Journal of Soft Computing Exploration

Optimization of breast cancer classification using feature selection on neural network

Jumanto[1], M Fadil Mardiansyah[2], Rizka Pratama[3], M Faris Al Hakim[4], Bibek Rawat[5]
[1,2,3,4]Department of Computer Science, Universitas Negeri Semarang, Indonesia
[5]Department of Computer Science, Chandigarh University, India

Article Info

Article history:

Received Sep 7, 2022
Revised Sep 16, 2022
Accepted Sep 29, 2022

Keywords:

Cancer
Breast cancer
Classification
Backpropagation
Neural network

ABSTRACT

Cancer is currently one of the leading causes of death worldwide. One of the most common cancers, especially among women, is breast cancer. There is a major problem for cancer experts in accurately predicting the survival of cancer patients. The presence of machine learning to further study it has attracted a lot of attention in the hope of obtaining accurate results, but its modeling methods and predictive performance remain controversial. Some Methods of machine learning that are widely used to overcome this case of breast cancer prediction are Backpropagation. Backpropagation has an advantage over other Neural Networks, namely Backpropagation using supervised training. The weakness of Backpropagation is that it handles classification with high-dimensional datasets so that the accuracy is low. This study aims to build a classification system for detecting breasts using the Backpropagation method, by adding a method of forward selection for feature selection from the many features that exist in the breast cancer dataset, because not all features can be used in the classification process. The results of combining the Backpropagation method and the method of forward selection can increase the detection accuracy of breast cancer patients by 98.3%.

Corresponding Author:

Jumanto,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: jumanto@mail.unnes.ac.id

1. INTRODUCTION

Cancer is currently one of the leading causes of death worldwide. One of the most common cancers, especially among women, is breast cancer. Breast cancer is the second leading cause of death due to cancer in women currently [1]. Breast cancer cases in women in developed countries are less than in developing countries, namely 794,000 cases, while in developing countries there are 833,000 cases. In developing countries, cancer is plaguing most people, and the 5-year survival rate of cancer is only 40.5% [2]. Breast cancer is cancer that forms in breast tissue. This cancer occurs when cells in the breast tissue grow uncontrollably and take over the healthy and surrounding breast tissue [3]. This heterogeneous disease is determined by molecular types and subtypes [4]. The hormonal risk for the development of BC is a well-proven fact, predominantly through the estrogen and progesterone receptors [5].

Recent advances in cancer research and diagnostics have largely depended on new developments in microscopic or molecular profiling techniques, offering a high level of detail concerning special molecular features [6]. There is a major problem for cancer experts in accurately predicting the survival of cancer patients. The presence of machine learning for further study has attracted a lot of attention in the hope of obtaining

accurate results, but the modeling method and its predictive performance remain controversial [7]. Predicting the recurrence of breast cancer with the use of machine learning algorithms allows doctors to access a large number of medical records regarding this cancer [8].

At this time the health sector has been supported by technology that is able to visualize and predict a patient's condition [9]. Several methods of machine learning are widely used to solve breast cancer prediction cases, especially classification algorithms including neural network [10], Naive Bayes [11][12]. The Naive Bayes model works on the Bayesian algorithm which classified the data by probability and statistical modeling [13]. Another method are Support Vector Machine (SVM) [14][15][16], and Backpropagation Neural Network [17]. Naive Bayes has the advantage of being simple but has high accuracy even though it uses not a lot of data, while SVM has the advantage of solving linear and non-linear classification and regression problems which can become a learning algorithm capability for regression and classification, but the Support Vector Machine (SVM) are not efficient in training large-capacity data [18]. Backpropagation has advantages over other Neural Networks, namely Backpropagation using supervised training [19].

Backpropagation is one of the methods of Artificial Neural Networks, which is capable of solving a problem. Backpropagation architecture is one of several ANN architectures that can be used to study and analyze past data patterns more precisely to obtain a more accurate output [20]. This method has been widely and successfully implemented in various applications, such as performance evaluation, location determination, and pattern recognition. The implementation of the Backpropagation algorithm goes through 2 processes, namely the process training and testing [4].

The majority of classification algorithms have weaknesses in handling classification with datasets that have class imbalance [21]. Several studies also show that researchers often do not pay attention to the distribution balance in the dataset class which causes difficulties in the classification algorithm used [22]. High-dimensional data is one of the obstacles in the application of machine learning and data mining techniques because it will have a negative effect on the analysis process [23]. One effort to reduce the features of high-dimensional data is to use feature selection. There are several feature selection methods that are widely recommended by world researchers, one of which is Forward Feature Selection [24].

Based on this background, this research was conducted to create a program that is able to optimize the classification of breast cancer using the Forward Feature Selection-Backpropagation ANN method. Where the Forward Feature Selection method is used to reduce data features by selecting several features to be used, so that the classification becomes optimum or gets a high accuracy value.

2. METHOD

The research conducted is designed coherently. To obtain effective research results. The steps for this research are a series of structured and planned processes to effectively achieve the aims of this research. The following steps were carried out for this research as shown in Figure 1.

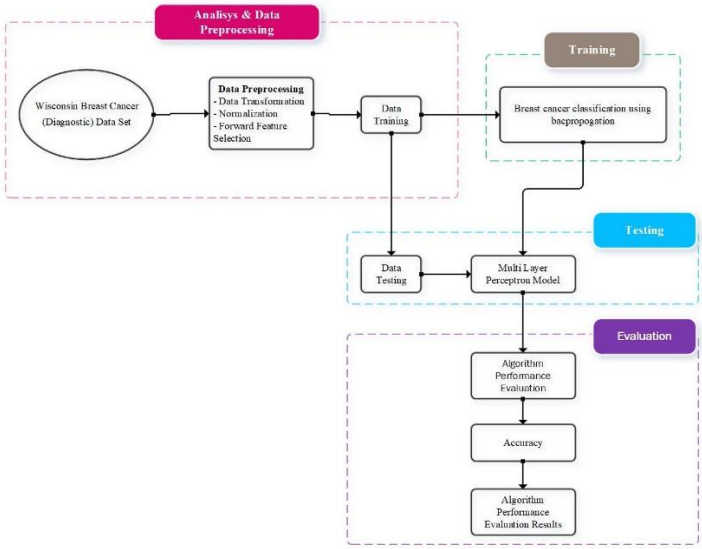

Figure 1. Research framework

The first stage in this research is data collection. The data used is the Wisconsin Breast Cancer (Diagnostic) Data Set. Data obtained from UCI Machine Learning Repository which has 569 cases, 2 classes (Malignant and Benign). The number of Benign classes is 357 and Malignant is 212. The next stage is the preprocessing stage. The preprocessing stages used in this research are data transformation, data normalization, and Forward Feature Selection. Data transformation is done using Label Encoder to convert categorical data into numerical data. Furthermore, data normalization was performed using the Min-Max Scaler method. This study divides the data into training data and testing data with a ratio of 80:20.

Furthermore, Forward Feature Selection is carried out using the Random Forest algorithm. Random Forest is done by merging trees (trees) by conducting training on the sample data owned. The Random Forest Algorithm is shown in Figure 2.

Algorithm 1: Pseudo code for the random forest algorithm

To generate c classifiers:
for i = 1 to c **do**
  Randomly sample the training data D with replacement to produce D_i
  Create a root node, N_i containing D_i
  Call BuildTree(N_i)
end for

BuildTree(N):
if N contains instances of only one class **then**
  return
else
  Randomly select x% of the possible splitting features in N
  Select the feature F with the highest information gain to split on
  Create f child nodes of N, N_1,..., N_f, where F has f possible values (F_1, … , F_f)
  for i = 1 to f **do**
    Set the contents of N_i to D_i, where D_i is all instances in N that match F_i
    Call BuildTree(N_i)
  end for
end if

Figure 2. Random forest algorithm

The next step is to classify using Backpropagation ANN. Backpropagation is one of the methods of an artificial neural network which is a supervised learning training method with a multi-layer network and has a special feature of minimizing errors in the output generated by the network. This classifier works by performing two calculation stages, namely forward calculations which will calculate the error value between the system output value and the correct value, and backward calculation to correct the weight based on the error value. The Backpropagation algorithm can be shown in Figure 3.

Algorithm 1 Backpropagation Algorithm

1: **procedure** TRAIN
2: $X \leftarrow$ Training Data Set of size mxn
3: $y \leftarrow$ Labels for records in X
4: $w \leftarrow$ The weights for respective layers
5: $l \leftarrow$ The number of layers in the neural network, 1...L
6: $D_{ij}^{(l)} \leftarrow$ The error for all l,i,j
7: $t_{ij}^{(l)} \leftarrow 0$. For all l,i,j
8: *For* $i = 1$ to m
9: $a^l \leftarrow feedforward(x^{(i)}, w)$
10: $d^l \leftarrow a(L) - y(i)$
11: $t_{ij}^{(l)} \leftarrow t_{ij}^{(l)} + a_j^{(l)} \cdot t_i^{l+1}$
12: **if** $j \neq 0$ **then**
13: $D_{ij}^{(l)} \leftarrow \frac{1}{m} t_{ij}^{(l)} + \lambda w_{ij}^{(l)}$
14: **else**
15: $D_{ij}^{(l)} \leftarrow \frac{1}{m} t_{ij}^{(l)}$
16: where $\frac{\partial}{\partial w_{ij}^{(l)}} J(w) = D_{ij}^{(l)}$

Figure 3. Backpropagation algorithm ANN training

The next step is to classify using Backpropagation ANN. Backpropagation is one of the methods of an artificial neural network which is a Supervised Learning training method with a multi-layer network and has a special feature of minimizing errors in the output generated by the network. This classifier works by performing two calculation stages, namely forward calculations which will calculate the error value between the system output value and the correct value and backward calculation to correct the weight based on the error value.

3. RESULTS AND DISCUSSIONS

This stage contains the results of the research stages based on the research framework. The stages of the research consist of collecting data from the UCI Machine Learning Repository. The next stage is data preprocessing with the Forward Feature Selection process and then classification using the Backpropagation ANN method. The last stage is performance evaluation based on the accuracy of the method applied using the Confusion Matrix.

The classification process is carried out after going through the data preprocessing stages. Classification is a process to determine an item from the dataset into a class label. The classification process is carried out using the Backpropagation ANN method on the dataset with the distribution of training data and testing data of 80:20. The dataset used in this study is the Wisconsin Breast Cancer (Diagnostic) dataset. After performing the classification process, the accuracy of the classification of the Wisconsin Breast Cancer (Diagnostic) dataset was obtained. The results are then carried out by the accuracy testing process with the confusion matrix method shown in Figure 4. Based on Figure 4 below, the results show an accuracy of 98.3%. And the Confusion Matrix is True Positive (TP) = 47, True Negative (TN) = 65, False Positive (FP) = 2, and False Negative (FN) = 0.

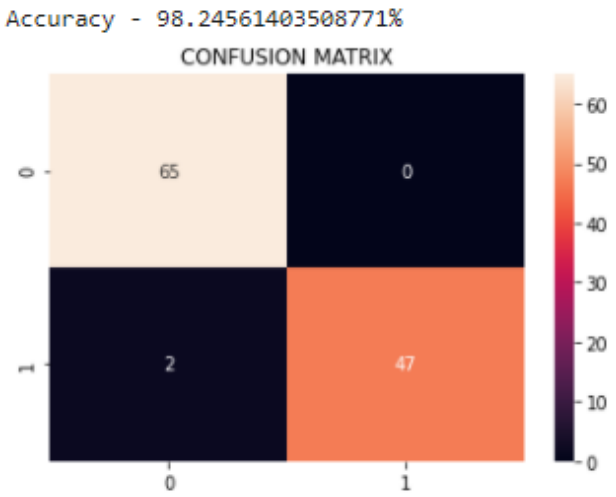

Figure 4. Accuracy and confusion matrix

4. CONCLUSION

Based on the testing results that have been carried out, it shows that the Forward Feature Selection-Backpropagation ANN method obtains an accuracy of 98.3%. Although the accuracy resulting from this study is lower than the previous study using Neural Networks with an accuracy of 99.20%, but the use of the Forward Feature Selection-Backpropagation ANN method in this study can be used to optimize the accuracy of breast cancer classification because it has best accuracy.

REFERENCES

[1] R. H. Saputra and B. Prasetyo, "Improve the Accuracy of C4.5 Algorithm Using Particle Swarm Optimization (PSO) Feature Selection and Bagging Technique in Breast Cancer Diagnosis," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 47–55, 2020, https://doi.org/10.52465/joscex.v1i1.9.

[2] G. Yu, Z. Chen, J. Wu, and Y. Tan, "A diagnostic prediction framework on auxiliary medical system for breast cancer in developing countries," *Knowledge-Based Syst.*, vol. 232, p. 107459, 2021, doi: 10.1016/j.knosys.2021.107459.

[3] M. Lilleborge, R. S. Falk, T. Sørlie, G. Ursin, and S. Hofvind, "Can breast cancer be stopped? Modifiable risk factors of breast cancer among women with a prior benign or premalignant lesion," *Int. J. Cancer*, vol. 149, no. 6, pp. 1247–1256, 2021, doi: 10.1002/ijc.33680.

[4] B. Prasetiyo, Alamsyah, M. A. Muslim, Subhan, and N. Baroroh, "Artificial neural network model for banckrupty prediction," *J. Phys. Conf. Ser.*, vol. 1567, no. 3, pp. 8–12, 2020, doi: 10.1088/1742-6596/1567/3/032022.

[5] R. B. Dickson and G. M. Stancel, "Estrogen receptor-mediated processes in normal and cancer cells.," *J. Natl. Cancer Inst. Monogr.*, no. 27, pp. 135–145, 2000, doi: 10.1093/oxfordjournals.jncimonographs.a024237.

[6] A. Binder *et al.*, "Morphological and molecular breast cancer profiling through explainable machine learning," *Nat. Mach. Intell.*, vol. 3, no. 4, pp. 355–366, 2021, doi: 10.1038/s42256-021-00303-4.

[7] J. Li *et al.*, "Predicting breast cancer 5-year survival using machine learning: A systematic review," *PLoS One*, vol. 16, no. 4 April, pp. 1–23, 2021, doi: 10.1371/journal.pone.0250370.

[8] H. Saleh, S. F. Abd-El Ghany, H. Alyami, and W. Alosaimi, "Predicting Breast Cancer Based on Optimized Deep Learning Approach," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/1820777.

[9] I. G. A. Suciningsih, M. A. Hidayat, and R. A. Hapsari, "Comparation analysis of naïve bayes and decision tree C4.5 for caesarean section prediction," *J. Soft Comput. Explor.*, vol. 2, no. 1, pp. 46–52, 2021, doi: 10.52465/joscex.v2i1.25.

[10] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3465–3469, 2009, doi: 10.1016/j.eswa.2008.02.064.

[11] N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052068.

[12] C. Agossou, M. N. Atchadé, A. Moussa Djibril, and S. V. Kurisheva, "Support Vector Machine, Naive Bayes Classification, and Mathematical Modeling for Public Health Decision-Making: A Case Study of Breast Cancer in Benin," *SN Comput. Sci.*, vol. 3, no. 2, pp. 1–19, 2022, doi: 10.1007/s42979-021-01008-6.

[13] M. Ibtasam, "Accuracy Measurements and Decision Making by NaÃ¯ve Bayes and Forward Chaining Method to Identify the Malnutrition Causes and Symptoms," *Sci. J. Informatics*, vol. 8, no. 2, pp. 320–324, 2021, doi: 10.15294/sji.v8i2.29317.

[14] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," *5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017*, vol. 2018-January, pp. 226–229, 2018, doi: 10.1109/R10-HTC.2017.8288944.

[15] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 3, pp. 293–299, 2019, doi: 10.1016/j.cegh.2018.10.003.

[16] N. Hidayat, M. F. Al Hakim, and J. Jumanto, "Halal Food Restaurant Classification Based on Restaurant Review in Indonesian Language Using Machine Learning," *Sci. J. Informatics*, vol. 8, no. 2, pp. 314–319, 2021, doi: 10.15294/sji.v8i2.33395.

[17] S. Senthil and B. Ayshwarya, "Lung Cancer Prediction using Feed Forward Back Propagation Neural Networks with Optimal Features," *Int. J. Appl. Eng. Res.*, vol. 13, no. 1, pp. 318–325, 2018, doi:10.37622/000000.

[18] R. Jayapermana, A. Aradea, and N. I. Kurniati, "Implementation of Stacking Ensemble Classifier for Multi-class Classification of COVID-19 Vaccines Topics on Twitter," *Sci. J. Informatics*, vol. 9, no. 1, pp. 8–15, 2022, doi: 10.15294/sji.v9i1.31648.

[19] D. I. Wijaya, M. K. Aulia, Jumanto, and M. F. Al Hakim, "Room occupancy classification using multilayer perceptron," *J. Soft Comput. Explor.*, vol. 2, no. 2, pp. 163–168, 2021, doi: https://doi.org/10.52465/joscex.v2i2.

[20] A. Agustyawan, T. G. Laksana, and U. Athiyah, "Combination of Backpropagation Neural Network and Particle Swarm Optimization for Water Production Prediction in Municipal Waterworks," *Sci. J. Informatics*, vol. 9, no. 1, pp. 84–94, 2022, doi: 10.15294/sji.v9i1.29849.

[21] R. S. Wahono, N. S. Herman, and S. Ahmad, "Neural network parameter optimization based on genetic algorithm for software defect prediction," *Adv. Sci. Lett.*, vol. 20, no. 10–12, pp. 1951–1955, 2014, doi: 10.1166/asl.2014.5641.

[22] N. Nikolaou, N. Edakunni, M. Kull, P. Flach, and G. Brown, "Cost-sensitive boosting algorithms: Do we really need them?," *Mach. Learn.*, vol. 104, no. 2–3, pp. 359–384, 2016, doi: 10.1007/s10994-016-5572-x.

[23] D. Guan, W. Yuan, Z. Jin, and S. Lee, "Undiagnosed samples aided rough set feature selection for medical data," *Proc. 2012 2nd IEEE Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2012*, pp. 639–644, 2012, doi: 10.1109/PDGC.2012.6449895.

[24] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation," *Environ. Model. Softw.*, vol. 101, pp. 1–9, 2018, doi: 10.1016/j.envsoft.2017.12.001.