Pancreatic cancer classification using logistic regression and random forest

Zuherman Rustam, Fildzah Zhafarina, Glori Stephani Saragih, Sri Hartini
Department of Mathematics, University of Indonesia, Depok, Indonesia

Article Info

Article history:

Received Mar 29, 2020
Revised Mar 2, 2021
Accepted Apr 6, 2021

Keywords:

Classification
Logistic regression
Machine learning
Pancreatic cancer
Random forest

ABSTRACT

In the medical field, technology machinery is needed to solve several classification problems. Therefore, this research is useful to solve the problem of the medical field by using machine learning. This study discusses the classification of pancreatic cancer by using regression logistics and random forest. By comparing the accuracy, precision, recall (sensitivity), and F1-score of both methods, then we will know which method is better in classifying the pancreatic cancer dataset that we get from Al-Islam Hospital, Bandung, Indonesia. The results showed that random forest has better accuracy than logistic regressions. It can be seen with maximum accuracy of logistic regressions 96.48 with 30% data training and random forest 99.38% with 20% of data training.

Corresponding Author:

Fildzah Zhafarina
Department of Mathematics
University of Indonesia
University of Indonesia, Depok 16424, Indonesia
Email: fildzah.zhafarina@sci.ui.ac.id

1. INTRODUCTION

One of the main diseases that cause death in the world is cancer [1], [2]. These diseases can attack all parts of the body [3]. One of the main causes of cancer-related deaths worldwide is pancreatic cancer. In the early stage, the diseases have no showing or symptoms. The most symptoms occur when the diseases in the final stage [4]. Pancreatic cancer is cancer that starts in the pancreas. The most common type of pancreatic cancer is pancreatic adenocarcinoma [5]. Location of the pancreatic organs behind the stomach. The pancreas is about 6 inches long and less than 2 inches wide in adults [6]. There are various treatments for pancreatic cancer, such as surgery, chemotherapy, radiation therapy, or a combination of these. The method of treatment is chosen based on the extent of cancer [7]. Information technology has an important role in the field of medicine. Cancer is a disease that can be detected by machine learning. Data is very useful in the medical field. It can be seen from the development of data mining in medical science is increasing rapidly. This increase can be seen from the high prediction results, can reduce treatment costs, increase the chances of recovery of patients, and decisions to save lives [8], [9]. Classification is a way to identify groups of categories to be part of observations [10]. One general classification is the continuous value of the predictive attribute. Whereas, ensemble classification is useful for increasing classification accuracy in ensemble applications [11].

2. RESEARCH METHOD

Pancreatic cancer dataset was obtained from Al-Islam Hospital, Bandung, Indonesia. This dataset consists of 79 non-cancer and 124 cancer samples with numerical characteristics described by 6 attributes, as shown in Table 1. This research uses logistic regressions and random forest for classification. This method is evaluated using 3-fold cross-validation, 45-random state, and later compared. Table sample of dataset is shown in Table 2.

Table 1. Pancreatic cancer dataset variable

Attributes	Description
Age	The number of age patients who are in check
CA 19-9	The number of cancer antigen units per milliliter of blood
Hemoglobin	The number of hemoglobin gram per deciliter of blood
Leukocyte	The number of leukocyte cell per uL of blood
Hematocrit	Hematocrit or the volume percentage of red blood cells
Thrombosis	The number of thrombosis cell per uL of blood

Table 2. Sample of dataset from Al-Islam Hospital, Bandung, Indonesia

Age	CA 19-9 < 37 (U/mL)	Hemoglobin 13-18 (g/dL)	Leukocyt 4000-10000 (sel/uL)	Hematokrit 40-54 (%)	Thrombosis 150000-450000 (sel/uL)
38	34.61	12.1	7600	36.9	244000
82	35.02	12.1	4900	36.7	253000
35	35.4	6.3	10100	23.4	496000
58	35.83	9.8	33500	29.1	467000
52	36	9.8	7600	29.9	613000
41	36.03	12.6	3400	38	203000
40	36.94	11.9	8900	39.8	430000
51	37.41	6.6	9500	23.5	259000
64	39.25	11.5	15500	35.3	230000

2.1. Logistic regressions

In some cases, the natural complement of ordinary linear is logistic regression. This happens when each target variable is categorized. Variable Y is a variable target and dependent with two class and variable X is a variable predictor and independent, let $g(x) = \Pr(X = x) = 1 - \Pr(X = x)$ the logistic regression model has a linear form for *Logit* with probability as follows [12]-[14]:

$$Logit[\ g(x)\ ] = \log\left(\frac{g(x)}{1-g(x)}\right) = \ \alpha + \beta x\ , where\ the\ odds\ \frac{g(x)}{1-g(x)} \tag{1}$$

The form of linear approximation and probability logarithm is derived from *Logit*. The rate of increase or decrease of the Shape *g(x)* curve is denoted by the parameter β [15].

2.2. Random forest

Random forest is a method developed by Breiman in 2001 [16], [17]. Random forest works when it reaches maximum accuracy, a decision tree can be used to avoid overfitting data [18]. The estimation process previously carried out by decision tree and CART was enhanced by Breiman, which was started by randomly selecting m variables from several independent variables. A decision tree or CART method is a tree that is grown without pruning. These trees will be selected with the highest accuracy. The procedure of random forest depends on the number of classifications [19]. There are some advantages of random forest [20], such as overcoming the problem of excessive compatibility, less sensitive to outlier data, parameters can be easily adjusted and therefore eliminate the need for tree pruning, and the importance of variables and accuracy are generated automatically. Random forest selected features are in agreement with existing domain knowledge (e.g. physiological knowledge Guan *et al*., 2012) [21]. Flowchart of random forest shown in Figure 1.

2.3. Confusion matrix

One of the methods used to calculate accuracy in the concept of data mining or decision support systems is confusion matrix [22]. It is balanced in the precision and sensitivity that distinguishes correct label classifications in different classes [23], [24]. Accuracy is the ratio of the true predictions in the whole data. Precision is a true positive prediction ratio compared to overall positive predicted results. In addition, the third is sensitivity is a true positive prediction ratio compared to overall true positive data. The last was denoted as F1-score, used to determine the balance between sensitivity and precision.

- True Positive (TP): Number of samples having pancreatic cancer diagnosed correctly
- False Positive (FP): Sum of healthy people that were incorrectly identified to have pancreatic cancer
- True Negative (TN): Number of healthy people correctly spotted
- False Negative (FN): Number of samples with pancreatic cancer that were incorrectly classified as healthy

From Table 3 it can build the formula for accuracy, precision, recall (sensitivity), and F1-score that are seen in (2)-(5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \tag{2}$$

$$Precision = \frac{TP}{TP+FP} x\ 100\% \tag{3}$$

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \tag{4}$$

$$F1Score = 2\ x\ \frac{(Precision\ x\ Recall\ )}{(\ Precision+Recall\ )} x\ 100\% \tag{5}$$

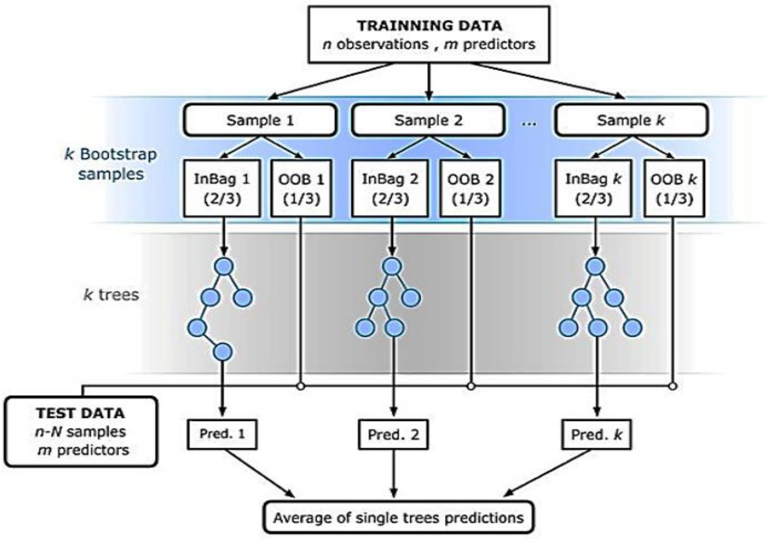

Figure 1. Flowchart of random forest [25]

Table 3. Confusion matrix

Actual Value	Recognize Value	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

3. RESULTS AND DISCUSSION

This research using Jupyter notebook as software for running the program of logistic regressions and random forest in processing pancreatic cancer classification problem. Testing the accuracy, precision, recall, and F1-score in this type of classification are by changing the amount of data training. In this test, the number of data training is equal to 10, 20, 30, 40, 50, 60, 70, 80, and 90 which will be used on the results of the dataset. The results of accuracy, precision, recall, and F1-score which are given by logistic regressions and random forest classifier method are shown in Table 4 and Table 5.

Based on Table 4, it is shown that the number of data training is affecting by the values of accuracy, precision, recall, and F1-score. In this research, the highest accuracy value was recorded when the data training is 30% with 96.48% while the lowest accuracy value was recorded when the data training is 70% with 91.49%. In precision, 70% and 90% of data training reached a maximum value that is 100%. For the recall, the recall of the highest value is 97.70% for 30% of data training. The last for F1-score, 30% of data training reached the highest value that is 96.29%.

Based on Table 5, it is shown that the number of data training is affected by the values of accuracy, precision, recall, and F1-score. In this research, the highest accuracy value was recorded when the data training is 20% with 99.38% while the lowest accuracy value was recorded when the data training is 90% with 89.68%. In precision, 10%, 20%, and 30% of data training reached a maximum value that is 100%. For the recall, the recall of the highest value is 99.10% for 10% of data training. The last for F1-score, 20% of data training reached the maximum value that is 100%.

Table 4. The results of pancreatic cancer classification using logistic regression

No.	Data Training	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1.	10	95.62	96.56	96.4	95.4
2.	20	95.68	96.02	96.97	95.45
3.	30	96.48	96.74	97.7	96.29
4.	40	95.94	97.44	96	95.73
5.	50	96.05	98.33	95.16	95.88
6.	60	95.01	96.06	95.83	94.76
7.	70	91.49	100	86.11	91.39
8.	80	95.05	96.3	95.83	94.77
9.	90	94.44	100	91.67	94.29

Table 5. The result of pancreatic cancer classification using random forest

No.	Data Training	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1.	10	98.91	100	99.1	98.85
2.	20	99.38	100	98.99	100
3.	30	99.29	100	98.85	99.26
4.	40	95.12	96.43	98.67	95.54
5.	50	97	96.9	98.41	96.85
6.	60	97.48	98.04	91.91	98.67
7.	70	91.73	97.44	94.44	96.57
8.	80	97.62	92.59	91.67	97.17
9.	90	89.68	93.33	91.67	70.83

4. CONCLUSION

After classifying pancreatic cancer with logistic regressions and random forest methods, it gets several results of accuracy, precision, recall, and F1-score. By comparing the values that are given from those methods (logistic regressions and random forest), it is possible to conclude that random forest generates a better result than logistic regression. The results of the two methods random forest gives the highest accuracy rate when the data training is 20% with 99.38%, while logistic regression reaches 96.48% when the data training is 30%. Because of the good results, random forest is suggested to help the medical staff to predict or classify a disease rather than logistic regression, especially for a dataset that is similar to this research.

ACKNOWLEDGEMENTS

This research full supported financially by University of Indonesia with a PUTI 2020 research scheme grant.

REFERENCES

[1] U. S. L. Wang *et al.*, "Signaling adaptor protein Crk is involved in malignant feature of pancreatic cancer associated with phosphorylation of c-Met," *Biochemical and biophysical research communications,* vol. 524, no. 2, pp. 378-384, 2020. https://doi.org/10.1016/j.bbrc.2020.01.105.

[2] World Health Organization, "10 Facts About Cancer," 2018. [Online]. Available: https://www.who.int/features/factfiles/cancer/en/. [Accessed 26 March 2021].

[3] M. Samandari, M. G. Julia, A. Rice, A. Chronopoulos, A. E. D. R. Hernandes, "Liquid biopsies for management of pancreatic cancer," *Translational Research,* vol. 201, pp. 98-127, 2018, doi: 10.1063/1.4991237.

[4] V. Panca and Z. Rustam, "Application of machine learning on brain cancer multiclass classification," *AIP Publishing LLC,* vol. 1862, p. 030133, 2017, doi: 10.1016/j.trsl.2018.07.008.

[5] J. Kokkinos *et al.*, "Targeting the Undruggable in pancreatic Cancer using Nano-based gene silencing drugs," *Biomaterials,* vol. 240, p. 119742, 2020, doi: 10.1016/j.biomaterials.2019.119742.

[6] American Cancer Society, "Pancreatic Cancer," [Online]. Available: https://www.cancer.org/cancer/pancreatic-cancer.html. [Accessed 26 March 2021].

[7] Mayoclinic, “Pancreatic Cancer,” 2019. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/pancreatic-cancer/symptoms-causes/syc-20355421. [Accessed 26 March 2021].

[8] H. Asri, H. Mousannif, H. Al Moatassime and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Comput. Sci*, vol. 83, pp. 1064-1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[9] D. L. S. Agrawal, R. Panda and A. Abraham, “Optimal breast cancer classification using Gauss–Newton representation based algorithm,” *Expert Systems with Applications,* vol. 85, pp. 134-145, 2017, doi: 10.1016/j.eswa.2017.05.035.

[10] J. Tang, S. Alelyani and H. Liu, “Feature selection for classification: A review,” *CRC Press,* 2014. [Online]. Available: http://www.math.chalmers.se/Stat/Grundutb/GU/MSA220/S18/featselect.pdf.

[11] K. Fawagreh, M. M. Gaber and E. Elyan, “Random forests: from early developments to recent advancements,” *Systems Science & Control Engineering: An Open Access Journal 2,* vol. 1, pp. 602-609, 2014, doi: 10.1080/21642583.2014.956265.

[12] C. Peng, K. Lee and G. Ingersoll, “An introduction to logistic regression analysis and reporting,” *The journal of educational research*, vol. 96, no. 1, pp. 3-14, 2002, doi: 10.1080/00220670209598786.

[13] K. Ghazvini, M. Yousefi, F. Firoozeh and S. Mansouri, “Predictors of tuberculosis: Application of a logistic regression model,” *Gene Reports,* vol. 17*,* p. 100527, 2019, doi: 10.1016/j.genrep.2019.100527.

[14] Y. Cao, X. Zhang, Y. Fu, Z. Lu and X. Shen, “Urban spatial growth modeling using logistic regression and cellular automata: A case study of Hangzhou,” *Ecological Indicators*, vol. 113, p. 106200, 2020, doi: 10.1016/j.ecolind.2020.106200.

[15] C. Aroef, R. Yuda, Z. Rustam and J. Pandelaki, “Multinomial Logistic Regression and Support Vector Machine for Osteoarthritis Classfisication,” *InJournal of Physics: Conference Series,* vol. 012012, p. 1417(1), 2019, doi: 10.1088/1742-6596/1417/1/012012.

[16] T. Octaviani and Z. Rustam, “andom forest for breast cancer prediction,” *In AIP Conference Proceedings,* vol. 2168, no. 1, p. 020050, 2019, doi: 10.1063/1.5132477.

[17] L. Breiman, “Random Forest,” *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[18] M. Huljanah, Z. Rustam, S. Utama and T. Siswantining, “Feature selection using random forest classifier for predicting prostate cancer,” *InIOP Conference Series: Materials Science and Engineering,* vol. 546, no. 5, p. 052031, 2019, doi: 10.1088/1757-899X/546/5/052031.

[19] T. Ho, “Random decision forests,” *In Proceedings of 3rd international conference on document analysis and recognition,* vol. 1, pp. 278-282, 1995, doi: 10.1109/ICDAR.1995.598994.

[20] J. Singh and A. S. Arora, “A framework for enhancing the thermographic evaluation on characteristic areas for paranasal sinusitis detection,” *Infrared Physics & Technology,* vol. 85, pp. 457-464, 2017, doi: 10.1016/j.infrared.2017.08.011.

[21] A. Arfiani and Z. Rustam, “Ovarian cancer data classification using bagging and random forest,” *InAIP Conference Proceedings,* vol. 2168, p. 020046, 2019, doi: 10.1063/1.5132473.

[22] V. Rodriguez-Galiano, M. Sanchez-Castillo, J. Dash, P. Atkinson and J. Ojeda-Zujar, “Modelling interannual variation in the spring and autumn land surface phenology of the European forest,” *Biogeosciences*, vol. 13, no. 11, pp. 3305-17, 2016, doi: 10.5194/bg-13-3305-2016.

[23] Z. Rustam, S. Hartini, N. Putri and J. Pandelaki, “Hierarchical Clustering Algorithm Based on Density Peaks using Kernel Function for Thalassemia Classification,” *in 2nd International Conference on Advanced Intelligent Systems for Sustainable Development, AI2SD 2019,* Marrakech, 2019, doi: 10.1007/978-3-030-36674-2_21.

[24] M. Sokolova, N. Japkowicz and S. Szpakowicz, “Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation,” *in Australasian joint conference on artificial intelligence*, vol. 4, pp. 1015-1021, 2006, doi: 10.1007/11941439_114.

[25] E. Izquierdo-Verdiguier and R. Zurita-Milla, “An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing,” *International Journal of Applied Earth Observation and Geoinformation,* vol. 88, p. 102051, 2020, doi: 10.1016/j.jag.2020.102051.

BIOGRAPHIES OF AUTHORS

Zuherman Rustam is an Associate Professor and a lecturer of the intelligence computation at the Department of Mathematics, University of Indonesia. He obtained his Master of Science in 1989 in informatics, Paris Diderot University, French, and completed his Ph.D. in 2006 from computer science, University of Indonesia. Assoc. Prof. Dr. Rustam is a member of IEEE who is actively researching machine learning, pattern recognition, neural network, artificial intelligence.

Fildzah Zhafarina is a final year student in the Department of Mathematics, University of Indonesia. She is currently working on her thesis, which is firmly about applied mathematics using machine learning. Also, Ms. Fildzah's specialties in research are mostly about machine learning in various fields, especially medical.

Glori Stephani Saragih was born in Medan, 17 January 1997. She is a Bachelor of Science from Department of Mathematics, Universitas Indonesia, who is completing the Master of Science at Universitas Indonesia and is currently pursuing a Ph.D. in intelligence computation. Ms. Glori is currently a Process Improvement Manager in PT. Aplikasi Karya Anak Bangsa (Gojek). Her current research is machine on machine learning and neural network in various fields, especially medical and finance.

Sri Hartini is a Bachelor of Science from the Department of Mathematics, University of Indonesia, who is also completing the Master of Science at the University of Indonesia and is currently pursuing a Ph.D. in intelligence computation. Ms. Hartini is passionately researching machine learning, computer vision, neural networks and deep learning in various fields.