JOURNAL OF DATA SCIENCE
eISSN:2805-5160
Vol.2021:002

Breast Cancer Prediction Model Using Machine Learning

Muhammad Amin Bakri [1], Inna Ekawati [2]

Email: muhammad.aminbakri@gmail.com[1], inna.ekawati@gmail.com[2]

[1]Electronic Engineering Universitas Islam 45 Bekasi, Indonesia

[2]Computer Engineering Universitas Islam 45 Bekasi, Indonesia

Abstract

Breast cancer requires early detection, hence it can be prevented earlier or treated more optimally. This article aims to demonstrate predictive modelling of breast cancer and evaluate the accuracy of its predictions using a machine learning approach. This study uses secondary data from the Wisconsin Breast Cancer Dataset (BCWD) which consists of predictive factors for breast cancer and labels for benign or malignant cancers that result. Modelling with machine learning is done by selecting three candidate algorithms, namely Random Forest, Support Vector Machine, and Logistic Regression. Evaluation of the classification performance of each algorithm is carried out by analysing its sensitivity, specificity, and accuracy. The experimental results show that Random Forest has better prediction accuracy (99.6%) followed by Support Vector Machine (98.7%), and Logistic Regression (93.9%).

Keywords: ***Prediction Model, Breast Cancer, Machine Learning Algorithm***

1. Introduction

Every year, no less than one million women are diagnosed with breast cancer. In fact, according to World Health Organization (WHO), half of them die due to late diagnosis (M. Amrane, S. et al. 2018). Breast cancer itself is a type of malignant tumour that activates breast cells so that it has the potential to spread to other parts of the body. Therefore, as early as possible this disease must be detected from the start. If it can be done well by doctors, it will provide progress in the prevention and efficiency of cancer treatment (A. R. Vaka, et al. 2020). Unfortunately, the research that has been done to predict this disease is dominated by the use of conventional statistical analysis.

Meanwhile, machine learning has the ability to analyse data and extract relationships and key characteristics from a dataset. In addition, machine learning is also able to build computational models that describe data better (M. Amrane, S. et al. ,2018) (E. A. Bayrak, et al. 2019) (H. Asri, et al. 2016) (Y. Li, 2018). The ability of machine learning to detect cancer automatically is also very much needed to help doctors do similar jobs in large quantities consistently (O. I. Obaid, et al. 2018).

With the capabilities and success that machine learning has shown, more and more academics and practitioners are now taking great interest in its predictive abilities (J. A. M. Sidey-Gibbons, C. J. Sidey-Gibbons, 2019). Thus, important fields such as the prevention and treatment of breast cancer are also increasingly in need of machine learning prediction applications, especially with the technical implementation that is easy for doctors to carry out. Therefore, this article aims to demonstrate the modelling and evaluation of machine learning-

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

based predictive models in diagnosing breast cancer by utilizing open source tools and public datasets that are not difficult to access.

2. Literature Study

Machine learning use in detecting breast cancer has been carried out by many researchers. Agarap (2019) conducted a comparative study of six machine learning algorithms to detect breast cancer. The six algorithms used are: GRU-SVM, Linear Regression, Multi-Layer Perceptron, Nearest Neighbor Search, Softmax Regression, and Support Vector Machine. This study utilizes the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to measure the classification accuracy of each algorithm as well as its sensitivity and specificity values. The results conclude that all the algorithms used provide classification performance with an accuracy above 90% (A. F. M. Agarap, 2018)

Ganggayah et.al (2019), used machine learning techniques to detect and visualize significant prognostic indicators of breast cancer survival. The dataset used was taken from the University Malaya Medical Centre, which consisted of 23 independent variables and 1 dependent variable. Prediction model is built using decision tree, random forest, neural networks, extreme boost, logistic regression, and support vector machine methods. The results show that the lowest number of accuracy is the decision tree (79.8%) and the highest accuracy is random forest (82.7%) (M. D. Ganggayah, et al. 2019).

Comparison of performance between different machine learning algorithms in predicting breast cancer was also carried out by Asri et.al (2016). The algorithms used are Support Vector Machine (SVM), Decision Tree (C4.5), Naïve Bayes, and k-Nearest Neighbors using the Wisconsin Breast Cancer dataset. The experimental results concluded that SVM has the highest accuracy (97.13%). Modelling simulation using WEKA tools . Bayrak et.al (2019) compared the performance of the Support Vector Machine (SVM) and Artificial Neural Network (ANN) in classifying the Wisconsin Breast Cancer dataset. The assessment indicators used are the value of precision, recall, and ROC. The results conclude that the accuracy of SVM is better than ANN (E. A. Bayrak, P. Kirci, and T. Ensari, 2019). The cross validation technique was used by Amrane et.al (2019) to test the accuracy of Naïve Bayes (NB) and k-nearest neighbor (KNN) in classifying breast cancer through the Breast Cancer Dataset. In conclusion, KNN provides a higher accuracy value than NB (M. Amrane, S. et al. 2018).

A slightly different experiment was carried out by Li and Chen (2018) by utilizing two different breast cancer datasets, namely the Coimbra Breast Cancer Dataset and the WBCD. Comparison of algorithm performance is carried out using three indicators: prediction accuracy value, F-measurement matrix, and AUC value. In conclusion, Random Forest has a better performance than its four rivals (Y. Li, 2018). .

The most unique study is the article by (Vaka et.al 2020) entitled "Breast cancer detection by leveraging Machine Learning." This article proposes a new method called DNSS (Deep Neural Network with Support Value). Unlike other methods, this technique is based on the support value of a deep neural network. The experimental results do show that the performance of the proposed DNSS is quite good compared to the previous method. These

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

findings are expected to provide significance for new medical systems in the future (A. R. Vaka, et al. 2020).

3. Methods

This study used breast cancer secondary data contained in the Wisconsin Breast Cancer (Diagnosis) data set. This dataset is publicly available at the University of California, Irvine machine learning repository (A. Sharma, et al.2017). The output variable is the result of a diagnosis of breast cancer which is categorized into two conditions, namely malignant (M) and benign (B). While the input variable is the value of each cell nucleus which consists of 10 features, i.e. radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensions. These features are calculated based on the results of digital FNA (fine needle aspirate) images of the breast mass. All of the features describe the characteristics of the cell nucleus in the figure.

The modelling consist of five stages: 1) dataset preparation, 2) pre-processing, 3) machine learning algorithm training, 4) machine learning algorithm testing, and 5) an assessment of the sensitivity, specificity, and accuracy of the algorithm. Dataset preparation, including data cleaning using the SPSS 26 statistical application. Meanwhile, the pre-process stage to algorithm performance assessment uses the open-source Orange tools. Figure 1 shows the research framework.

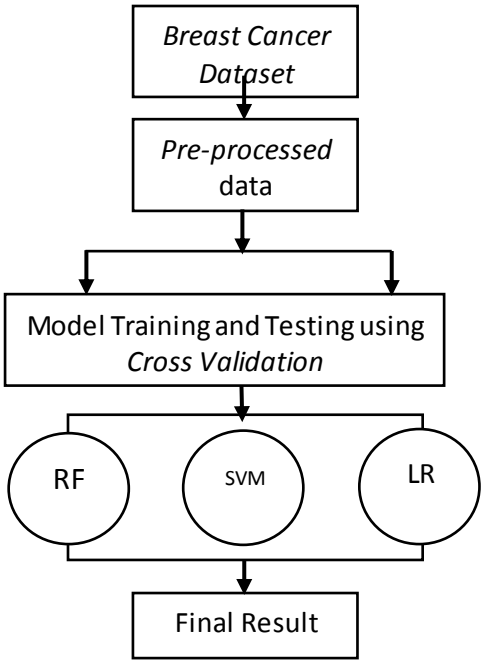


Figure 1. Research Framework

Machine learning algorithms used to obtain breast cancer prediction models are: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR). The three algorithms were trained, tested, and then evaluated for their classification performance.

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

Evaluation of the resulting prediction model is done by analysing the receiver-operation (ROC/AUC) curve, the precision level of the model, the recall rate, and the confusion matrix.

4. Result and Discussion

After data pre-processing, training and performance validation of each algorithm was conducted to conclude the predictive performance of each model. By using the "prediction" widget on the Orange tool, the prediction results are obtained as shown in Figure 2.

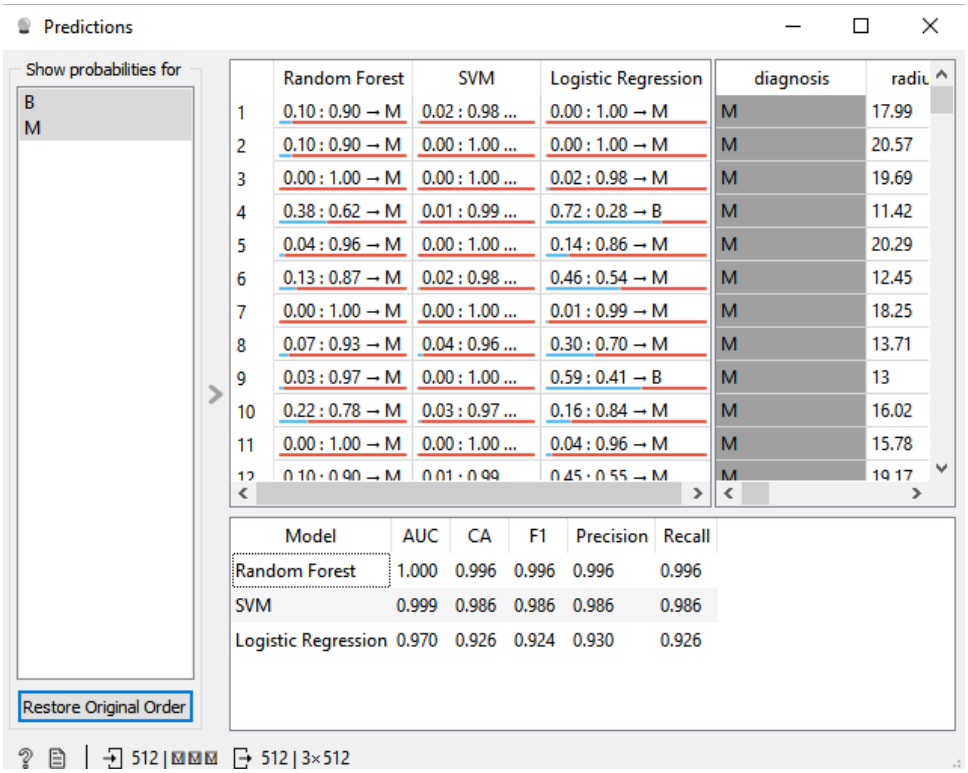
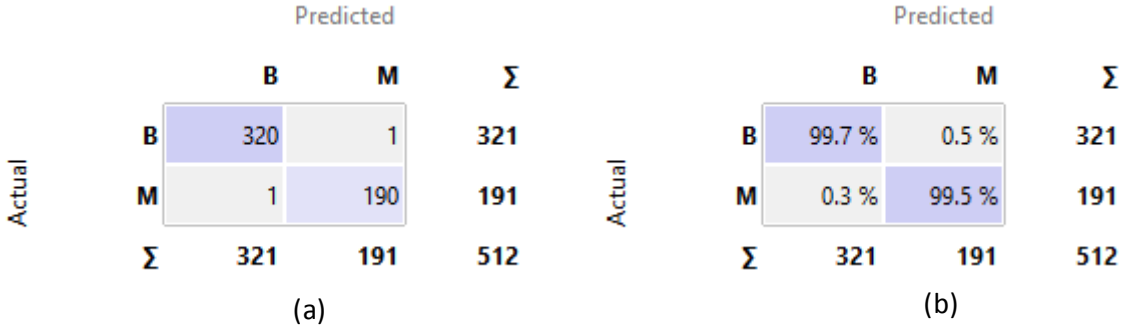


Figure 2. Performance of each model in predicting breast cancer using the BCWD dataset

The prediction results show the advantages of Random Forest over SVM and Logistic Regression based on the parameters AUC, CA, F1, Precision, and Recall. Tests result analysis using a confusion matrix also gave the same results. Figures 3 a and b show the confusion matrix of Random Forest's prediction performance, Figures 4 a and b show the confusion matrix of SVM prediction performance, and Figures 5 a and b show the confusion matrix of predictive performance from Regression Logistics.



Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

Figure 3. Random Forest's confusion matrix

The confusion matrix shows that Random Forest's predictive ability for benign cancer (B) is 99.7% and for malignant cancer (M) is 99.5%. Thus, the Random Forest Model has a prediction accuracy of 99.6%.

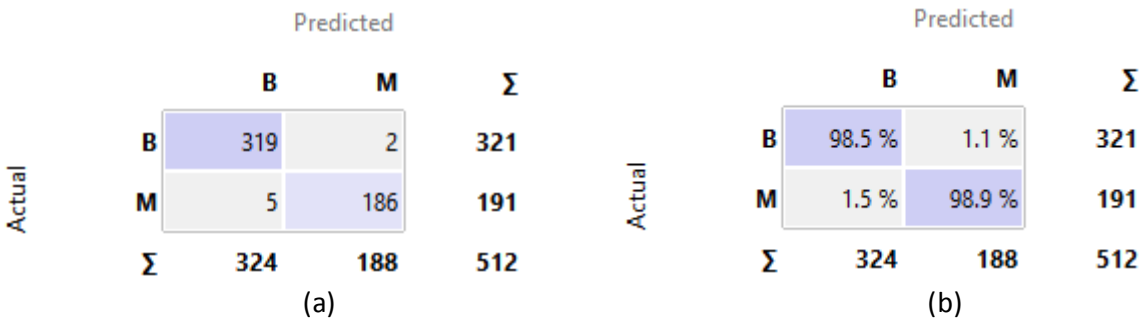


Figure 4. SVM's Confusion Matrix

The confusion matrix shows that the ability of the SVM model to predict benign cancer (B) is 98.5%, whereas for malignant cancer (M) is 98.9%. Therefore, the accuracy of the SVM model in predicting breast cancer is 98.7%.

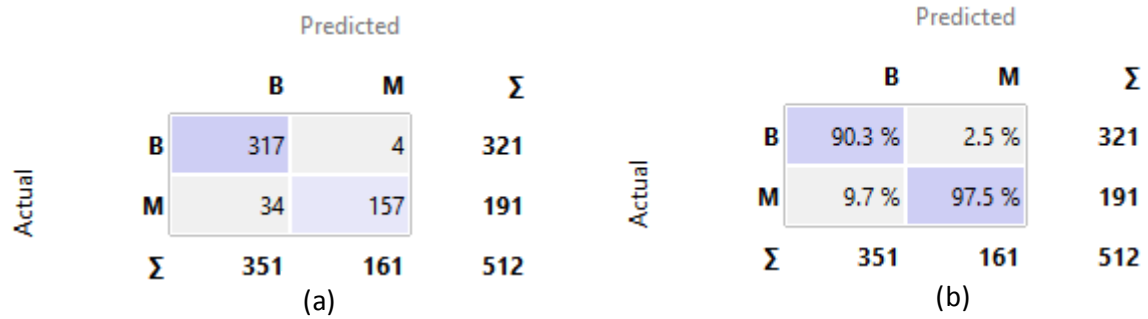


Figure 5. Logistic Regression's Confusion Matrix

The confusion matrix shows the ability of Logistic Regression in predicting benign cancer (B) by 90.3% and against malignant cancer (M) by 97.5%. Hence, the prediction accuracy of Logistic Regression in predicting breast cancer is 93.9%.

From the calculation of the confusion matrix for the three prediction models, the results show that the best prediction accuracy is shown by Random Forest (99.6%), followed by SVM (98.7%), and finally Logistic Regression (93.9%).

The ROC Analysis widget on Orange also shows how Random Forest performs better than SVM and Logistic Regression. Figure 6 shows the ROC Analysis view of the performance of the three prediction models. The closer the curve to the left and the upper limit, or the wider the area formed by the curve at the bottom, the better the prediction performance of the model. Thus, the area under the widest curve is that given by Random Forest, followed by SVM and Logistic Regression.

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

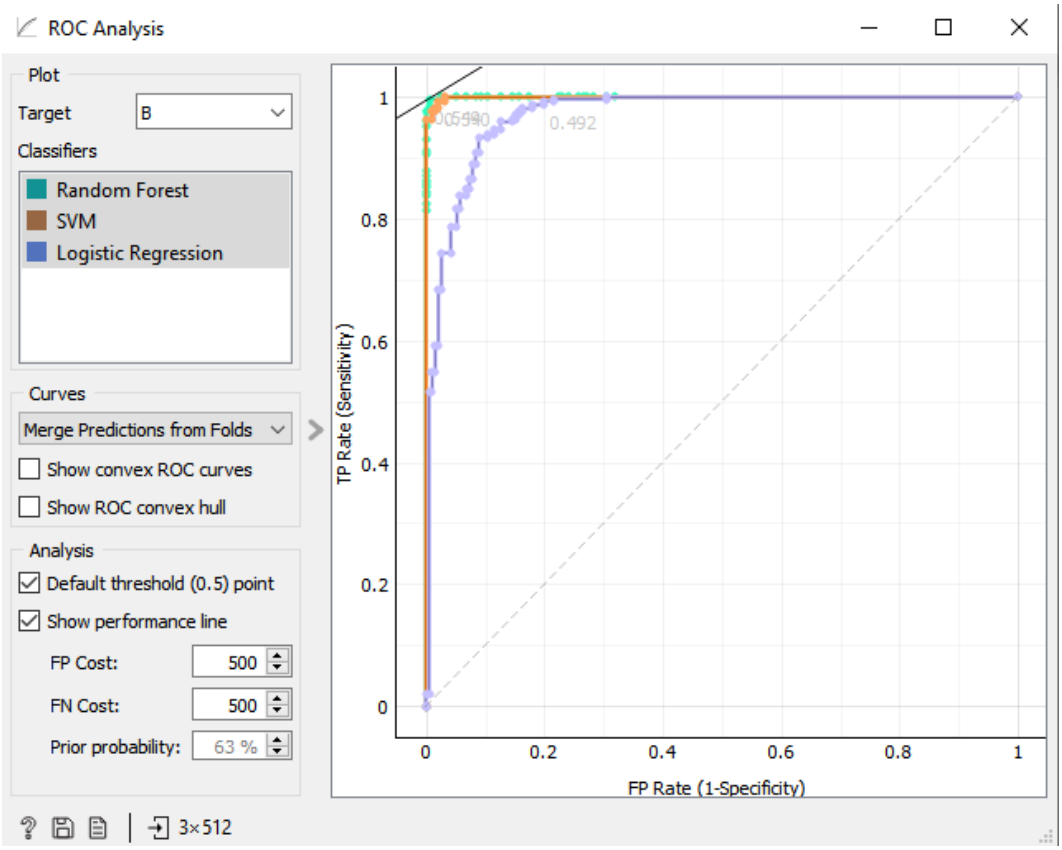


Figure 6. ROC Analysis results for prediction models

Overall, breast cancer prediction models have been successfully obtained through a machine learning approach using BCWD data. The three algorithms tested, namely Random Forest, SVM, and Regression Logistics, showed very good accuracy, i.e. all three were more than 90%. Of the three algorithms, Random Forest has the highest level of accuracy. The predictive performance of this model is superior in sensitivity, specificity, accuracy, and recall (A. Sharma, et al.2017) (A. Osareh and B. Shadgar, 2010) (H. M and S. M.N, 2015)

The use of cross validation techniques at the modelling stage showed the satisfactory results. So far, this technique is very reliable to test the generalizability of predictive models, as well as their ability to overcome the problem of overfitting. In this case, the selection of 10 levels of validation is the decision that gives the best results in most cases (D. Berrar, 2018).

The results obtained from this experiment at the same time strengthen and complement a number of similar studies with the same dataset that has been done previously (M. Amrane, S. et al. 2018), (E. A. Bayrak, P. Kirci, and T. Ensari, 2019) (H. Asri, et al. 2016) (A. F. M. Agarap, 2018). However, these four studies did not include Random Forest in their experiments. In addition, the tools used make more choices on WEKA. The same kind of research that includes Random Forest in the experiment is one written by (Y. Li, 2018). The difference is that this study uses two data sets, namely BCWD and BCCD. The result also supports Random Forest with an accuracy of 96.1%.

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

5. Conclusions

The results showed the accuracy of the three predictive models of breast cancer are above 92%. This value makes the three models worthy of consideration in predicting new data, as well as for further improvement and for the other implementation. To enrich the choices, it is necessary to continue this experiment by including other algorithms which are also expected to have a strong classification performance in predicting breast cancer.

6. REFERENCES

M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, (2018) "Breast cancer classification using machine learning," *Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2018*, pp. 1–4, 2018, doi: 10.1109/EBBT.2018.8391453.

A. R. Vaka, B. Soni, and S. R. K., (2020) "Breast cancer detection by leveraging Machine Learning," *ICT Express*, vol. 6, no. 4, pp. 320–324, 2020, doi: 10.1016/j.icte.2020.04.009.

E. A. Bayrak, P. Kirci, and T. Ensari, (2019) "Comparison of machine learning methods for breast cancer diagnosis," *Sci. Meet. Electr. Biomed. Eng. Comput. Sci. EBBT 2019*, pp. 4–6, 2019, doi: 10.1109/EBBT.2019.8741990.

H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, (2016) "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, doi: 10.1016/j.procs.2016.04.224.

Y. Li, (2018) "Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction," *Appl. Comput. Math.*, vol. 7, no. 4, p. 212, doi: 10.11648/j.acm.20180704.15.

O. I. Obaid, M. A. Mohammed, M. K. Abd Ghani, S. A. Mostafa, and F. T. Al-Dhief, (2018) "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer," *Int. J. Eng. Technol.*, vol. 7, no. 4.36 Special Issue 36, pp. 160–166, 2018, doi: 10.14419/ijet.v7i4.36.23737.

J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, (2019) "Machine learning in medicine: a practical introduction," *BMC Med. Res. Methodol.*, vol. 19, no. 1, pp. 1–18, doi: 10.1186/s12874-019-0681-4.

A. F. M. Agarap, (2018), "On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset," *ACM Int. Conf. Proceeding Ser.*, no. 1, pp. 5–9, doi: 10.1145/3184066.3184080.

M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, (2019), "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–17, doi: 10.1186/s12911-019-0801-4.

A. Sharma, S. Kulshrestha, and S. Daniel, (2017) "Machine learning approaches for breast cancer diagnosis and prognosis," *Int. Conf. Soft Comput. its Eng. Appl. Harnessing Soft Comput. Tech. Smart Better World, icSoftComp 2017*, vol. 2018-Janua, pp. 1–5, 2018, doi:

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University

10.1109/ICSOFTCOMP.2017.8280082.

A. Osareh and B. Shadgar, (2010) "Machine learning techniques to diagnose breast cancer," *2010 5th Int. Symp. Heal. Informatics Bioinformatics, HIBIT 2010*, pp. 114–120, 2010, doi: 10.1109/HIBIT.2010.5478895.

H. M and S. M.N, (2015), "A Review on Evaluation Metrics for Data Classification Evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

D. Berrar, (2018) "Cross-validation," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978-0-12-809633-8.20349-X.

Submission: 16 August 2021
Acceptance: 23 August 2021
©INTI International University