TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 18, No. 2, April 2020, pp. 815~821
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v18i2.14785

Comparing random forest and support vector machines for breast cancer classification

Chelvian Aroef, Yuda Rivan, Zuherman Rustam
Department of Mathematics, University of Indonesia, Indonesia

Article Info

Article history:

Received Aug 31, 2019
Revised Jan 15, 2020
Accepted Feb 14, 2020

Keywords:

Breast cancer
Random forest
Support vector machines

ABSTRACT

There are more than 100 types of cancer around the world with different symptoms and difficulty in predicting its appearance in a person due to its random and sudden attack method. However, the appearance of cancer is generally marked by the growth of some abnormal cell. Someone might be diagnosed early and quickly treated, but the cancerous cell most times hides in the body of its victim and reappear, only to kill its sufferer. One of the most common cancers is breast cancer. According to Ministry of Health, in 2018, breast cancer attacked 42 out of every 100.000 people in Indonesia with approximately 17 deaths. In addition, the Ministry recorded a yearly increase in cancer patients. Therefore, there is adequate need to be able to determine those affected by this disease. This study applied the Boruta feature selection to determine the most important features in making a machine learning model. Furthermore, the Random Forest (RF) and Support Vector Machines (SVM) were the machine learning model used, with highest accuracies of 90% and 95% respectively. From the results obtained, the SVM is a better model than random forest in terms of accuracy.

Corresponding Author:

Zuherman Rustam,
Department of Mathematics,
University of Indonesia, Jl. Margonda Raya, Pondok Cina,
Kecamatan Beji, Kota Depok, Jawa Barat 16424, Indonesia.
Email: rustam@ui.ac.id

1. INTRODUCTION

Cancer is one of the deadliest diseases in the world. According to World Health Organization (WHO), in 2018, it is the second leading cause of death globally and responsible for approximately 9.6 billion deaths in 2018. There are over 100 different type of cancer that affect human. However, this study, aims to analyze the breast cancer, a disease in which cells grow out of control to form a tumor which tends to affect another part of the body. There are three common parts of breast whose cells has the ability to turn into cancer namely lobules, ducts, and the connective tissue. Lobules are glands which produces milk while the ducts are thin tubes that carry milk away from the lobule. The connective tissue consists of fibrous and fatty tissues which holds the breast and gives it shape and size. However, in most cases, it begins in the lobules or ducts.

The exact causes of breast cancer are still not known, but experts are of the opinion that an interaction between genes with lifestyle, environment, and hormone, tends to provoke abnormal cell growth. There are several factors that increase the risk of getting breast cancer such as age. According to research, most cases people are diagnosed after the age of 50. Men still have a risk of getting breast cancer even though it is a lot lower than women. Someone who had early menstrual periods before the age of 12 and starting menopause after 55, stand a higher risk of being affected. Radiation therapy is also another factor which makes

cell grows abnormally. Furthermore, women have a higher risk of getting breast cancer assuming her first relative (mother, daughter, or sister) was diagnosed with it, which in most cases is unchangeable. There are also some factors for instance overweight women after menopause stand a higher risk than those with normal body weight. Care should therefore be taken by those with increased hormones after menopause, as it raises the risk of getting affected by breast cancer. When someone has all the above mentioned factors, doesn't mean they are sufferers, and vice versa.

Symptoms of breast cancer differ from persons. However, some common symptoms include skin changes, such as swelling, redness, visible differences in one or both breasts, appearance of a lump which doesn't go away after some period, feeling of pain or burning sensation around the breast area even with no pressure, a change in the nipple, itches, etc. Once you come across any of these symptoms, consult a doctor immediately.

The treatment for breast cancer is different and dependent on the type, the tumor size, and how far it has spread in the body (stage of the cancer). The most common treatment method is surgery, which is used to remove the tumor and tissues known as lumpectomy or the whole breast called mastectomy. In addition, once the cancer has already spread in the body, the common treatment is radiation therapy, the intention is to kill its cells using high energy waves. The other way to kill cancer cells is Chemotherapy, which is the use of drugs, however, this treatment also has its side effect such as hair loss, early menopause, and fatigue. The use of medicine to prevent hormones, especially estrogen, also works as a treatment. But sadly, currently there is no cure for cancer completely. Therefore, the sooner the better to know someone is suffering cancer or not, so it can be treated early.

Many machine learning methods have been applied for breast cancer classification, such as Support Vector Machines [1] and Network-based [2]. However, this research compares both in terms of accuracy. SVM is already a widely known method used for classification such acute sinusitis [3], face identification [4], predicting bank failure [5], Intrusion Detection System [6, 7, 8], Classification of Schizophrenia [9], Detection of Traffic Incident [10], and Face Recognition [11, 12]. Some previous studies utilized random forest for gene selection and classification [13], classification of android malware [14], predict bank failure [15], predict prostate cancer [16], and osteoarthritis classification [17]. This research is expected to help the health sector to classify breast cancer sufferers.

2. RESEARCH METHOD

2.1. Data

The data in this study was taken from UCI machine learning repository [18]. The data consists of nine features, as follows:

- Age (Years)
- BMI (kg/m²)
- Glucose (mg/dL)
- Insulin (µU/mL)
- HOMA
- Leptin (ng/mL)
- Adiponectin (µg/mL)
- Resistin (ng/mL)
- MCP 1 (pg/mL)

There are two classification class, with 116 observations consisting of 52 healthy (1) and 64 patients (2).

Table 1. Data of breast cancer from UCI machine learning repository

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP 1	Class
76	27.1	110	26.2	7.11	21.78	4.93	8.49	45.84	1
77	25.9	85	4.58	0.96	13.74	9.75	11.77	488.82	1
45	21.3	102	13.852	3.48	7.64	21.05	23.03	552.44	2
45	20.83	74	4.56	0.83	7.75	8.23	28.03	382.95	2

2.2. Supervised learning

Supervised learning is a method that provides discrete prediction, called classification. It split the data into training, which is used to predict model to obtain the best parameter and the test data where the obtained

results are applied. Supervised learning keeps updating itself to make the best model possible, and by using it, a new data has the ability of being inputted and classified.

2.3. Decision tree

Decision tree is a model diagram that consists of node and edge. There are several types of nodes such as, root, parent (internal), and child nodes (leaf). Root node is the beginning of a node that makes another branch, known as the parent node. This makes another branch known as child node, which consists of right and left nodes. Furthermore, when the child node doesn't have any branch, it's called terminal node. Figure 1 shows a simple decision tree consists of root node, internal node and leaf node.

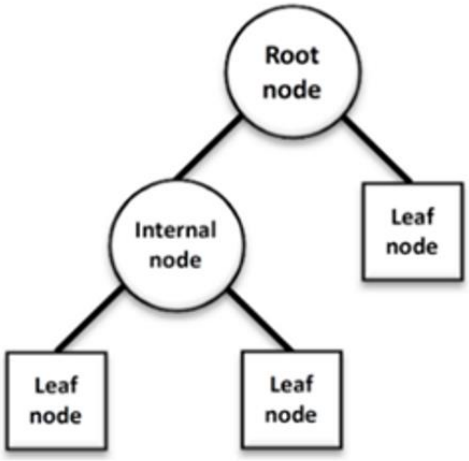

Figure 1. Diagram of decision tree consists of root node, internal node, and leaf node

Decision tree is one of the examples of machine learning model, which is easy to understand because it is visualized. It follows the rule of Boolean algebra. Tree is made of binary recursive process of the whole data so the variable from each data will be homogenous [19]. Making a decision tree model involve these processes [19]:

- Split the parent node into child node with goodness of split criterion
- State the terminal node which is the last node of the decision three
- Class assignment

Algorithm for making the decision tree model is using divide and conquer recursively [20]. The algorithm follows the steps below:

- Choose a feature to be named as root node and make branch for all possible feature.
- Divide the training set with one set for a branch.
- Recursively repeat the process for every branch using its data.
- Stop when all the data has same class.

2.4. Boruta feature selection

Boruta feature selection is built around the random forest classification algorithm, which is carried out without tuning of parameters and numerical estimate of the important feature. Random forest is a classification method which is performed by voting of multiple unbiased decision trees built from samples of the training set [21]. The importance of the feature is obtained from the loss of accuracy of classification. In addition, the decision of a tree isn't influenced by another in the forest. The Boruta Feature Selection algorithm [22]:

- Build extended system information by adding copies of all feature randomly permuted across objects.
- Shuffle the system to minimize the correlations with the response
- Perform random forest to obtain the Z scores computed.
- Determine the maximum Z score for the extended feature and assign the feature assuming it has a better score than the extended. Furthermore, run two-sided test of equality using the Z score for the extended feature of each attribute with undetermined importance.
- Label the feature with lower Z score for the extended 'unimportant' feature and remove it from the system.

– Label the feature with higher Z score for the extended 'important' feature.
– Remove all copied feature and repeat the procedure till none is removed

2.5. Random forest

Random forest was first introduced by Ho in 1995 to split nodes. It is the ensemble of many decision trees using bootstrapping and random feature selection. Random forest is suitable for this study because it performs well on large datasets. Figure 2 shows a diagram of random forest which built of many decision tree.

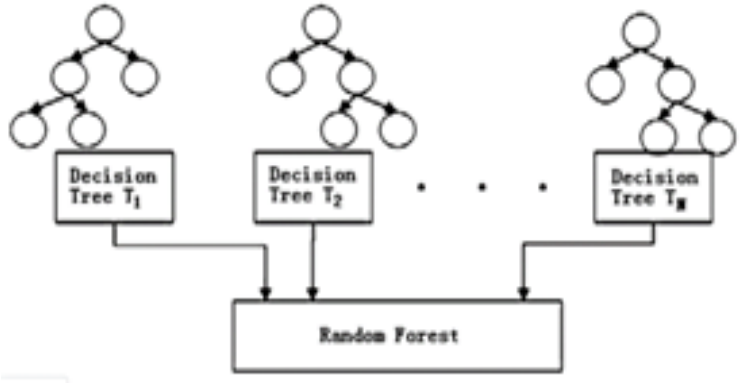

Figure 2. Diagram of random forest which consists of many decision trees

Random forest is a classifier consisting of classification tree $\{\boldsymbol{T}(\boldsymbol{u}, \boldsymbol{v}_i), i = 1, \dots, I\}$ where $\{\boldsymbol{v}_i\}$ is a vector with independently and identically distributed with each tree vote at the input u. The accuracy from decision tree is more stable and accurate. However, the random forest performs better accuracy which makes the correlation of tree significant. By forming a lot of tree into random forest the risk of getting over fitting is reduced, with the error and converge into some value generalized. Given ensemble classifier $T_1(\boldsymbol{x}), T_2(\boldsymbol{x}), \dots, T_I(\boldsymbol{x})$ with random training data obtained from vector **X** and Y, the function of the margin is written as:

$$mg(\boldsymbol{X}, Y) = av_b P(T_I(\boldsymbol{X}) = Y) - \max_{j \neq Y} av_b (T_I(\boldsymbol{X}) = j) \quad (1)$$

Where P is the indicator function and av_b the average, with $T_i(\boldsymbol{X}) = Y$ the result of classification, where Y is the class prediction and $T_i(\boldsymbol{X}) = j$ is the result of classification with j. Margin is used to determine the average value of votes **X**, Y. A greater margin value means a more accurate classification.

$$\xi = W_{\boldsymbol{X},Y}(mg(\boldsymbol{X}, Y) < 0) \quad (2)$$

ξ denotes the generalization error, while $W_{\boldsymbol{X},Y}$ indicate that the probability is more than **X**, Y dimension [23].

2.6. Support vector machines

Support Vector Machines is one of the supervised learning methods widely introduced by Vapnik in the late 90s. During its early days, it was used only for classification, however, it has developed, and capable of solving regression problems. SVM try to solve the classification problem by forming a hyperplane which maximizes the margin by dividing the data into classes. The nearest distance from the hyperplane to the point of each class is known as margin. Figure 3 is an illustration of SVM model.

Given a dataset $P = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_v, y_v)\}, \boldsymbol{x}_u \in \boldsymbol{X}, y_u \in Y = \{-1, 1\}$, SVM try to solve the following equation:

$$\min_{G,b} \frac{1}{2} ||\boldsymbol{G}||^2 \quad (3)$$

$$y_u(\boldsymbol{G}^t \boldsymbol{x}_u + b) \geq 1, \quad u = 1,2,\dots,v$$

For some error cases, parameter $S > 0$ and slack variable $\mathcal{E} \geq 0$ was added to the equation:

$$\min_{G,b} \frac{1}{2} ||\boldsymbol{G}||^2 + S \sum_{u=1}^{v} \varepsilon_u \tag{4}$$

$$y_u(\boldsymbol{G}^t \boldsymbol{x}_u + b) \geq 1 - \mathcal{E}_u, \quad u = 1,2,\dots,v$$

Kernel function is used for some problem which can't be solved using linear hyperplane. Its function is defined as:

$$\varkappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)^T \varphi(\boldsymbol{x}_j) \tag{5}$$

These are the most common kernel function [24, 25]:

- Linear : $\varkappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$
- Radial Basis Function : $\varkappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp\left(-\varphi \left|\left|\boldsymbol{x}_i - \boldsymbol{x}_j\right|\right|^2\right), \varphi > 0$
- Polynomial : $\varkappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\boldsymbol{x}_i^T \boldsymbol{x}_j + 1\right)^m, m > 0$

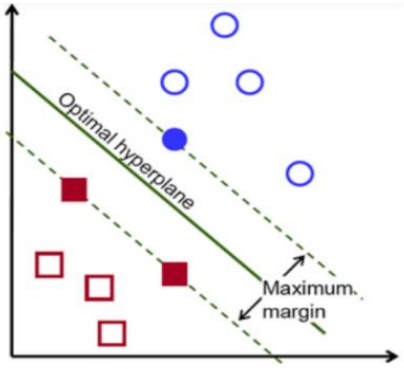

Figure 3. SVM solve classification problem by performing a hyperplane which maximize the margin of the data

2.7. Model performance validation

This study applied Hold-Out Validation to validate the performance of the model. It functions by splitting the data into training and testing data with the model built from the training data and tested with the testing data. A different percentage of training data was applied, to overcome the weakness of Hold-Out Validation, which was performed nine times with different percentage of the training data utilized. The performance of the model is obtained from the accuracy with the formula written as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{6}$$

For TP, TN, FP, FN represents True Positive, True Negative, False Positive, and False Negative.

3. RESULTS AND ANALYSIS

This study used software Rstudio version 1.1.463 for both random forest and support vector machines.

3.1. Result of boruta feature selection

The result of Boruta Feature Selection determines whether the feature is important or not as shown in Table 2. According to the result shown on the table, the important features are Age, BMI, Glucose, HOMA, and Resistin. However, the features which are labeled important to make the machine learning model are random forest and support vector machines.

Table 2. Result of boruta feature selection

Feature	Result
Age	Important
BMI	Important
Glucose	Important
Insulin	Unimportant
HOMA	Important
Leptin	Unimportant
Adiponectin	Unimportant
Resistin	Important
MCP. 1	Unimportant

3.2. Breast cancer classification using random forest

According to the result shown on the Figure 4, the model performed best with 80% of training data, resulting in an accuracy of 90.90%. Conversely, the worst accuracy was recorded at 74.75% with 10% training data.

Figure 4. Results of accuracy of breast cancer classification performed by random forest

3.3. Breast cancer classification using support vector machines

According to the result shown on the Figure 5, the model performed best using 80% of the training data which resulted in an accuracy of 95.45%. Conversely, the worst accuracy is recorded at 72.81% with 10% training data.

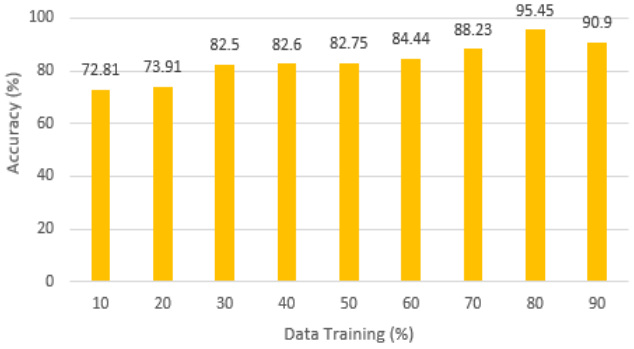

Figure 5. Results of accuracy of breast cancer classification performed by support vector machines with RBF kernel, parameter $C = 1$ and $\sigma = 0.328524$

4. CONCLUSION

This study used Random Forest (RF) and Support Vector Machines (SVM) as the machine learning methods to classify breast cancer. Furthermore, the Hold-Out Validation was used to validate and evaluate the performance of the model, from the simulation for SVM with Radial Basis Function (RBF) kernel, with

$C = 1$ and $\sigma = 0.328524$ as the best parameter of the model. According to the experiment result, RF scored the best accuracy at 90.90% using 80% training data, while SVM had better accuracy at 95.45% using 80%. These results show that the performance of SVM is better than RF in terms of accuracy.

ACKNOWLEDGEMENTS

This research supported financially by the Ministry of Research and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with a PTUPT 2020 research grant scheme, ID number 1621/UN2.R3.1/HKP.05.00/2019

REFERENCES

[1] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert systems with applications*, vol. 36, no. 2, pp. 3240-3247, 2009.
[2] H-Y. Chuang, E. Lee, Y-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Mol. Syst. Biol.*, vol. 3, no. 140, pp. 1-10, 2007.
[3] Arfiani, Z. Rustam, J. Pandelaki, and A. Siahaan, "Kernel spherical K-means and support vector machine for acute sinusitis classification," *IOP Conference Series: Materials Science and Engineering*, vol. 546, pp. 1-9, 2019.
[4] Tzotsos, Angelos & Argialas, Demetre, "Support Vector Machine Classification for Object-Based Image," *Object-Based Image Analysis*, pp. 663-677, January 2008.
[5] Z. Rustam, F. Nadhifa and M. Acar, "Comparison of SVM and FSVM for preditcing bank failures using chi-square feature selection," *IOP Conference Series: Journal of Physics*, vol. 1108, no. 1, pp. 1-7, 2018.
[6] J. Maharani and Z. Rustam, "The application of multi-class support vector machines on intrusion detection system with the feature selection using information gain," *1st Annual International Conference on Mathematics, Science, and Education (ICoMSE)*, vol. 218, pp. 1-2, 2017.
[7] Z. Rustam and N. Olievra, "Comparison of fuzzy robust kernel C-means and support vector machine for intrusion detection systems using modified kernel nearest neighbor feature selection," *AIP Conference Proceedings: 3rd International Sympossium on Current Progress in Mathematics and Sciences (ISCPMS)*, pp. 020215-1-6, 2017.
[8] Z. Rustam and N. P. A. A. Ariantari. Comparison Between Support Vector Machine and Fuzzy Kernel C-Means as Classifier for Intrusion Detection System using Chi-Square Feature Selection, *AIP Conference Proceedings: 3rd International Sympossium on Current Progress in Mathematics and Sciences (ISCPMS)*, vol. 2023, 2018.
[9] T. V. Rampisela and Z. Rustam, "Classification of schizophrenia data using support vector machine (SVM)," *IOP Conference Series: Journal of Physics*, vol. 1108, pp. 1-7, 2018.
[10] J. Xiao, "SVM and KNN ensemble learning for traffic incident detection," *Physica A: Statistical Mechanics and its Applications*, Elsevier, vol. 517, pp. 29-35, 2019.
[11] Z. Rustam and A. A. Ruvita, "Application support vector machine on face recognition for gender classification. *Conference Series: Journal of Physics*, vol. 1108, pp. 1-5, 2018.
[12] Rustam, Zuherman & Faradina, Ridhani. (2018). Face Recognition to Identify Look-Alike Faces using Support Vector Machine. Journal of Physics: Conference Series. 1108. 012071. 10.1088/1742-6596/1108/1/012071.
[13] R. Diaz-Uriarte and S. A. de Andres, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 3, pp. 1-13, 2006.
[14] M. Alam and S. T. Vuong, "Random forest classification for detecting android malware," *IEEE Int. Conf. on Green Comp. and Comm. and IEEE Internet of Things and IEEE Cyber, Physical and Soc. Comp.*, pp. 663-669, 2013.
[15] Z. Rustam and G. Saragih, "Predicting Bank Financial Failures using Random Forest," *International Workshop on Big Data and Information Security (IWBIS)*, pp. 81-86, 2018.
[16] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature selection using random forest classifier for predicting prostate cancer," *IOP Conference Series: Materials Science and Engineering*, vol. 546, pp. 1-8, 2019.
[17] U. Aprilliani and Z. Rustam, "Osteoarthritis disease prediction based on random forest," *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 237-240, 2018.
[18] UCI Machine Learning Repository, Accessed 10 August 2019, Available at https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra.
[19] B. S. Everitt, "Classification and regression trees," John Wiley & Sons, Inc., 2005.
[20] C. Bishop, "Pattern recognition and machine learning," Springer, New York, 2006.
[21] M. B. Kursa and W. R. Rudnicki, "Feature selection with Boruta package," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1-13, 2010.
[22] M. B. Kursa, A. Jankowski, and W. Rudnicki, "Boruta-a system for feature selection," *Fundam. Inform.*, vol. 101, no. 4, pp. 271-285, 2010.
[23] L. Breiman, "Random forest," *Machine Learning*, vol. 45, pp. 5-32, 2001.
[24] N. Cristianini and J. S. Taylor, "An introduction to support vector machines and other kernel-based learning methods," *Cambridge University Press*, 2000.
[25] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Methods in Molecular Biology* (N. J. Clifton), vol. 609, pp. 223-239, 2010.