Comparative Study of Breast Cancer Diagnosis using Data Mining Classification

Yopie Noor Hantoro
Faculty of Computer Science and Information Technology
Gunadarma University
Depok, Indonesia

Abstract—**Breast Cancer is often suffered by women and it is an enemy to millions of women all over the world. The most important strategy to prevent death is to do early detection of the breast cancer and provide modern treatments. Along with the development of medical technology and information technology, various methods have been developed to detect the presence of breast cancer, one of which is the machine learning classification technique. In this study, performance comparison is conducted on three machine learning algorithm i.e. Multilayer Perception (MLP), Random Forest (RF) and Support Vector Machine (SVM). The dataset is sourced from Wisconsin Breast Cancer Diagnostic (WBCD). The performance comparison is evaluated by measuring accuracy, precision and recall values. Result of this study confirm that, using the k-fold cross validation technique, the MLP algorithm has the highest performance.**

Keywords— Data Mining, Multilayer Perception (MLP), Random Forest (RF), Support Vector Machine (SVM), Breast Cancer

I. INTRODUCTION

Cancer, kind of Non-communicable disease (NCD), is the leading cause of death and the single most important barrier to increasing life expectancy in every country of the world [1]. In according GLOBOCAN 2018 report, female breast cancer is most frequent in terms of new cases in the majority (154 countries) of countries. Fig. 1 show the most commonly diagnosed cancer and leading causes of cancer death in female. World Health Organization (WHO) has stated that the breast cancer is a type of cancer that is often suffered by women and it is an enemy to millions of women all over the world, but the positive trend is that the death rate is gradually decreasing after 1990 due to awareness, screening early detection, and continuous improvement in treatment [2].

The organs and tissues of the body consist of cells [3]. These cells largely repair and reproduce themselves in the same way [3]. Normally, cells divide regularly and are controlled [3]. But these process gets out of control for some reason and the cells continue to divide and develop into lumps called tumors [3]. Breast tumors usually grow due to overgrowth of cells lining the breast ducts [3]. It is very important to conduct an accurate diagnosis of tumors. Most tumors are the result of benign (non-cancerous), but it will cause serious problems if a malignant tumor is diagnosed [4]. Breast cancer.org reported that stages of the breast cancer depends on the size and type of tumor and amount the tumor cells have been penetrated in the breast tissues [5]. The most important strategy to prevent death is to do early detection of the breast cancer and provide modern treatments. It will be easier to treat early, small, and non-spreading breast cancer.

The most reliable way to detect breast cancer early is by having regular screening tests. Breast cancer diagnosis have the main role i.e. to distinguish between the Malignant and Benign breast masses, while the prognosis estimates recurrence of disease, predicts survival of patient, and helps in establishing a treatment plan by predicting the outcome of a disease [6]. According to breast cancer diagnosis, doctors will propose different treatment projects for therapy [7].
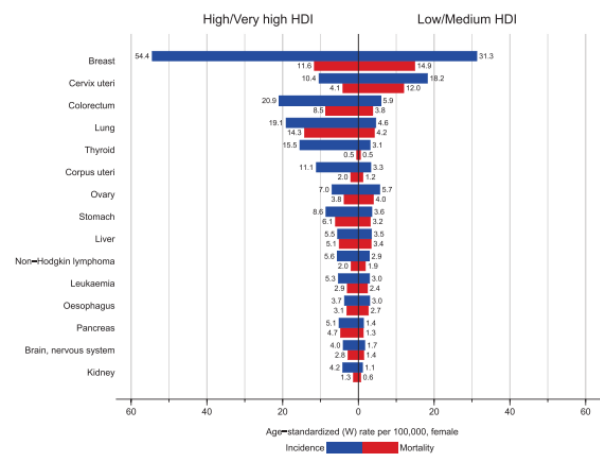

Fig. 1 Incidence and Mortality Age-Standardized Rates in High/Very-High Human Development Index (HDI) Regions Versus Low/Medium HDI Regions Among Women in 2018 [1]

Data mining as sub-field of Information and Communication Technologies can play potential role in breast detection. Data mining approaches are applied to medical science topics rise rapidly due to high performance in prediction, reducing costs, promoting patients' health, improving healthcare value and quality and make real time decision to save people's lives [8]. Data mining is one of data analysis techniques to uncover previously undetected relationships among data items [9]. Data Mining techniques may be classified in two categories: (1) machine learning techniques which are based on artificial intelligence techniques such as artificial decision/regression trees, neural networks and case-based reasoning; and (2) non-machine learning techniques which are mainly based on statistical techniques such as principal component analysis and linear regression [6]. There are various major data mining techniques i.e. classification, clustering, and association rules [10]. Classification technique assigns items in a collection to target categories or classes. Some of the classification methods are Bayesian Network, Artificial Neural Network, Rule-Based

Classification, Decision Tree, Associative Classification, Support Vector Machine, Genetic Algorithm, Rough Set Approach, Fuzzy Set Classification and K-Nearest Neighbors [10]. Clustering technique make a group of object based on their characteristics and aggregate them according to their similarities. Some of the clustering methods are Fuzzy Clustering, K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Expected Maximization (EM) [10]. Association technique discovers new relations between variables and focuses on finding frequent patterns among database [10].

Today, medical data recording has increased rapidly. Medical data, collected from examinations, measurements, prescriptions, etc., are stored in databases continuously [3]. Traditional methods can't analyze and search for interesting patterns and information that is hidden in this enormous amount of data [3]. Therefore, data mining techniques are needed to find interesting patterns and hidden knowledge of these data sources. This data mining techniques can predict the risk of cancer, may play a important role in the diagnosis process and effective prevention strategies or play a role in cancer screening [11].

II. RELATED WORK

Several studies have been conducted on the detection of breast cancer. These studies have compared several methods to achieve high classification accuracy. Some of previous studies are given in the following:

A. Al Bataineh [12] compare the performance criterion on five nonlinear machine learning algorithms viz Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Classification and Regression Trees (CART), Gaussian Nave Bayes (NB) and Support Vector Machines (SVM) to find the best classifier in Wisconsin Breast Cancer Dataset. His research result shows that MLP model has the highest performance in terms of accuracy, precision, and recall of 99.12%, 99.00%, and 99.00% respectively.

K. Rajendran, M. Jayabalan, V. T. An, and V. Sivakumar [13] has conduct feasibility study on data mining techniques in diagnosis of breast cancer. They have reviewed a lot of paper to provide a holistic view of the types of data mining techniques used in prediction of breast cancer. The result shows that the data mining techniques that are commonly used include Decision Tree, Naïve Bayes, Association rule, Multilayer Perceptron (MLP), Random Forest, and Support Vector Machines (SVM). The overall performance of the techniques differ for every dataset. On Wisconsin Breast Cancer Dataset, random forest classifier produced better performance with accuracy 99.82%.

M. K. Keles [14] has conduct comparative study on breast cancer prediction and detection using data mining classification. He run and compare all the data mining classification algorithms in Weka tool against an antenna dataset. His comparative result shows that random forest algorithm become the most successful algorithm with 92.2% accuracy rate.

Deneshkumar, Manoprabha and Senthamarai [15] have predicted breast cancer using five prediction algorithms i.e. Naive Bayes, Logistic regression, Decision tree, Random forest and Support vector Machine. The prediction was done on Wisconsin breast cancer dataset. The result shows that, without any feature selection, support vector machine as the best algorithm with an accuracy of about 95.6%. While logistic regression showed a better performance compared to other algorithms with feature selection, which was nearly 97% [15].

III. MATERIALS AND METHODS

One of the important applications of data mining is Classification [16]. In this paper, we used Multilayer Perception (MLP), Random Forest, and Support Vector Machine (SVM) techniques for classification of breast cancer. These classification algorithms are selected because they are the best model in the research mentioned above. Therefore it will have potential to yield good results. RapidMiner tool is used to evaluate the performance of these techniques. These methods are applied using the 10-fold validation technique

A. 10-fold cross validation

To perform training and testing data, the predictive performance of the models are assessed using k-fold cross-validation. This method makes every experimental design data point gets to be in a test subpart and train subpart; this helps to prevent overfitting on training data [17]. This method create partition on data into k equally (or nearly equally) sized segments or folds. k iterations of training and validation are done such that within each iteration a different data fold is held-out for validation while the k - 1 folds are used for training [18]. In this study, 10-fold cross validation is used. The dataset is splitted into 10 parts, 9 parts for training and 1 part for testing, then repeated for all combinations. Fig. 2 Shown how k-fold cross validation work on dataset.
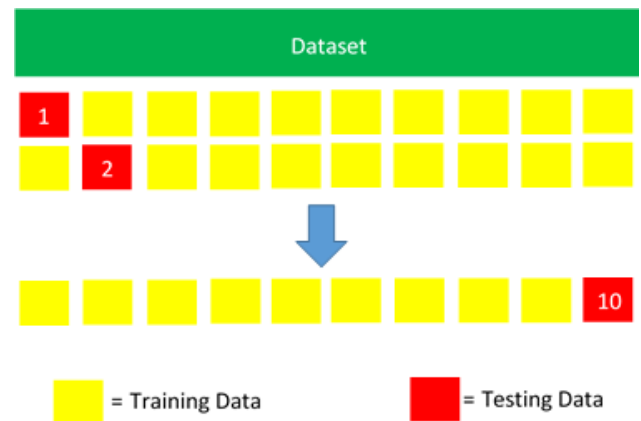

Fig. 2 10-Fold Cross Validation (regenerated from [12])

B. Breast Cancer Wisconsin Dataset

The data used in this study are Wisconsin Breast Cancer Dataset (WBCD). This dataset have been collected by Dr. William H. Wolberg at University of Wisconsin. It obtained from UCI Machine Learning Repository. The feature of dataset are evaluated from a digitized image of a fine needle aspirate (FNA) of breast mass. They give characteristics of the cell nuclei present in the image. This dataset has 699 instances, 2 classes (benign and malignant) and 9 integer-valuated attributes (see Table 1). Value 10 describe the most abnormal condition. We do data cleansing by removing 16 instances with missing value. Therefore, we have a new

dataset with 683 instances. Class distribution of new dataset: benign = 444 (65.01%), malignant = 239 (34.99%). See Fig. 3 for detail class and attributes distribution.

Table 1 Wisconsin Breast Cancer Dataset

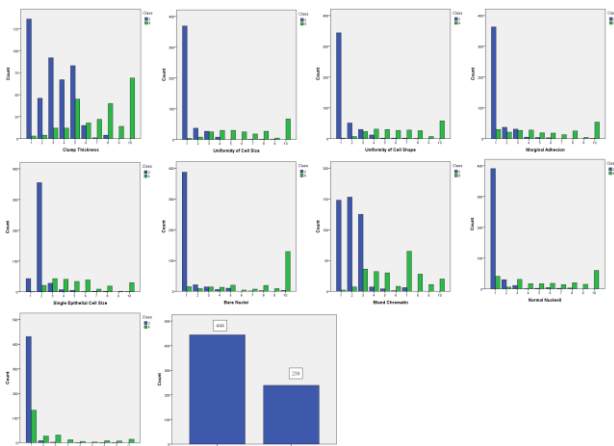| No	Attribute	Value
1	Sample code number	Sample code number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2 = benign, 4 = malignant

Fig. 3 Class and attribute distribution

C. Multilayer Perceptron (MLP)

80% of ANNs researches focused on using Multilayer Perceptron (MLP) [19]. Multilayer Perceptron (MLP) has one or more hidden layers along with the input and output layers [20]. The neurons are arranged in layers, their connections are directed from lower layers to upper layers, the neurons are not interconnected in the same layer [21]. See Fig. 4 for the structure of a multilayer perceptron. An artificial neuron calculate the weighted sum from its input and then activate function to get a signal that will be transmitted to the next neuron [22].

MLP use back-propagation training. It minimize the error between the correct output value and the target value by adjusting the weight values that are calculated from input-output mappings and [12]. This method iteratively calculates the weight value using the gradient descent algorithm. The weight of the output vector is computed by the gradient descent rule by following equation:

$$\Delta w_{j} = \alpha \, y_j \, (\, 1 - y_j \,) \, (\, t_j - y_j \,) \, x_j \qquad (1)$$

where α represent the training level, $y_j \, (\, 1 - y_j \,)$ represent a derivative of the activation function and $(\, t_j - y_j \,)$ represent an error with t_j as a target [12].

The activation function in the MLP classification can be described as the following:

Input Layer: there is no activation function

Hidden Layer: using the following simplified sigmoid function:

$$\left(\varphi(x) = \frac{1}{1 + e^{-x}} \right) \qquad (2)$$

Output Layer: using the following sigmoid function:

$$\varphi_0(x) = \varphi(x) \qquad (3)$$

or hyperbolic tangent as follows :

$$\varphi_0(x) = \tanh(x/2) = \frac{1 - e^{-x}}{1 + e^{-x}} \qquad (4)$$
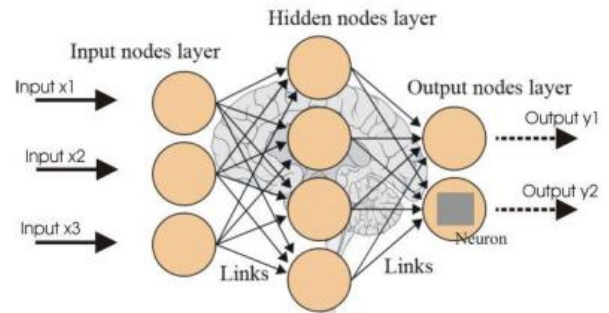

Fig. 4 Multi Layer Perceptron [20]

D. Random Forest (RF)

Random forest, based on decision trees and combined with aggregation and bootstrap ideas, was first introduced by Breiman [23]. He explains a method for creating a unrelated trees forest using a Classification and Regression Treelike (CART-like) procedure that is combined with randomized node optimization and bagging [14]. It applies two mechanisms i.e. build an unsemble of trees via bagging with replacement (bootstrap) and select features at each tree node randomly [24]. The first mechanisms means that the selected training set can be selected again [24] and each tree is grown using the obtained bootstrap sample [24]. The second one makes random selection of a small fraction of features and further separation using the best feature from this set [24]. It provides excellent performance on a number of practical problems, mainly because it is not sensitive to noise in the data set, and prevents overfitting [25]. During classification process, this algorithm use more than one decision tree to find the classification value. Random Forest Algorithm divides each node by using the best variable among the randomly selected variables in each node, they don't divide each node into branches by using the best branch among all the variables [14].

The random forest algorithm is as follows: k indicates the number of decision tree in the random forest, n suggests the number of training data-set sample that each decision tree corresponds to, M represents the feature number of sample, m refers to the number of features when carrying out segmentation on a single node of a decision tree, m<<M [26]:
(1) Take sample for N times from all the training data-set with a repeated sampling method, from k group training set (namely bootstrap sampling). Each constructs a decision tree for each training set, the sample not selected from k group data out of bag [26];

(2) For each node of the decision tree, choose m features randomly based on this node, and calculate the best segmentation characteristics according to the m characteristics [26];

(3) Every decision tree truly grows without pruning [26];

(4) Shape a random forest model from several decision trees and identify and classify unknown data by using the model [26].

E. Support Vector Machine (SVM)

Support Vector Machines (SVM) is a learning model that is widely used with related learning schemes, which are used for classification and regression analysis [27]. Basically, the way SVM works is to function as a linear separator between two data points to identify two different classes in a multidimensional environment [28]. This method makes a line to separate the two classes by determining a linear classifier. This separation is called the optimal hyperplane separator [12]. This is selected from the set of hyperplanes to classify patterns that maximizes the hyperplane margin i.e. the distance from the hyperplane to the closest point of each patterns [29]. SVM construct two parallel hyperplanes on both side of the optimal hyperplane separator. An assumption is built that the greater the margin or distance between these parallel hyperplanes will be the better the generalization error of the classifier [30].

Fig. 5 is a simple model of SVM technique. There are two different patterns and the purpose of SVM is to separate these two patterns. This model consists of three different lines. The margin of separation or marginal line is represented as the line w.x-b = 0. There are two line on both sides of the margin line i.e. lines w.x - b = 1 and w.x - b = -1. Hyper plane built by these three lines that separate the given pattern and the pattern located at the edge of the hyper plane is named the support vector. The perpendicular distance between the edge of the hyper plane and the margin line is called as the margin. One of the goals of SVM is to maximize this margin to get better classification. The classification process gets better with a larger margin and hence minimizes the occurrence of errors.
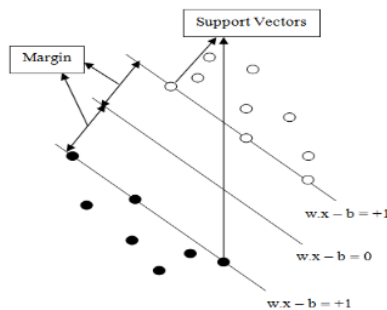

Fig. 5 SVM Model [29]

IV. RESULTS AND DISCUSSION

In this section an evaluation and comparison of each algorithm is performed based on the measurement of accuracy, precision and recall.

A. Accuracy

Accuracy is the overall correctness of the model and is calculated from the ratio of the number of samples predicted correctly divided by the total number of samples in the data set. Mathematically, accuracy can be written as follows:

$$Accuracy = \frac{x}{n} \times 100\% \qquad (5)$$

x refer to the number of samples correctly predicted, and n is the total number of samples in the dataset. In terms positives and negatives, accuracy can be calculated as follows:

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \times 100\% \qquad (6)$$

Where TN = True Negative (Prediction is malignant, and actual output is also malignant), TP = True Positive (Prediction is benign, and actual output is also benign), FN = False Negative (Prediction is malignant, and actual output is benign), FP = False Positive (prediction is benign, and actual output is malignant). Table 2 shown accuracy result of compared methods

Table 2 Accuracy Result

Algorithm	Accuracy (%)
Multilayer Perceptron (MLP)	95.96
Support Vector Machine (SVM)	95.26
Random Forest (RF)	95.61

We can see that the MLP algorithm has the highest accuracy value that is equal to 95.96%.

B. Precision

Precision, or the positive predictive value, is the ratio of correctly predicted positive observations to the total predicted positive observations. It can be described as follows:

$$Precision = \frac{TP}{TP + FP} \times 100\% \qquad (7)$$
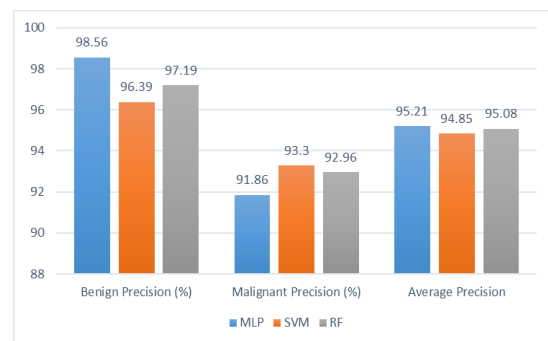
Fig. 6 shown precision result of compared methods:

Fig. 6 Precision Result

We can see that the MLP algorithm has the highest average precision that is equal to 95.21%.

C. Recall

Recall, also called sensitivity, is the ratio of correctly predicted positive observations to the all observations in actual class. It can be described as follow:

$$Recall = \frac{TP}{TP + FN} \times 100\% \qquad (8)$$

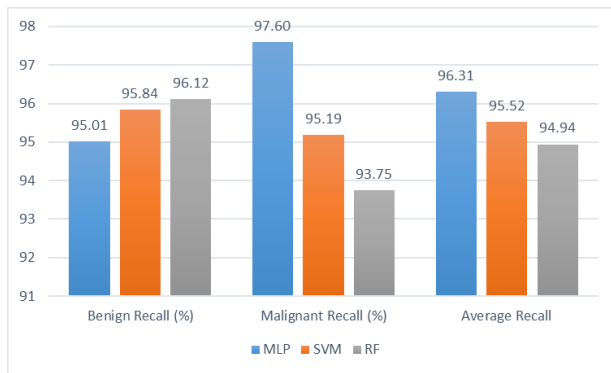Fig. 7 shown recall result of compared methods:

Fig. 7 Recall Result

We can see that the MLP algorithm has the highest average recall that is equal to 96.31%.

Overall, Fig. 8 shows the performance of the three algorithms
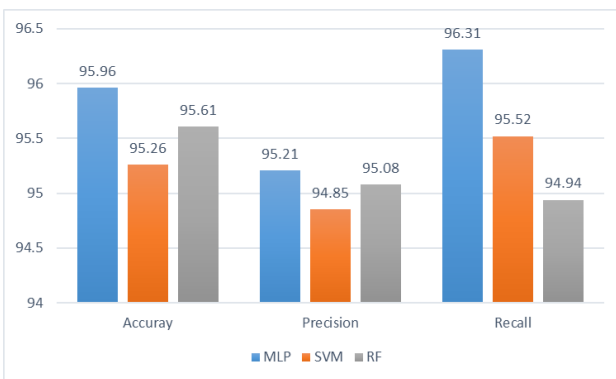

Fig. 8 Performance Result

V. CONCLUSION

The use of analytical data in the world of health to predict and diagnose breast cancer is increasingly interesting to develop. Mortality from breast cancer can be reduced through early detection. Through research in recent years, shows that machine learning techniques play an important role in diagnosing breast cancer. In this paper, three of populer machine learning techniques are applied for breast cancer drtection.i.e. MLP, SVM and Random Forest. Wisconsin Breast Cancer Diagnostic (WBCD) dataset is used to compare the performance of the proposed techniques. Result of this study confirm that, using the k-fold cross validation technique, the MLP algorithm has the highest performance in terms of accuracy, precision and recall of 95.96%, 95.21% and 96.31% respectively.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 68, no. 6, p. 394-424, 2018, doi: 10.3322/caac.21492.

[2] A. K. Dubey, U. Gupta, and Sonal Jain, "Breast cancer statistics and prediction methodology: a systematic review and analysis," *Asian Pacific J. Cancer Prev.*, vol. 16, no. 10, p. 4237-4245, 2015, doi: 10.7314/apjcp.2015.16.10.4237.

[3] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithm.*

[4] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthc.*, vol. 8, no. 2, p. 111, 2020, doi: 10.3390/healthcare8020111.

[5] M. Akram, M. Iqbal, U. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," *Biol Res*, vol. 50, no. 1, p. 33, 2017, doi: 10.1186/s40659-017-0140-9.

[6] A. Idri, I. Chlioui, and B. El Ouassif, "A systematic map of data analytics in breast cancer," in *Proceedings of the Australasian Computer Science Week Multiconference*, 2018, p. 10, doi: 10.1145/3167918.3167930.

[7] Y. Li and Z. Chen, "Performance evaluation of machine learning methods for breast cancer prediction," *Appl. Comput. Math.*, vol. 7, no. 4, pp. 212–216, 2018, doi: 10.11648/j.acm.20180704.15.

[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] B.Padmapriya and T.Velmurugan, "Classification algorithm based analysis of breast cancer data," *Int. J. Data Min. Tech. Appl.*, vol. 5, no. 1, pp. 43–49, 2016, doi: 10.20894/IJDMTA.102.005.001.010.

[10] R. Ghorbani and R. Ghousi, "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction," *Int. J. Data Netw. Sci.*, vol. 3, pp. 47–70, 2019, doi: 10.5267/j.ijdns.2019.1.003.

[11] A. Atashi, S. Sohrabi, and A. Dadashi, "Applying two computational classification methods to predict the risk of breast cancer: A comparative study," *Multidiscip. Cancer Investig.*, vol. 2, no. 2, pp. 8–13, 2018, doi: 10.30699/acadpub.mci.2.2.8.

[12] A. Al Bataineh, "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 248–254, 2019, doi: 10.18178/ijmlc.2019.9.3.794.

[13] K. Rajendran, M. Jayabalan, V. T. An, and V. Sivakumar, "Feasibility study on data mining techniques in diagnosis of breast cancer," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 328–333, 2019, doi: 10.18178/ijmlc.2019.9.3.806.

[14] M. K. Keles, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *Teh. Vjesn.*, vol. 26, no. 1, pp. 149–155, 2019, doi: 10.17559/TV-20180417102943.

[15] D. V, M. M, and S. K. K, "Comparison of datamining techniques for prediction of breast cancer," *Int. J. Sci. Technol. Res.*, vol. 8, no. 8, pp. 1331–1337, 2019, [Online]. Available: https://www.ijstr.org/final-print/aug2019/Comparison-Of-Datamining-Techniques-For-Prediction-Of-Breast-Cancer.pdf.

[16] A. K. Shrivas and A. Singh, "Classification of breast cancer diseases using data mining techniques," *Int. J. Eng. Sci. Invent.*, vol. 5, no. 12, pp. 62–65, 2016, [Online]. Available: http://www.ijesi.org/papers/Vol(5)12/version-2/J05120206265.pdf.

[17] K. Srinivasan, A. K. Cherukuri, D. R. Vincent, A. Garg, and B.-Y. Chen, "An efficient implementation of artificial neural networks with k-fold cross-validation for process optimization," *J. Internet Technol.*, vol. 20, no. 4, pp. 1213–1225, 2019, doi: 10.3966/160792642019072004020.

[18] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross-validation," in *Encyclopedia of database systems*, Springer US, 2009, pp. 532–538.

[19] K. Ncibi, T. Sadraoui, M. Faycel, and A. Djenina, "Multilayer perceptron artificial neural networks based a preprocessing and hybrid optimization task for data mining and classification," *Int. J. Econom. Financ. Manag.*, vol. 5, no. 1, pp. 12–21, 2017, doi: 10.12691/ijefm-5-1-3.

[20] S. J. Livingston, B. S. T. Selvi, M. Thabeetha, C. P. Grena, and C. S. Jenifer, "A neural network based approach for sentimental analysis on amazon product reviews," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6S, pp. 469–473, 2019, [Online]. Available: https://www.ijitee.org/wp-content/uploads/papers/v8i6s/F60970486S19.pdf.

[21] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: architecture optimization and training," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 4, no. 1, pp. 26–30, 2016, doi: 10.9781/ijimai.2016.415.

[22] W. Castro, J. Oblitas, R. Santa-Cruz, and H. Avila-George, "Multilayer perceptron architecture optimization using parallel computing techniques," *PLoS One*, vol. 12, no. 12, pp. 1–17, 2017, doi: 10.1371/journal.pone.0189369.

[23] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, and N. Vialaneix, "Random

forests for big data," *Big Data Res.*, vol. 9, pp. 28–46, 2017, doi: 10.1016/j.bdr.2017.07.003.

[24] N.Venkatesan and G.Priya, "A study of random forest algorithm with implemetation using weka," *Int. J. Innov. Res. Comput. Sci. Eng.*, vol. 1, no. 6, p. 2015, 2015, [Online]. Available: http://www.ioirp.com/Doc/IJIRCSE/i6/JCSE242.pdf.

[25] P. R. Patil and S. Kinariwala, "Automated diagnosis of heart disease using random forest algorithm," *Int. J. Adv. Res. Ideas Innov. Technol.*, vol. 3, no. 2, pp. 579–589, 2017, [Online]. Available: https://www.ijariit.com/manuscripts/v3i2/V3I2-1197.pdf.

[26] Q. Ren, H. Cheng, and H. Han, "Research on machine learning framework based on random forest algorithm," in *AIP Conference Proceedings*, 2017, p. 080020, doi: 10.1063/1.4977376.

[27] F.Leenavinmalar and A.Kumarkombaiya, "Application of data mining techniques in early detection of breast cancer," *Int. J. Eng. Trends Technol.*, vol. 56, no. 1, pp. 43–45, 2018, doi: 10.14445/22315381/IJETT-V56P208.

[28] P. Janardhanan, L. Heena, and F. Sabika, "Effectiveness of support vector machines in medical data mining," *J. Commun. Softw. Syst.*, vol. 11, no. 1, pp. 25–30, 2015, doi: 10.24138/jcomss.v11i1.114.

[29] A. Pradhan, "Support vector machine-A survey," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, 2012.

[30] H. Bhavsar and M. H. Panchal, "A review on support vector machine for data classification," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 10, pp. 185–189, 2012, [Online]. Available: http://ijarcet.org/wp-content/uploads/IJARCET-VOL-1-ISSUE-10-185-189.pdf.