., 2020. Sequential Feature Selection and Machine Learning Algorithm-Based Patient’s Death Events Prediction and Diagnosis in Heart Disease. *SN Computer Science*, 1(6), pp.1-16.

[19] Ho, T.K., 1995, August. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.

[20] Ho, T.K., 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), pp.832-844.

[21] Friedman, J., Hastie, T. and Tibshirani, R., 2001. *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

[22] Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273-297.

[23] Piryonesi, S.M. and El-Diraby, T.E., 2020. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation*...
Engineering, Part B: Pavements, 146(2), p.04020022.

[24] Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[25] Fisher, R.A., 1950. The use of multiple measurements in taxonomic problems, Annual Eugenics, 7, Part II, 179-188 (1936); also in Contributions to Mathematical Statistics.

[26] McLachlan, G.J., 2004. Discriminant analysis and statistical pattern recognition (Vol. 544). John Wiley & Sons.

[27] Breiman, L., 1997. Arcing the edge (technical report). University of California, Berkeley, CA.

[28] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. and Zhou, Z.H., 2008. Top 10 algorithms in data mining. Knowledge and information systems, 14(1), pp.1-37.

[29] Barnett, T.P. and Preisendorfer, R., 1987. Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. Monthly Weather Review, 115(9), pp.1825-1850.

[30] Aha, D.W. and Bankert, R.L., 1996. A comparative evaluation of sequential feature selection algorithms. In Learning from data (pp. 199-206). Springer, New York, NY.

[31] UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/]. Irvine, CA: University of California, Center for Machine Learning and Intelligent Systems.

[32] Nakazato, Y., Sugiyama, T., Ohno, R., Shimoyama, H., Leung, D.L., Cohen, A.A., Kurane, R., Hirose, S., Watanabe, A. and Shimoyama, H., 2020. Estimation of homeostatic dysregulation and frailty using biomarker variability: a principal component analysis of hemodialysis patients. Scientific Reports, 10(1), pp.1-12.

[33] Chicco, D. and Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), p.6.