Classification Breast Cancer Revisited with Machine Learning

Hanna Arini Parhusip [a,1,*], Bambang Susanto [a,2], Lilik Linawati [a,3],Suryasatriya Trihandaru [a,4], Yohanes Sardjono [b,5], Adella Septiana Mugirahayu [b,6]

[a] Mathematics Department, Universitas Kristen Satya Wacana, Jl.Diponegoro 52-60, Salatiga, 50711, Indonesia
[b] BATAN , Jl. Babarsari, Tambak Bayan, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281, Indonesia
[1] hanna.parhusip@uksw.edu; [2]bambang.susanto@uksw.edu; [3]lilik.linawati@uksw.edu; [4]suryasatriya@uksw.edu; [5]sardjono.batan@gmail.com; [6]662017002@student.uksw.edu
* corresponding author

ARTICLE INFO

Article history
Received April 21, 2020
Revised April 30, 2020
Accepted May 10, 2020

Keywords:
machine learning
accuracy
breast cancer

ABSTRACT

The article presents the study of several machine learning algorithms that are used to study breast cancer data with 33 features from 569 samples. The purpose of this research is to investigate the best algorithm for classification of breast cancer. The data may have different scales with different large range one to the other features and hence the data are transformed before the data are classified. The used classification methods in machine learning are logistic regression, k-nearest neighbor, Naive bayes classifier, support vector machine, decision tree and random forest algorithm. The original data and the transformed data are classified with size of data test is 0.3. The SVM and Naive Bayes algorithms have no improvement of accuracy with random forest gives the best accuracy among all. Therefore the size of data test is reduced to 0.25 leading to improve all algorithms in transformed data classifications. However, random forest algorithm still gives the best accuracy.

1. Introduction

Study on breast cancer has been developed in the last 30 years historically from education and support to partnership in scientific research in USA [1] to improve awareness and social support which are also done in other countries in the world. In Indonesia, the research on breast cancer is also growing since this disease is the second cause mortality for mostly women in Indonesia. The research group in Science and Mathematics Faculty has involved in the breast cancer study which namely learning and investigating the used of Boron Neutron Capture Therapy since 2014. One result was measuring the internal dose of radiation for workers at Boron Neutron Capture Therapy (BNCT) facility based on Cyclotron 30 MeV with BSA The internal dose exposed to the radiation worker is 9.08E-9 µSv [2]. The boron accumulation in EMT-6 tumors of different sizes was evaluated and resulted that in larger tumors (approximately 400 mg) boron accumulated but was significantly higher in smaller tumors (approximately 100 mg) [3]. There have been many other results in the study of Boron Neutron Capture Therapy in the group,such as [4][5] which have no study on classification data. Since later on classification will be needed to improve the reseach of using BNCT facility to treat breat cancer, the article here focus on classification on big data which may appear in the measurement.

Classification on big data and the related algorithms are now rapidly growing since development of big data has been supported by existence of “Open Data” provided by various organizations and made available to citizens and businesses [6]. Data from medical field for instance breast cancer has been several times were discussed by many researchers and the research on this topic is still developed since breast cancer (BC) is one of the most common cancers among women worldwide [7].

This paper employed 7 Machine Learning algorithms using data set breast cancer provided by UCI Machine Learning Repository available in internet where tha data of breast cancer has been classified into 2 classes,i.e. Malignant (M) and Benign (B) breast cancer. The used algorithms are Logistic Regression,Nearest Neighbor, Support Vector Machine (SVM), Kernel SVM, Naive Bayes, Decision Tree, dan Random Forest Classification (RFC). There have been several journals discussed those algorithms and applied to the case of breast cancer [6] [8] [9] [10]. However those journals have not presented all algorithms in one report as shown here. Using these classifications, the paper here proposes accuracy of each algorithm.

2. Machine Learning Method

2.1. Data Identification

There are 30 features in the data in 3 cathegories obtained from 569 persons who are 212 persons Benign (B) breast cancer and 357 persons are Malignant Cancer (M) .The features are collected into 3 classes,i.e. Feature mean, Feature standard error (se), Feature Worst taken from internet created by University of Wisconsin, Clinical Sciences Center in 1995 (Web 1).

Additionally, 2 types of text data in one column to identify the patients belong to, dataset Malignant (M) and Benign (B) breast cancer. The given data are visualized by using Kde Plot and Violin plot to distiquised 2 chategorised more eaily as shown in Figure 1. Figure 1 illustrates the distribution of all features for both categories of breast cancer. Concave_point mean, concavity mean, area mean, and perimeter mean have indicated different distributions for both breast cancers compared to the other features. This conclusion is also depicted in Figure 2. By drawing in different sides (left and right), we can distiquish the distributions more clearly. The results showed that the texture_mean, median of Malignant Breast Cancer (M) and Benign Breast cancer is most likely separated which indicated that it is good for classification. However, fractal_dimension_mean has unclear separation leading to bad feature for classification. With these initial study, we employ each algorithm to the given data.

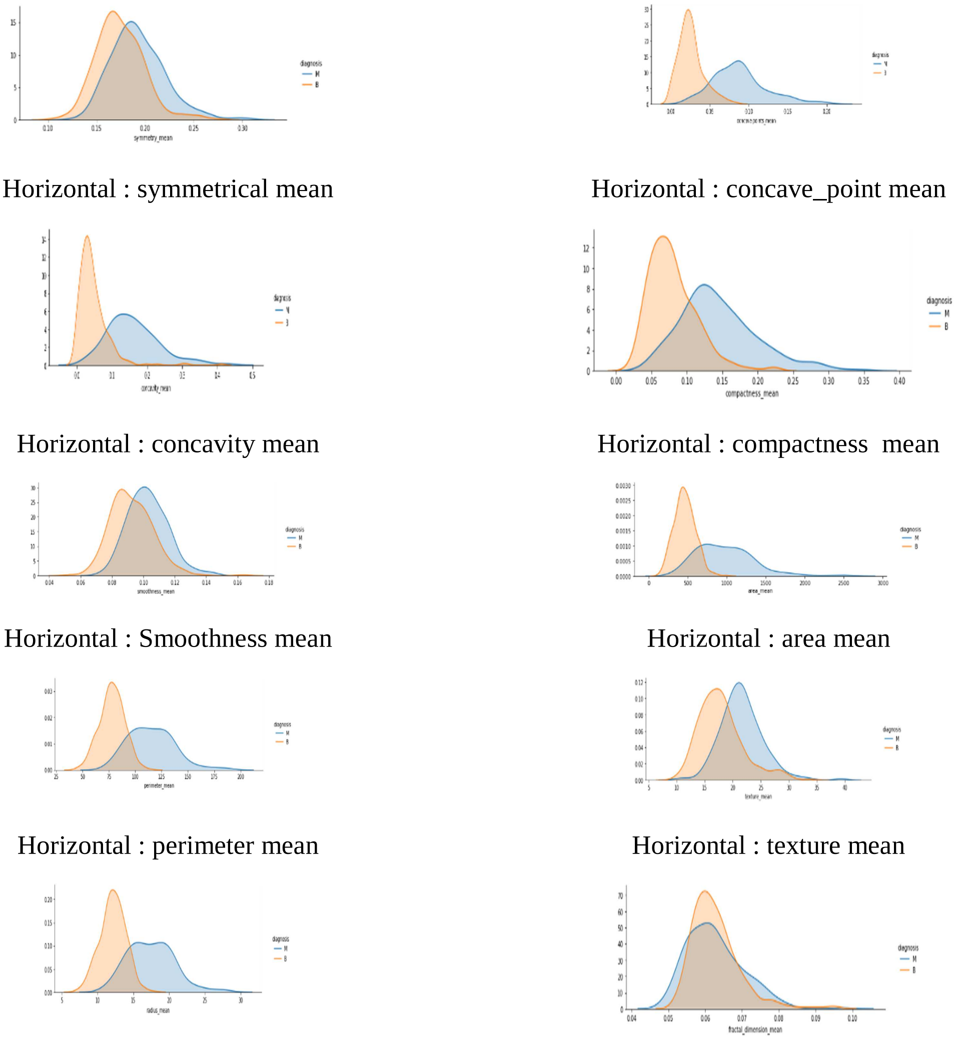
2.2. Summary of the Used Algorithms

2.2.1. Logistic Regression

Logistic regression is one of classification algorithms to relate a feature as an input with a discrite output with a certain probability limited to values between 0 and 1. . This logistic regression is similar to Ordinary Least Squares (OLS) regression where target variable has dicotomy scales such as yes and no, good and bad, low and high. The curve is given by natural algorithm of target variable. The logistic regression can be formulated as $\frac{\mathrm{p}}{1-\mathrm{p}} = \exp\left(\beta_0 + \beta_1 \mathrm{X}\right)$. By taking the natural logarithm of both side,one yields $\ln \frac{\mathrm{p}}{1-\mathrm{p}} = \beta_0 + \beta_1 \mathrm{X}$. Since we have several features variables, we may write

$$p = \frac{1}{1+ e^{-\left(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p\right)}} \tag{1}$$

The parameters are determined uses maximum likelihood estimation (MLE).



Horizontal : symmetrical mean

Horizontal : concave_point mean

Horizontal : concavity mean

Horizontal : compactness mean

Horizontal : Smoothness mean

Horizontal : area mean

Horizontal : perimeter mean

Horizontal : texture mean

Horizontal : radius mean

Horizontal : fractal dimension mean

Fig 1. Kde plot of data distribution

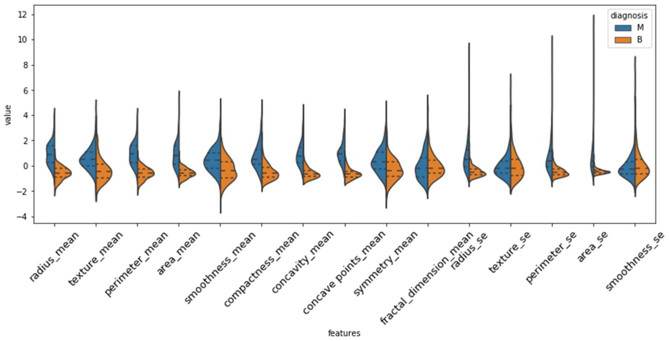


Fig 2. Distribution of Malignant Breast Cancer (M) and Benign Breast cancer for each feature.

2.2.2. Support Vector Machine

Support Vector Machine (SVM) is one of methods in supervised learning,i.e the data set have been labeled. This algorithm separates data into different classes by hyperplane in linear case or a curve in a nonlinear case by maximing the margin leading to optimization problem. The optimal separating hyperplane or other more general functions in multidimensional cases are defined into a special coding called a kernel that can be set up depending on the given problem. Additionally, data can also contain outliers. Therefore the algorithm adds some regularization parameters. This algorithm has been tested for breast cancer data in several literatures [11][10] [12] and also for other big data [13]. The paper here shows that the algorithm is used in different sizes of test data with different kernels.

2.2.3. K-Nearest Neighbor

Another classification algorithm is the k-nearest-neighbors (K-NN) algorithm and it can be used also for regression. Several authors have used the K-NN for classification and regression [14] It is developed due to some behaviour of data may have similar things in close proximity based on the used distance definition. The given data are splitted into sample data dan query data. We examine the closeness of the sample data dan query data. By sorting the obtained distances from the smallest to the largest , one has a list of distances. Take the first-K of these distances and collect the obtained labels. The most appeared labels are then collected to identify the classes.

2.2.4. Naive Bayes

Naïve Bayes Classifier (NBC) is a method classification based on Bayesian theorem. Naive stands for relying on the assumtion that all features are unrelated in the class though they can be related. All properties are considered to have possible contribution independently. Bayes theorem relies on the event probability from the prior as the assumption knowledge. If we have hypothesis H and given an event E, Bayes states that the relation probability before the event occur as P(H) and probability from hypothesis after the event occur is called P(H|E) and formulated as P(H|E) is a certain proportion of P(H). This proportion is named as likelihood that gives a kind of similarity for a certain probability and the likelihood is given by P(E|H)/P(E). Thus one yields

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{2}$$

where P(H|E) posterior probability. In the computation, one has to define a certain radius to guarantee that some elements in the given dataset on the same class. In this case, the Euclidean distance is used. As usual in the machine learning, data set are splitted into 2,i.e. training data and test data. Prior probability is defined as the number of element in current class divided by the number of the element in all classes. We define also margin probability P(X) to determine the elements in the circle divided by the total elements. Furthermore, the likelihood is then can be prescribed as P(|current class) is equal to the number of elements in the current class divided by the toral elements. Thus the posterior probability in the algorithm is :P(current class|X)= likelihood*prior probability/ margin probability. Though Naive Bayes Classifier has been employed in breast cancer data by other authors [12][9] , this paper revisits the algorithm to learn how the algorithm works compared to other algorithms in Machine Learning. One may find the simple presentation of this method in the level of education [15].

2.2.5. Decision Tree Algorithm

When the atributes are categorical and numeric then decision tree algorithm has ability to classify with these type of data. The algorithm constructs terminal Nodes, Recursive Splitting, Building a Tree. The terminal nodes contain the maximum tree depth and the minimum node records. The first node is called the root node of the tree and from the root node other nodes are made until to the maximum tree depth is obtained. There must be minimum nodes satisfied in the algorithm. Thus the number of nodes is growing until the terminal nodes. A node may have zero children (a terminal node), one child (one side makes a prediction directly) or two child nodes. We will refer to the child nodes as left and right in the dictionary representation of a given node.

By recursive splitting the tree is made starting from 2 groups of data,i.e. two lists of rows given the index of an attribute and a split value for that attribute which then the cost function evaluating the split is called the Gini index. The spliting is succesful if the Gini score is 0, and the worst case in splitting will have Gini score 0.5 and the function is called **gini_index()** The rows in the first group all belong to class 0 and the rows in the second group belong to class 1, so it’s a perfect split. Given a dataset, we must check every value on each attribute as a candidate split, evaluate the cost of the split and find the best possible split we could make. Once the best split is found, we can use it as a node in our decision tree.

2.2.6. Random Forest Classification

Random Forest Classification can be considered as a collection of decision trees which is also a supervised learning algorithm. . By averaging out the impact of several decision trees, random forests tend to improve prediction. Randomness is governed by selecting data samples randomly from a given dataset where the best tree will be obtained by voting. The final prediction is taken from the most votes. Compared to decision trees, Random forest has no problem on overfitting with more time consuming due to complexity of trees. To calculate the importance of each feature Gini index is used by droping a variable [8].

The accuracy of each algorithm is checked using Confusion matrix which is frequently used to predict accuracy of classification algorithm and we use the library in **sklearn** to do the computation of accuracy matrix.

3. Result and Discussion

3.1 Correlation studies

All features are studied to have the correlation between 2 different features and we obtained the matrix of correlations.The positive correlations are shown by several pairs of features,i.e. perimeter mean and radius worst, area mean and radius worst, texture mean and texture worst, area worst and radius worst which are shown in Figure 3.

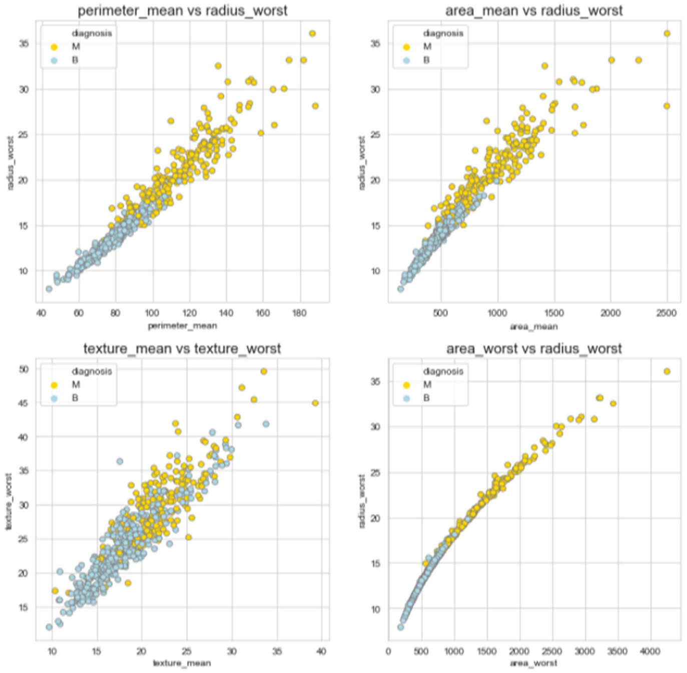


Fig 3. Positive correlation for Perimeter mean vs Radius worst, Area mean vs Radius worst, Texture mean and Texture Worst, Area worst and radius worst.

3.2 Classification Result

The algorithms have been provided in the python library callled sklearn. We employ all the algorithms above for the studied data. As usual in machine learning algorithm, one must split data into training data and test data. For the first study, the test data is 0.3 part of the original data. Additionally, if data should also be transformed to avoid big range between different features.

Using the given algorithm, Table 1 has presented the result of logistic regression for classification for data cancer. The obtained confusion matrix is CM = $\begin{bmatrix} 103 & 5 \\ 1 & 62 \end{bmatrix}$

Table 1. Result on logistic regression

	precision	recall	f1-score	support
B	0.99	0.95	0.97	108
M	0.93	0.98	0.95	63
accuracy			0.96	171
macro avg	0.96	0.97	0.96	171
weighted avg	0.97	0.96	0.97	171

The algorithm of Logistic Regression has been employed. The table shows that the Benign Breast Cancer (C) contains 108 persons where 5 persons are missclasified, and hence the variable recall in the table has defined the prediction has 0.95 value or 95 % correctness. Moreover the Malignant Breast Cancer (M) has 63 perrson with 1 misclassification. Therefore the recall column gives 0.98 ,that is the number of correct predicted divided by the total data. For all data, the algorithm gives 0.965 accuracy or 96.5% accurary. Futhermore, the algorithm is again used where the data are transformed by min-max normalization. . The obtained confusion matrix for the transformed data is as follows,i.e. CM = $\begin{bmatrix} 93 & 15 \\ 0 & 63 \end{bmatrix}$.

Table 2. The result of logistic regression with min-max normalization

	precision	recall	f1-score	support
B	1.00	0.86	0.93	108
M	0.81	1.00	0.89	63
accuracy			0.91	171
macro avg	0.90	0.93	0. 91	171
weighted avg	0.93	0.91	0. 91	171

By normalization Min-Max for the given data, we obtain breast cancer with 108 data of Benign Breast Cancer (C) and 15 persons have been misclassified. The Malignant Breast Cancer (M) contains 63 persons with none misclassification. As a result, we get 91,2 % accuracy.

The k-nearest neighbor algorithm is the second used algorithm. We the same splitting parameters where the test size data is 0.3 , one yields the CM matrix, i.e . CM = $\begin{bmatrix} 103 & 5 \\ 4 & 59 \end{bmatrix}$.

The same as in the previous algorithms, the table has shown the Benign Breast Cancer (C) has 108 data with 5 misclassification and Malignant Breast Cancer (M) contains 63 data with 5 misclassification. Finally, one concludes that total accuracy is 94.7%. For each algorithm, we redo the algorithm where the given data are normalized. Other algorithms are not shown separately. The final studies for all algorithms are listed in Table 3.

Table 3. Comparison accuries given by each algorithm

No	Algorithm	Accuracy using original data	Data transformed using	Accuracy using transformed data
1	Logistic Regression	96,5%	min-max normalization	91,2%
2	Nearest Neighbor	94,7%	min-max normalization	91,8%
3	SVM	63,2%	min-max normalization	94,2%
4	Kernel SVM	63,2%	min-max normalization	94,2%
5	Naive Bayes	92,4%	min-max normalization	78,9%
6	Decision Tree	92,4%	Model Overfitting	94,2%
7	RFC	96,5%	Standard Scale	97,1 %

Table 3 shows that the RFC has given best accuracy to predict breast cancer classification. indicating best accuracy is still obtained. Compared to other authors , some differencies achieved,e.g. logistic regression has 96.5% accuracy in this research where as other author obtained 97.18%. RFC in this reseach has 96,5% and other researcher obtained 95.25% [8]. In the case of KNN, efficiency for different values of K has been shown by other author [16].

Table 4. Comparison accuries given by each algorithm for different test size

No	Algorithm	Accuracy using original data Test size =0.3	Accuracy using transformed data Test size =0.3	Accuracy using transforming data Test size =0.25.
1	Logistic Regression	96,5%	91,2%	95.86 %
2	Nearest Neighbor	94,7%	91,8%	95.10 % (minkownski distance)
3	SVM	63,2%	94,2%	97.20% (linear separation)
4	Kernel SVM	63,2%	94,2%	96,50% (rbf separation)
5	Naive Bayes	92,4%	94,2%	91.60%
6	Decision Tree	92,4%	94,2%	95.80% (criterion=entropy)
7	RFC	96,5%	97,1 %	98.60% (criterion =entropy)

The result of Table 4 is obtained by using test size =0.3,i.e the training data are 70% from the given data. However, transforming data may improve the accuracy as seen in the case of SVM. In the case of KNN, the accuracy is reduced. We expect the accuracy should be better in all algorithms after data are rescaled. We tried to use smaller test size. We show that the test size in transforming data may give different accurary and confusion matrix as listed in Table 4.

Reducing test size has led to improve accurary in all algorithms as we expected that the accuracy should rely on the trasforming data to be classified since we want to have reasonable range of data in all features. Futhermore, one may study the impact for each feature to the algorithm to the result prediction .The result shows that the importantness of features are concave points_worst, radius_worst, and perimeter_worst as the highest three first scores .One may also illustrate in a histogram as shown in Figure 4.

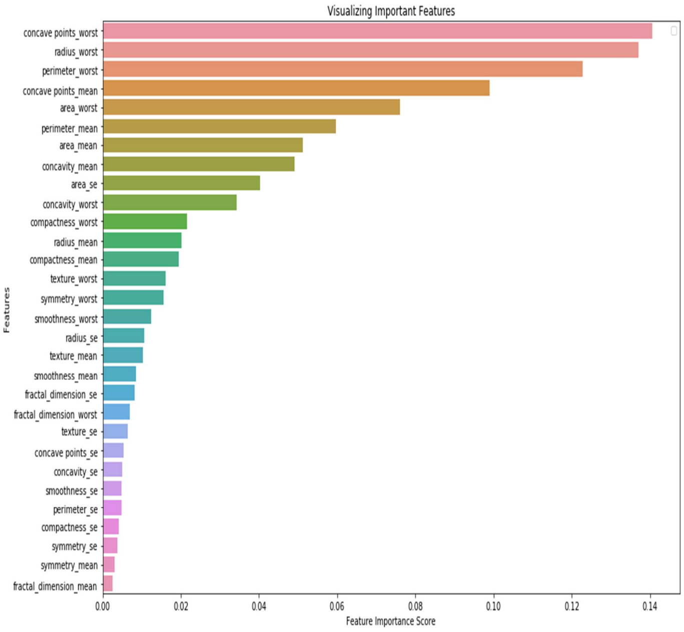


Fig 4. The important levels for features

Compared to other author, the result on Naive Bayes Classifier gave 94.762% using 10 features in the study [9] .This may happen since the features used in the research here are 31 features . Other author stated that accurary was 95.5% in other literature [12] where Naive Bayes Classifier was improved with cross validation.

4. Conclusion

This paper has shown that some algorithms in Machine Learning are tested to use for analysis observing data of breast cancer obtained from UCI Machine Learning Repository available in internet. The algorithms are logistic regression, support vector machine, ,k-nearest neighbourhood, naive bayesian, decision tree and random forest classification. After data are transformed, the 30% of test data give no improvement on accuracy of classification. By reducing 5% test data from 30%, accuracy has been improved and random forest classification gives the best accuracy.

Acknowledgment

The paper here supported by the internal research center from Universitas Kristen Satya Wacana SK no.121/Pen./Rek./3/V/2020 entitled *Klasifikasi kanker dengan metode Clustering dan COTHONS (Coding Python for Schools) period 2019/2020.*

References

[1] R. Janet et al., “A Historical Perspective on Breast Cancer Activism in the United States: From Education and Support to Partnership in Scientific Research,” J Womens Heal., vol. 21, no. 3, pp. 355–362, 2012, doi: 10.1089/jwh.2011.2862.

[2] M. Muhammad, A. W. Harto, and Y. Sardjono, “Monte Carlo N Particle Extended (MCNPX) Radiation Shield Modelling on Boron Neutron Capture Therapy Facility Using D-D Neutron Generator,” vol. 4, no. 2, pp. 58–65, 2019, [Online]. Available: https://ejournal.uksw.edu/ijpna/article/view/3318/1404.

[3] A. A. Khan, C. Maitz, C. Quanyu, and F. Hawthorne, “BNCT induced immunomodulatory effects contribute to mammary tumor inhibition,” PLoS One, vol. 14, no. 9, pp. 1–14, 2019, doi: 10.1371/journal.pone.0222022.

[4] S. Dyah, P. Bagaswoto, and S. Yohannes, “In Vitro and In Vivo Test of Boron Delivery Agent for BNCT,” Indones. J. Phys. Nucl. Appl., vol. 4, no. 2, 2019, doi: https://doi.org/10.24246/ijpna.v4i2.39-44.

[5] S. G. Pinasti, “Measurement of Yttrium-90 Biodistribution in Selective Internal Radiation Therapy (SIRT): a Comparison Between PET AND SPECT IMAGING,” vol. 4, no. 2, pp. 45–57, 2019, [Online]. Available: https://ejournal.uksw.edu/ijpna/article/view/2769/1403.

[6] I. Issam, J. Stéphane, N. Karl, and M. Carole, “The Big Data Revolution for Breast Cancer Patients,” Eur J Breast Heal., vol. 14, no. 2, pp. 61–62, 2018, doi: 10.5152/ejbh.2018.0101.

[7] H. G. Russnes, O. C. Lingjærde, A. L. Børresen-Dale, and C. Caldas, “Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters,” Am. J. Pathol., vol. 187, no. 10, pp. 2152–2162, 2017, doi: 10.1016/j.ajpath.2017.04.022.

[8] S. Jabeen and K. Jilani Abdul, “Predicting Breast Cancer using Logistic Regression and Multi-Class Classifiers,” Int. J. Eng. Technol., vol. 7, no. 4.20, pp. 22–26, 2018.

[9] R. Megha and S. Arun Kumar, “Breast Cancer Prediction using Naïve Bayes Classifier,” Int. J. Inf. Technol. Syst., vol. 1, no. 2, pp. 77–80, 2012, [Online]. Available: https://www.researchgate.net/publication/308934053_Breast_Cancer_Prediction_using_Naive_Bayes_Classifier.

[10] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[11] Moh. Yamin Darsyah, “Menakar Tingkat Akurasi Support Vector Machine Study Kasus Kanker Payudara,” Stat. Univ. Muhammadiyah, vol. 1, no. 1, pp. 15–20, 2013.

[12] Kathija and N. Shajun, “Breast Cancer Data Classification Using SVM and Naïve Bayes Techniques,” Int. J. Innov. Res. Comput. Commun. Eng., vol. 3297, no. 6, pp. 11449–11455, 2016, doi: 10.15680/IJIRCCE.2016. 0412129.

[13] L. Demidova, E. Nikulchev, and Y. Sokolova, “Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 5, pp. 294–312, 2016, doi: 10.14569/ijacsa.2016.070541.

[14] B. Sadegh, Imandoust, and B. Mohammad, “Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background,” Int. J. Eng. Res. Appl., vol. 3, no. 5, pp. 605–610, 2013, [Online]. Available: https://wwijera.com/papers/Vol3_issue5/DI35605610.pdf.

[15] M. . Bárcena, M. A. Garín, A. Martín, A. . Tusell, and E. Unzueta, “A Web Simulator to Assist in the Teaching of Bayes’ Theorem,” J. Stat. Educ., vol. 27, no. 2, 2019, doi: https://doi.org/10.1080/10691898.2019.1608875.

[16] Z. Shichao, L. Xuelong, Z. Ming, Z. Xiaofeng, and W. Ruili, “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” IEEE Trans. Neural Networks Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2018, doi: 10.1109/TNNLS.2017.2673241.