

Exploring the performance of feature selection method using breast cancer dataset

Tsehay Admassu Assegie¹, Ravulapalli Lakshmi Tulasi², Vadivel Elanangai³, Napa Komal Kumar⁴

¹Department of Computer Science, College of Natural & Computational Science, Injibara University, Injibara, Ethiopia

²Department of Computer Science and Engineering, R.V.R, College of Engineering, Guntur, India

³Department of Electrical and Electronics Engineering, St. Peter's Institute of Higher Education and Research, Avadi, India

⁴Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Avadi, India

Article Info

Article history:

Received Apr 27, 2021

Revised Oct 25, 2021

Accepted Nov 19, 2021

Keywords:

Breast cancer

Breast cancer detection

Breast cancer prediction

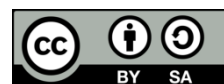
Feature selection

Sequnetial feature selection

ABSTRACT

Breast cancer is the most common type of cancer occurring mostly in females. In recent years, many researchers have devoted to automate diagnosis of breast cancer by developing different machine learning model. However, the quality and quantity of feature in breast cancer diagnostic dataset have significant effect on the accuracy and efficiency of predictive model. Feature selection is effective method for reducing the dimensionality and improving the accuracy of predictive model. The use of feature selection is to determine feature required for training model and to remove irrelevant and duplicate feature. Duplicate feature is a feature that is highly correlated to another feature. The objective of this study is to conduct experimental research on three different feature selection methods for breast cancer prediction. Sequential, embedded and chi-square feature selection are implemented using breast cancer diagnostic dataset. The study compares the performance of sequential embedded and chi-square feature selection on test set. The experimental result evidently shows that sequential feature selection outperforms as compared to chi-square (X^2) statistics and embedded feature selection. Overall, sequential feature selection achieves better accuracy of 98.3% as compared to chi-square (X^2) statistics and embedded feature selection.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tsehay Admassu Assegie

Department of Computer Science, College of Natural and Computational Science, Injibara University

Injibara, Ethiopia

Email: tsehayadmassu2006@mail.com

1. INTRODUCTION

Breast cancer is the most common cause of death among women throughout the global population [1], [2]. Breast cancer causes the second prevalent number of deaths in women [3]. Thus, early prediction of breast cancer is vital to reduce mortality caused by breast cancer. Despite the advances in mammography screening systems for early prediction of breast cancer, interpretation of X-rays and limited number of experienced oncologist in developing nations such as Ethiopia, high variability of experts' knowledge on breast cancer prediction makes breast cancer prediction more complicated. The decision making process during breast cancer prediction needs high accuracy as the outcome is highly risky because false positives leads to anxiety and false negatives leads to complications and patient suffers due to lack of treatment due to false negative outcome.

Redundant and duplicate input feature in the Wisconsin's breast cancer diagnostic dataset increases the computational time required for training and testing predictive model for breast cancer detection [4].

Furthermore, very large input feature increases the volume of dataset and larger dataset requires higher storage space. High correlation between features also affects the performance of model on breast cancer prediction [5].

The objective of feature selection is to extract representative feature for describing each of the dataset observation [6], [7]. In addition, feature selection reduces the number of dataset feature required to describe an observation in dataset. Hence, feature selection essentially reduces the number of input feature required to train a model. The reduction of the number of input feature in dataset decreases the computational time for training and testing model [8]. Hence, feature selection helps in developing more effective and faster model. Different researchers have proposed different types of feature selection methods. Feature selection methods are classified into three groups [9], [10] namely filter, embedded and wrapper method. The performance of feature selection methods varies across different datasets. The objective of this research is to investigate the performance of sequential, embedded and chi-square for breast cancer prediction. Overall, this research aims to investigate the answers to the following research questions:

- What is the performance of sequential feature selection for breast cancer prediction?
- What is the performance of embedded feature selection for breast cancer prediction?
- What is the performance of chi-square for breast cancer prediction?
- How to improve the performance of random forest model for breast cancer prediction?

The rest of this research is organized as follows: In section 2, the state of the art is presented, section 3 discusses the methodology, section 4 presents the result and discusses the comparative results and section 5 finally concludes the research.

2. LITERATURE SURVEY

Different researches have been carried out to solve the problem of breast cancer prediction using automated machine learning model and there have been different automated model for breast cancer prediction. However, breast cancer prediction still needs to be studied as the performance of the existing model have larger scope for improvement. Zhang *et al.* [11], the researchers developed decision tree based model for breast cancer prediction, which achieved an accuracy of 97.48%. The experiment on various feature subset evidently shows that feature selection is important to obtain good result on breast cancer prediction using decision tree model.

Feature selection is important to enhance the classification accuracy of the predictive accuracy of a model for breast cancer prediction. Assegie *et al.* [12], the researchers have suggested that, the performance of decision tree, adaptive boosting model greatly improves when the model is trained on optimal input feature. Moreover, optimal feature selection is significant to get insights into dataset and discover important feature from breast cancer dataset. The experimental result reveals that accuracy of the developed model is 92.53%.

Automated predictive model is proved important for breast cancer prediction at early stage and increases survival rate of breast cancer patient. Automated model is more accurate than inexperienced human experts or oncologist for breast cancer prediction [13]. The authors developed convolutional neural network based predictive model for breast cancer detection with predictive accuracy of 97%. In addition to accuracy, automated breast cancer prediction model avoids human errors, time and cost incurred for breast cancer identification [14]. Moreover, automated breast cancer prediction model avoids extra overload on oncologist especially where the number of breast cancer patient is higher.

In recent years, automated intelligent breast cancer prediction system is implemented with different supervised learning algorithms such as, k-nearest neighbor (KNN) and artificial neural network (ANN) [15]. However, the performance of the developed model still has scope for improvement for more accurate breast cancer prediction. Thus, we are motivated to study the existing work and propose more accurate model for breast cancer prediction by employing different feature selection methods, such as chi-square, sequential and embedded feature selection method.

3. RESEARCH METHODOLOGY

The dataset for this study is obtained from Wisconsin's breast cancer diagnostic dataset collected from Kaggle data repository. To evaluate the developed model, we have employed accuracy (number of correct predictions) with 5-fold cross validation. Chi-square, sequential feature selection and model based or embedded feature selection using random forest model is evaluated on breast cancer dataset. We have trained random forest model on original 30 input features representing 569 observations in the breast cancer dataset. Then the model is trained on optimal feature selected by chi-squared sequential feature selection and model based feature selection and accuracy is compared on model trained on original features before applying the feature selection.

3.1. Chi-square (X^2) statistic

Chi-square is statistical method for feature selection. Chi-square is a typical example for filter based feature selection [16]. Chi-square compares two input features and examines if they are related. Mathematically, chi-square is defined as (1).

$$X^2 = \frac{\sum(O_i - E_i)^2}{E_i} \quad (1)$$

Where O denotes observed value and E denotes expected value. The summation symbol shows that the calculation is performed on every input feature in dataset. Chi-square test is shows relationship between two variables in dataset [17]. Lower value of chi-square shows high correlation between input feature and target class or variable.

3.2. Sequential feature selection

Sequential feature selection (SFS) method is wrapper feature selection [18], [19]. Sequential feature selection selects the last feature or the first feature in the dataset initially. Then one of input feature from the remaining input feature is selected randomly and the model performance is compared. The process is repeated for all input features and the corresponding accuracy for each input feature subset is calculated. The input subset that produces highest accuracy is considered as optimal input feature.

Sequential feature selection is used to reduce an original N-dimensional input feature sub set to a d-dimensional feature set for d.

Initialize: Subset = 0, M = 0

for $i < d, i = i + 1$

 Compute model performance on: Subset = $f_{\delta_{ub}}[i]$

 compute $m = m + 1$

$f_{\delta_{ub}} = f_{\delta_{ub}} + 1$

stop when $i = m$

3.3. Embedded feature selection method

Tree based model such as decision tree and random forests (ensemble of trees) are used for feature selection [20]. Decision tree and random forest model is used to calculate feature importance when developing a model for determining the best feature and leaving unsuitable feature, with lower feature importance score [21], [22]. Random forest is an ensemble model, used as an embedded feature selection method, where each decision tree model in the ensemble is implemented by using observations of data from the complete dataset.

3.4. Performance metric

Accuracy is the most widely employed performance measure for validating the predictive performance of classification model [23]-[25]. Hence, in this study we have employed the predictive accuracy to evaluate the performance of the developed model. Mathematically, classification accuracy is defined as the number of correct predictions (true positives TP and true negatives TN) over all samples in the validation set N.

$$\text{Accuracy} = \frac{(TP+TN)}{N} \quad (2)$$

3.5. Dataset features

The original breast cancer dataset consists of 33 features. The ranking of the original 33 breast cancer diagnostic feature is demonstrated in Figure 1. Figure 1 worst concave points, worst area, mean concave points, mean concavity and worst radius has higher importance for breast cancer prediction. Overall, each feature of the original breast cancer dataset describing each sample in the dataset are demonstrated in Figure 1.

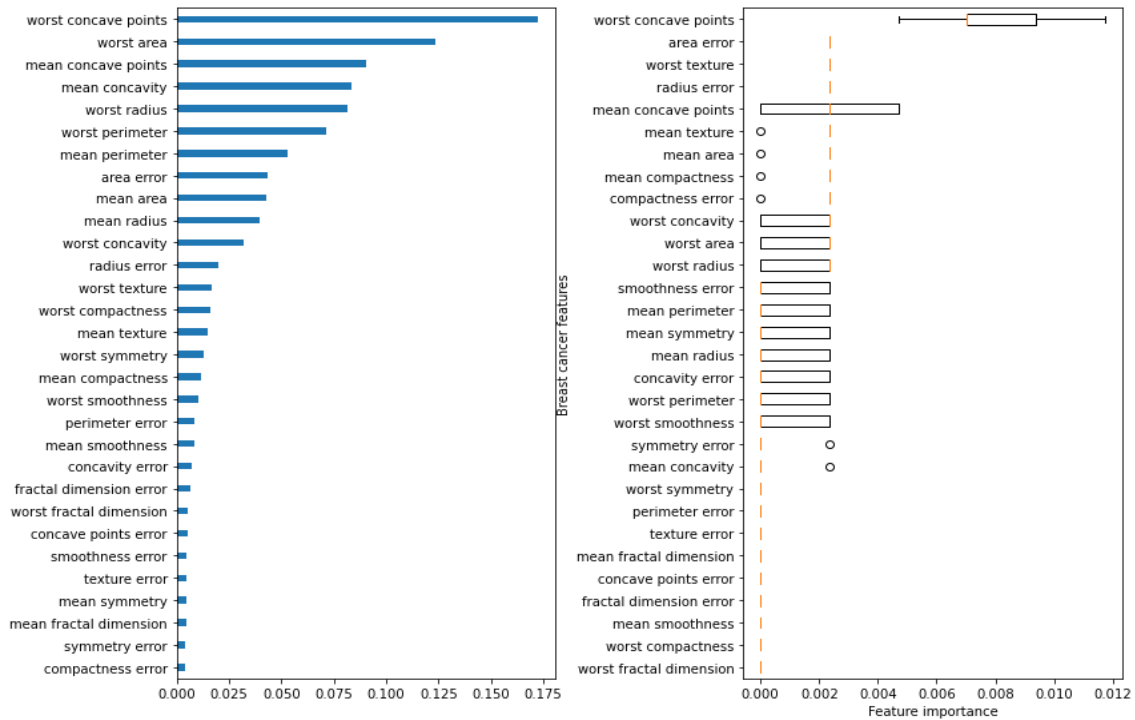


Figure 1. Breast cancer dataset features and their importance

4. RESULT AND DICUSSIONS

In this section, the experimental result on the performance of different feature selection method is presented. Specifically, the optimal feature selected by chi-square, sequential and model based or embedded feature selection with random forest algorithm is presented. The performance of feature selection model is evaluated against the predictive accuracy achieved when particularly feature selected by the feature selection method is used for training random forest model. The experimental result evidently appears to prove that the number of features selected by each of the feature selection method is different.

4.1. Comparison on the performance of feature selection method

Random forest model is trained on selected input feature subset by chi-square, sequential and feature importance and five-fold cross validation accuracy is employed to compare the model performance on each of the input feature subset. The comparative result on the performance of sequential embedded and chi-square feature selection is summarized in Table 1. As shown in Table 1, the highest accuracy (98.3%) is achieved with the feature subset selected sequential feature selection method as compared to chi-square with accuracy (95.7%) and embedded or model based feature importance with breast cancer detection accuracy of 96.3%.

Table 1. The performance of different feature selection method for breast cancer prediction

Feature selection method	No. input features selected	Accuracy
Chi square	8	95.78%
Sequential	8	98.3%
Embedded	8	96.30%

Different feature selection methods such as chi-square (X^2), sequential feature and embedded method selects different set of input features as optimal feature. Thus, the performance of a base classifier is different for different feature selection methods. Overall, all of the feature selection methods a better and more accuracy on breast cancer detection as compared to model trained on the original input feature. Sequential feature selection is better method to achieve better performance. The ranking methods such as embedded feature selection is good compared to statistical method such as chi-square statistics. We observe from Table 1 that the five-fold cross validation accuracy of the proposed random forest model performs better on breast cancer detection with feature subset selected using wrapper based sequential feature selection method as compared to chi-square statistical method and the embedded method.

The performance of sequential, chi-square and embedded feature selection method for breast cancer prediction is demonstrated in Figure 2. To compare the performance of the feature selection methods, we employed predictive accuracy as performance measure. We observe in Figure 2 that, sequential feature selection method outperforms as compared to chi-square and embedded method. Embedded feature selection performed better as compared to chi-square statistic being the least performing feature selection method.



Figure 2. Performance of feature selection methods for breast cancer prediction

5. CONCLUSION

In this study, we have explored the performance of embedded, chi-square and sequential feature selection by employing breast cancer dataset. The original breast cancer dataset includes 33 features. However, after feature selection, we drop this number from 33 to 8 with accuracy 98.3% using sequential feature selection method. The experimental result evidently shows that the accuracy is different for different feature selection methods for embedded and chi-square feature selection the accuracy is 95.78% and 96.30% respectively.

ACKNOWLEDGEMENTS

This work is partially supported by Inijbara University. The authors are thankful to Inijbara University for providing laboratory equipment, laptop computer for conducting this research.





REFERENCES

- [1] T. A. Assegie, "An optimized K-Nearest Neighbor based breast cancer detection," *Journal of Robotics and Control*, vol. 2, no. 3, pp. 115-118, May 2020, doi: 10.18196/jrc.2363.
- [2] R. A. I. Alhayali, M. A. Ahmed, Y. M. Mohialden, and A. H. Ali, "Efficient method for breast cancer classification based on ensemble hoeffding tree and naïve Bayes," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1074-1080, May 2020, doi: 10.11591/ijeecs.v18.i2.pp1074-1080.
- [3] H. Dhahri, E. A. Maghayreh, A. Mahmood, and W. Elkilani, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms," *Hindawi Journal of Healthcare Engineering*, vol. 2019, pp. 1-11, doi: 10.1155/2019/4253641.
- [4] Z. Uyu and L. Choridah, "Feature Selection Mammogram based on Breast Cancer Mining," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 1, pp. 60-69, February 2018, doi: 10.11591/ijece.v8i1.pp60-69.
- [5] T. S. Lim, K. G. Tay, A. Huong, and X. Y. Lim, "Breast cancer diagnosis system using hybrid support vector machine-artificial neural network," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3059-3069, August 2021, doi: 10.11591/ijece.v11i4.pp3059-3069.
- [6] Y. Guoa, B. Zhanga, Y. Sunb, K. Jiang, and K. Wu, "Machine learning based feature selection and knowledge reasoning for CBR system under big data," *Pattern Recognition*, vol. 112, 2021, doi: 10.1016/j.patcog.2020.107805.
- [7] N. Maleki, Y. Zeinali, and T. A. Seyed, "A K-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Systems with Applications*, vol. 164, 2021, doi: 10.1016/j.eswa.2020.113981.
- [8] S. Punitha, F. Al-Turjman, and Thompson, "An automated breast cancer diagnosis using feature selection and parameter optimization in ANN," *Computers and Electrical Engineering*, vol. 90, 2021, doi: 10.1016/j.compeleceng.2020.106958.
- [9] K. Zhu and J. Yang, "A cluster-based sequential feature selection algorithm," *2013 Ninth International Conference on Natural Computation*, 2013, doi: 10.1109/ICNC.2013.6818094.
- [10] L. Wang, C. Shen, and H. Richard, "On the Optimal of Sequential Forward Feature Selection Using Class Separability Measure," *International Conference on Digital Image Computing: Techniques and Applications*, 2021, doi: 10.1109/DICTA.2011.41.
- [11] J. Zhang, L. Chen, and F. Abid, "Prediction of Breast Cancer from Imbalance Respect Using Cluster-Based Undersampling Method," *Hindawi Journal of Healthcare Engineering*, vol. 2019, pp. 10, doi: 10.1155/2019/7294582.
- [12] T. A. Assegie, R. L. Tulasi, and N. K. Kumar, "Breast cancer prediction model with decision tree and adaptive boosting," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 1, pp. 184-190, 2021, doi: 10.11591/ijai.v10.i1.pp184-190.184.
- [13] S. A. Alanazi et al., "Boosting Breast Cancer Detection Using Convolutional Neural Network," *Hindawi Journal of Healthcare Engineering*, vol. 2021, pp. 1-11, doi: 10.1155/2021/5528622.




- [14] T. A. Assegie and P. S. Nair, "The Performance of Different Machine Learning Models on Diabetes Prediction," *International Journal of Scientific & Technology Research*, vol. 9, no. 01, pp. 2491-2494, January 2020. [Online]. Available at: <https://www.ijstr.org/final-print/jan2020/The-Performance-Of-Different-Machine-Learning-Models-On-Diabetes-Prediction-.pdf>
- [15] Y. A. Mohammed and E. G. Saleh, "Comparative study of logistic regression and artificial neural networks on predicting breast cancer cytology," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 2, 2021, pp. 1113-1120, doi: 10.11591/ijeecs.v21.i2.pp1113-1120.
- [16] A. Khamparia, S. Bharati, P. Podder, D. Gupta, A. Khanna, T. K. Phung, and H. Thanh, "Diagnosis of breast cancer based on modern mammography using hybrid transfer learning," *Multidimensional Systems and Signal Processing*, vol. 32, pp. 747-765, 2021, doi: 10.1007/s11045-020-00756-7.
- [17] R. R. Janghel, A. Shukla, R. Tiwari, and R. Kala, "Intelligent Decision Support System for Breast Cancer," *Soft Computing and Expert System Laboratory*, pp. 351-358, 2010, doi: 10.1007/978-3-642-13498-2_46.
- [18] Z. Rustam, Y. Amalia, S. Hartini, and G. S. Saragih, "Linear discriminant analysis and support vector machines for classifying breast cancer," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 1, pp. 253-256, March 2021, doi: 10.11591/ijai.v10.i1.pp253-256.
- [19] M. S. essiane *et al.*, "Feature selection based on dialectics to support breast cancer diagnosis using thermographic images," *Research on Biomedical Engineering*, pp. 1-22, 2021, doi: 10.1007/s42600-021-00158-z.
- [20] A. Ridok, N. Widodo, W. F. Mahmudy, and M. Rifa, "A hybrid feature selection on AIRS method for identifying breast cancer diseases," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, pp. 728-735, February 2021, doi: 10.11591/ijece.v11i1.pp728-735.
- [21] M. Mahmood, B. Al-Khateeb, and W. M. Alwash, "A review on neural networks approach on classifying cancers.," *IAES International Journal of Artificial Intelligence*, vol. 9, no. 2, pp. 317-326, June 2020, doi: 10.11591/ijai.v9.i2.pp317-326.
- [22] W. N. Ibeni, M. Z. Salikon, A. Mustapha, S. A. Daud, and M. N. Salleh, "Comparative analysis on Bayesian classification for breast cancer problem," *Bulletin of Electrical Engineering and Informatics*, vol. 8, no. 4, pp. 1303-1311, December 2019, doi: 10.11591/eei.v8i4.1628.
- [23] Y. A. Mohammed and E. Saleh, "An enhancement of mammogram images for breast cancer classification using artificial neural networks," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 332-345, 2021, doi: 10.11591/ijai.v10.i2.pp332-345.
- [24] S. Bagchi, K. G Tay, A. Huong, dan S. K. Debnath, "Image processing and machine learning techniques used in computer-aided detection system for mammogram screening-A review," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, pp. 2336-2348, June 2020, doi: 10.11591/ijece.v10i3.pp2336-2348.
- [25] G. Saranya and A. Pravin, "A comprehensive study on disease risk predictions in machine learning," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 4217-4225, August 2020, doi: 10.11591/ijece.v10i4.pp4217-4225.

BIOGRAPHIES OF AUTHORS






Tsehay Admassu Assegie     is Lecturer at College of Natural & Computational Science, Injibara University, Ethiopia. He Holds a M.Sc., degree in Computer Science. His research areas are machine learning, medical image analysis and pattern recognition. He has published over 26 research articles in referred and Scopus indexed international journals. He can be contacted at email: tsehayadmassu@inu.edu.et.






Dr. Ravulapalli Lakshmi Tulasi    is currently working as a Professor in the Department of Computer Science and Engineering, R.V.R & J.C College of Engineering, Guntur, Andhra Pradesh, India. Her research interests include Machine Learning, Data Mining, Information Retrieval Systems, and Semantic Web. She can be contacted at email: rtulasi.2002@gmail.com.



Vadivel Elanagai    is currently working as Assistant Professor in the Department of Electrical and Electronics Engineering at St. Peter's Institute of Higher Education and Research, AVADI, Chennai. She has 11 years of Teaching Experience. She is currently doing her research in Image Processing. Her current research interest includes Image Processing, VLSI Design, Fuzzy logic, Artificial Neural Network. She has also published research papers in reputed journals and conference proceedings. She can be contacted at email: elanagai123@gmail.com.



Napa Komal Kumar    is currently working as Assistant Professor in the Department of Computer Science and Engineering at St. Peter's Institute of Higher Education and Research, Avadi, Chennai. His research interests include Machine Learning, Data Mining, and Cloud Computing. He can be contacted at email: komalkumarna@gmail.com.