Random forest prognostic factor in colorectal cancer

G Anuraga¹, J W Fernanda² and Pebrianty³
¹Department of Statistics, Faculty of Science and Mathematics, PGRI Adi Buana University, East Java, Indonesia.
²Department of D3 Medical Record and Health Information Management, Faculty of Health Science, Institute of Health Sciences Bhakti Wiyata, East Java, Indonesia.
³Department of Public Health, Faculty of Health Science, Institute of Health Sciences Bhakti Wiyata, East Java, Indonesia.
E-mail: g.anuraga@unipasby.ac.id

Abstract. In developing countries such as Indonesia, colorectal cancer cases in women are the third largest after breast cancer and cervical cancer, whereas, in men, cancer ranks second after lung cancer, followed by the third is prostate cancer. This study aims to determine the factors that affect the survival of colorectal cancer patients in the city of Makassar, Indonesia. The data used in this study including colon cancer patients diagnosed first in 2012 in 4 hospitals in Makassar City and observed survival until 2015. Predictor variables consisted of comorbidity, stage of cancer, age, treatment status, the location of cancer, sex, and history of metastasis of patients with colorectal cancer. The samples used in this study were as many as 38 cancer patients. In this study, we are using random forest which is an algorithm used in data classification through tree merging by training on sample data. Random Forest also an ensemble method consisting of several decision trees as classifiers. In a nutshell, the accuracy of this models can be justified by the value of classification by Area Under Curve (AUC) equal to 50%. Moreover, the most influential variable on the survival of colorectal cancer patient is a history of metastasis of colorectal cancer patient, cancer location and gender respectively.

1. Introduction
Cancer is one of the crucial groups of non-communicable diseases in Indonesia, where sufferers bear a considerable economic burden and have a high risk of death. This disease arises due to abnormal growth of body tissue cells that turn into cancer cells. There are several risk factors for behavior and diet that can cause cancer, including high body mass index, lack of fruit and vegetable consumption, lack of physical activity, smoking and excessive alcohol consumption. According to data from the Ministry of Health of the Republic of Indonesia in 2013. It was known that the prevalence of cancer reached 0.14 percent or 347,792 people from the total population[1] The World Health Organization (WHO) also predicts the number of people living with cancer in Indonesia will increase seven-fold by 2030, the most deaths caused by cancer are cases of lung cancer, breast cancer, and colorectal cancer. Colorectal cancer is a health problem or disease that occurs in many developed countries and some large cities in developing countries[2]. The disease begins in the large intestine or rectum. This disease has a high incidence and prevalence; sufferers also bear substantial psychological and economic burdens with many variations of treatment in handling them in healthcare institutions. According to the 2017 National Guidelines for the Management of Colorectal Cancer Treatment. There are two risk factors that can increase the risk of colorectal cancer, namely modifiable factors such as, a history of...
adenoma or polyp KKR and chronic inflammatory disease of the intestine; and factors that cannot be modified are inactivity, obesity, high consumption of red meat, smoking and consumption of alcoholic beverages [3]. The Global Burden Cancer (Globocan) in 2012 stated that the incidence of colorectal cancer in Indonesia was 12.8 per 100,000 population in adulthood, with a mortality of 9.5% of all cancer cases. Which cause by changes in the diet[4] of the Indonesian people, both as a consequence of increasing prosperity and a shift towards western ways of eating (westernization) which is higher in fat and low in fiber[5]. Various studies on the incidence of colorectal cancer among them have been conducted by [6] obesity also associated with an increased risk of all cancer-specific deaths and deaths in colorectal cancer patients. However, the profile of tumor samples concluded that functional analysis of the gene sign used for subtyping showed the relationship of patients with colorectal cancer [7].

Random forest is one machine learning algorithm that is easy to use and works fast in computational iteration processes. Random forest works with Classification and Regression Trees (CART) procedures as the base learner or weak learner, which is combined with optimizing random nodes and bagging, meaning that the predictors are taken randomly to be used in the classification modeling process and then to do the resampling process with returns [8]. This method can provide a new practice in classification modeling, namely through out-of-bag errors as an estimate of generalization errors and can measure essential variables through permutation [9]. Random forest popular analysis in ecological data as well as classifying invasive plant species [10], species counts [11], medical statistics and bioinformatics [12]. The random forest can be using and describe the classification of the risk of diabetic retinopathy (DR) [13]. Based on the description above, this study aims to determine the factors that influence the survival of colorectal cancer patients in the city of Makassar. This study uses the first data in 2012 in 4 hospitals in Makassar City and observed survival until 2015.

2. Materials and Methods

This study is a retrospective cohort study, which is by looking at the history of patients with colorectal cancer, then followed in the future to see the survival of cancer patients. The data used in this study are secondary data in the form of data about colon cancer patients who were first diagnosed in 2012 in 4 Makassar City Hospitals and observed their survival until 2015. Predictor variables used were comorbidity, stage of cancer, age, treatment status, the location of cancer, sex and history of metastasis in colorectal cancer patients. The dependent variable in this study is the survival of patients with colorectal cancer. Cancer stage is divided into two, namely the category of early and advanced stages. The age of the patient is categorized with an age range of fewer than 50 years and is greater or equal to 50 years. Comorbidity is categorized as heavy and light. Treatment status is categorized into two, namely complete if treatment (surgery, chemotherapy, radiation) has been carried out according to the doctor's recommendations recorded in the medical record. Treatment status is categorized as incomplete if the patient does no treatment according to the doctor's recommendation stated in the medical record. The location of cancer consists of two, namely the colon and rectum. Metastatic history is a history of cancer cells that have reached other organs [14]. There are two categories in metastatic. First, the development of primary tumor cells to other organs of the laboratory results from support, and second, no metastatic history if no primary tumor cell development found in other organs.

The Random Forest was developed with the idea that sample data which can be seen in figure 1. The data will be collected randomly (bagging) to form a classification tree, in addition to the predictor variables, were also taken randomly and only selected as the best sorter when determining tree sorting [9]. Random forest classification is carried out through the incorporation of trees by conducting training on the sample data. Random forest uses the Decision Tree to do the selection process. The built tree is divided recursively from the data in the same class. Solving is used to share data based on the type of attribute used. The following is an illustration of a random forest.
The Random Forest algorithm generally has the following stages:
1. Take $n$ sample data from the initial dataset by bootstrap.
2. In each bootstrap result dataset, the classification tree form without pruning where the best node sorting is done based on randomly selected predictor variables.
3. Predict classification of data samples by combining the prediction results of classification trees based on majority voting.

Moreover, table 1 can represent as accuracy calculation in classification to see how large the grouping of data is adequately classified in the group. The Receiver Operating Characteristic (ROC) curve is considered more informative in measuring accuracy\cite{15}. The ROC curve can illustrate how a classification method can measure sensitivity and specificity \cite{16}. The area under the curve reflects the goodness of a classification method. Grouping a set of data to be classified correctly in the classification case can be illustrated in the table as follows:

| Prediction class | $l_1 = -1$ | $l_2 = 1$ |
|------------------|------------|------------|
| Actual Class $y_1 = -1$ | $n_{11}$ | $n_{12}$ |
| Class $y_2 = 1$ | $n_{21}$ | $n_{22}$ |

where:
- $n_{11} =$ The number 11 of the exact observations is classified as $y_1$
- $n_{12} =$ The number 11 of the incorrect observations are classified as $y_2$
- $n_{21} =$ Amount of 12 from observations that are incorrectly classified as $y_1$
- $n_{22} =$ The number 12 of the exact observations is classified as $y_2$

Calculation of sensitivity and specificity can be describe by equation (1) and (2):

\[
Sensitivity = \frac{n_{11}}{n_{11} + n_{12}} \times 100\% \quad (1)
\]

\[
Specificity = \frac{n_{22}}{n_{21} + n_{22}} \times 100\% \quad (2)
\]

3. Results and Discussions
In figure 2A, it can be seen that the survival factor of colorectal cancer patients who are less than one year for respondents male sex is 14 respondents and women are 15 respondents, while the survival of more than one year for male gender as many as seven respondents and women as many as two respondents. For indicators of treatment status, respondents with complete treatment status are 21
respondents and incomplete as many as 17 respondents. Cancer stage indicator is divided into two categories: early stage as many as ten respondents and advanced stages of 28 respondents.

Figure 2.A Descriptive analysis of survival factors for colorectal cancer by gender, treatment status and stage of cancer

Figure 2.B Descriptive analysis of survival factors for colorectal cancer by location of cancer, metastasis and age
Based on comorbidities on figure 2.A and 2.B, it was known that respondents with mild comorbidity amount 33 respondents and severe comorbidities amount five respondents. Patients with cancer in the colon amount 14 respondents and in the rectum were 24 respondents. Patients with a metastatic history of 21 respondents and without a metastatic history of 17 respondents, and based on age known as many as eight respondents under the age of 50 years and 30 respondents aged more than 50 years. Age, gender, family history of colorectal cancer, smoking, BMI are some of the factors that influence colorectal cancer [17]. Risk factors for colorectal cancer were identified using the logistic regression model.

Based on Figure 3 the importance variable analysis on random forest it is known that six variables have a contribution in determining the classification of survival of people living with colorectal cancer. Four crucial variables in the case of classification of patients with colorectal cancer survival include the metastatic history of patients with colorectal cancer, the location of cancer, sex, and comorbidity. However, the results of the classification accuracy in the random forest can be known based on the ROC curve. The area under Curve (AUC) is useful for knowing the accuracy of a random forest. The results of the study using random forest found that the value of classification accuracy based on Area Under Curve (AUC) is 50%.

Figure 3. ROC-Curve and importance of the variables in random forest

4. Conclusions
The results of the analysis using random forest produced four essential variables in the case of classification of patients with colorectal cancer survival, namely a history of metastasis with colorectal cancer, the location of cancer, gender, comorbidity, with classification accuracy based on Area Under Curve (AUC) of 50%.

Acknowledgments
This research was supported by Jerhi Wahyu Fernanda, Pebrianty (Faculty of Health Science, Institute of Health Sciences Bhakti Wiyata). We thank our colleagues for providing insight and expertise that greatly helped in the completion of this article.

References
[1] Kementerian 2015 Kesehatan Pusat Data dan Informasi Kesehatan RI.
[2] Pardamean B, Baurley J W, Pardamean C I and Figueiredo J C 2016 International Journal of Colorectal Disease 31 1537.
[3] Komite Penanggulangan Kanker Nasional 2014 Pedoman Nasional Pelayanan Kedokteran Kanker Kolorektal no c.
[4] Kurniawan M Y, Briawan D and Caraka R E 2015 J. Gizi Klin. Indones. 11 105.
[5] Ferlay J et al. 2015 Int. J. Cancer 136 E359.
[6] Billy M, Sholihah H, Andanni K, Anggraeni M I, Siregar S M and Mirtha L T 2016 Acta Med Indones. 48 242.
[7] Rashid M, Vishwakarma R, Deeb A, Hussein M A and Aziz M A 2018 Molecular Classification of Colorectal Cancer Using Gene Expression Profile of Tumor Samples.

[8] Breiman L 2004 Tech. Rep. 670. Stat. Dep. Univ. Calif. Berkeley p. 10.

[9] Breiman L 1999 Mach. Learn 45 1.

[10] Cutler D. R et al. 2007 Ecology 88 2783.

[11] Caraka R E et al. 2018 Procedia Comput. Sci. 135 377.

[12] Pardamean B, Budiarto A and Caraka R E 2014 Bioinformatika dengan R Tingkat Lanjut, 1st ed. (Yogyakarta: Teknosains).

[13] Casanova R, Saldana S, Chew E Y, Danis R P, Greven C M and Ambrosius W T 2014 PLoS One 9.

[14] Alsarraj J and Hunter K W 2013 Genomic and Personalized Medicine 2 776.

[15] A. Agresti, An Introduction to Categorical Data Analysis. 2007.

[16] Hernández-Orallo J 2013 Pattern Recognit. 46 3395.

[17] Kaminski M F, Polkowski M, Kraszewska E, Rupinski M, Butruk E and Regula J 2014 Gut. 63 1112.