Classification Algorithm Analysis for Breast Cancer

Arief Agus Sukmandhani[1*], *Lukas*[2†], *Yaya* Heryadi[1‡], *Wayan* Suparta[3§], *Antoni* Wibowo[1**]

[1]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
[2]Cognitive Engineering Research Group (CERG), Faculty of Engineering, Universitas Katolik Indonesia Atmajaya, Indonesia
[3]Department of Electrical Engineering, Institut Technologi Nasional Yogyakarta, Yogyakarta, Indonesia 55281

Abstract. Breast cancer in women is a type of disease that is the main cause of death in women according to world breast cancer data. Therefore, early detection of breasts is needed significantly to improve life. If a woman has been identified, then rehabilitation and treatment on an incentive basis are needed to reduce the worse. This study used a dataset collected by the University of Wisconsin Hospitals, Madison (https://atapdata.ai/). This research conducted experiments using several data mining classification strategies to predict breast cancer using machine learning algorithms. The Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naive Bayes, Random Forest, Decision Tree, Deep Learning (H2O), and Neural Network are used to classify algorithms. From these algorithms' classification, we compare accuracy, best classification, and compare algorithm performance with curve ROC (RapidMiner Studio Core) to see which performance algorithm has the best quality for classification. From the analysis results, the deep learning algorithm with Tanh and Exprectifier activation function has a good accuracy of 93.14%, and the best classification with 89.62%. In addition, deep learning has found the best quality from the ROC curve results on the dataset used in this research.

1 Introduction

Cancer is a disease that causes the most deaths worldwide. According to data from the Center for Data and Information, in 2012, there were around 8.2 million deaths[1]. The types of cancer that cause the most deaths include lung, liver, stomach, colorectal, and breast cancer. WHO data states that an estimated 84 million people died from cancer from 2005 to 2015 [2]. This disease is characterized by abnormalities in the cell cycle that lead to the ability of these cells to cause cell division beyond normal limits. The consequences of cell division are attacking nearby biological tissues and can migrate to body tissues through blood circulation [3]; [4].

Based on WHO data, breast cancer is one type of cancer that attacks women. According to Anggorowati, it is estimated that 8-9% of women have breast cancer[5]. Until now, one of the standard methods of treatment for people with cancer is surgery and chemotherapy. It's just that the treatment carried out can give less significant results if the severity level reaches the tau tolerance limit, which is usually referred to as reaching the final stage. Therefore, early detection of this disease is needed so that the bad things that occur can be minimized [6]. Many women do not realize whether she has breast cancer or not. The essential thing that can be done is to do a manual check by detecting whether there are abnormal lumps found around the breast area. If found, more incentives, it need to be done to identify whether it includes breast cancer or not. Further checks are required with the help of a doctor, but many people sometimes ignore this so that it becomes more severe in the future.

Researchers see several studies using the cancer breast Wisconsin dataset (WBC) classification of Machine Learning Algorithms. Sharma's research uses RapidMiner tools with Naïve Bayes (NB) algorithm, K-Nearest Neighbor (KNN), Logistics Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Xgboost (XG), Decision Tree and the accuracy obtained by Xgboost is 98.24% while in SVM the accuracy obtained is 96.49% [7]. Research on Badr and Abou uses python with genetic algorithm and cuckoo search algorithm with 84% accuracy [8]. Suthagar's research using Python tools with a Classification Algorithm of AdaBoost, Kneighbor Classifier, DT Classifier, RF Classifier, SVM Classifier, Classifier Logistics Regression, Gaussian NB Classifier, and Gradient Improve Classifier with SVM Classifier has gained the highest accuracy of 99.1% [9]. Monirujjaman's research uses Python with Random Forest, Logistics Regression, Decision Tree, and K-Nearest Neighbor algorithms with 98% accuracy [10]. Tounsi's research uses python with SVM and KNN algorithms [11]. Deva's research uses

---

[*] Corresponding author: arief.sukmandhani@binus.ac.id
[†] Corresponding author: lukas@atmajaya.ac.id
[‡] Corresponding author: yayaheryadi@binus.edu
[§] Corresponding author: wayan@itny.ac.id
[**] Corresponding author: anwibowo@binus.edu

python using a linear algorithm of Discriminant Analysis (LDA) and Principal Component Analysis (PCA) which is commonly used for feature extraction, where Naive Bayes Classifier (NBC), Random Forest (RF), Neural Networks (NNS) and Support Vector Machine SVM) with LDA and NN produce the best results in terms of accuracy, sensitivity, specificity, and kappa coefficient [12]. Priyadarshini's research uses python with the Logistics Regression Algorithm, Random Forest Tree, Supervised Vector Machine (SVM), and KNN [13]. Bensouchauha's research using Mathlab using the Decision Tree (DT) algorithm, Discriminated Analysis (DA), Naive Bayes (NB), Support Vector Machine (SVM), K Nearest Neighbor (KNN), Ensemble Methods (EMS), and Multi-Layer Perceptron (MLP) with the highest accuracy in the SVM and MLP algorithms with 96.52%accuracy [14]. Ilhan's research uses python with deep learning algorithms, and the accuracy is 98.42% [15]. Research on Mekha and Teeyukaet using rapidminer with Deep Learning Algorithm, Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), Vote (DT+NB+SVM), Random Forest (RF), and AdiBoost with the highest accuracy on Deep Learning Expectifier Activation Function Algorithm [16]. Wassim's research uses a rapidminer with the K Nearest Neighbors (K-NN) algorithm, Naive Bayes (NB), and the Support Vector Machine (SVM) with the best accuracy owned by SVM with a size of 97.07% [17].

Data mining can be applied in various fields, for example in the health sector to help doctors extract relevant information and help to make decisions [18]. In data mining, we can classify data using algorithms and they are many types of algorithms for classification. Before determining the algorithm to use, we need to analyze which algorithm has the best accuracy and classification. In this study, researchers conducted an experiment to carry out the classification process by comparing several algorithms found in machine learning to find out whether the patient had breast cancer or not using a dataset sourced from the University of Wisconsin Hospital, Madison by Dr. William H. Wolberg uploaded to the source dataset (https://atapdata.ai/). In this experimental research, the researcher tries using several existing algorithms in machine learning. The algorithm is processing with ten times trying to get stable results for the analysis.

2 Literature Review

Classification is the separation or sorting of objects into classes[19]. Classification has been applied in many fields such as medical, astronomy, commerce, biology, media, etc. In the classification algorithm there are two phases that are carried out, among others: the first phase is the algorithm trying to find a model for the class attribute as a function of other variables from the data set. the next phase, applies the previously designed model into a new dataset to determine the associated class of each record [19, 20].

SVM (Support Vector Machine) is a classification model that can be applied to detect cancer [21]. In this model, all samples are grouped into vectors and then distributed in classes and separated by making a linear classification function [22]. This classifier functions by maximizing the distance between the decision hyperplane and the nearest data point commonly called the finite distance [23]. Support Vector Machine(SVM) algorithm has the advantages of complete theory, global optimization, strong adaptability, and good generalization ability because it is on the basis of Statistical Learning Theory (SLT) [24].

The Naive Bayes method or commonly called Gaussian Naive Bayes, is a method that is easy to make without repeated parameter estimates and is used to predict the probability of a data set to be part of a class using Bayes' rules [18, 25]. The advantage of using Naive Bayes is that this method only requires a small amount of training data (Training Data) to determine the parameter estimates needed in the classification process. Naive Bayes often performs much better in most complex real-world situations than one might expect [26].

The K-Nearest Neighbor method is an algorithm that can classify objects depending on most of the k closest neighbors. This algorithm is a non-parametric technique because the classification of test data points depends on the closest training data point without considering the data point parameters. The accuracy of this model can increase with the increase in the number of nearest neighbors [25]. The k-NN algorithm is a method that uses a supervised algorithm. The purpose of the k-NN algorithm is to classify new objects based on attributes and training samples. Where the results of the new test sample are classified based on most of the categories in the k-NN. In the classification process, this algorithm does not use any model to match and is only based on memory [20, 26].

In a random forest, features are selected at random in each decision split. Correlation between trees is reduced by randomly selecting features that can increase predictive power and result in higher efficiency. The advantage of using this model is that it can overcome the problem of overfitting [27, 28].

The neural network is a data processing model inspired by biological neural processes, such as the brain. This model has a unique structure for data processing systems because it consists of a network of artificial neurons (nodes) [18, 29]. Each processing data are interconnected (nodes), operating simultaneously to answer problems. Through learning methods, these methods are configured for specific uses, such as pattern identification or data analysis. Studying biological processes requires adjusting existing synaptic associations between neurons [29].

Decision tree is a model that applies the type of classification or regression in the form of a tree diagram. this model divides the dataset into smaller subsets, and at the same time each related subset is developed incrementally to get a decision tree.[18]. The advantage of using a decision tree in classifying data is that it is easy to understand and interpret [18]. However, decision trees have drawbacks such as requiring that the target

attribute will only have discrete values and decision trees using the "divide and conquer" method, they tend to perform well if there are a few highly relevant attributes, but less so if there are many complex interactions [30, 31].

Cross-validation is an additional method of data mining technique that aims to obtain maximum accuracy results. This method is often also called k-fold cross-validation, where the experiment is k times for one model with the same parameters.[32].

The confusion matrix analyzes how well the classifier recognizes tuples from different classes [33], The Confusion Matrix contains tables to represent the fulfillment of the classification model, (or "classifier") on the test data set. The performance of the method is measured by utilizing the data in the model [34] see table 1.

Table 1. Confusion Metrix

	Predicted	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Classifier accuracy is a standard of how reliable the performance of an algorithm is [34] and An evaluation method based on precision (number 1), recall (number 2) and F-Size (number 3) [34].

$$(1) \quad Accurac \quad (2) \qquad\qquad (3)$$

$$\frac{TP+TN}{TP+FN+FP} \quad F - Measure \qquad Precision$$

$$= \frac{2 \; x \; Precision \; x \; R}{Precision + Rec} = \frac{TP}{TP + FP}$$

$$Recall$$
$$= Sensitivi$$
$$= \frac{TP}{TP + FN}$$

The ROC curve measures the Area Under Curve (AUC). The ROC curve divides positive results between the y-axis and negative results on the x-axis so that the larger the area under the curve, the better the prediction results. The Receiver Operating Characteristics (ROC) curve is used to evaluate the classifier's accuracy and compare different classification models [35], so the larger the area under the curve, the better the prediction results (see table 2).

Table 2. Quality Classifier [36]

No	Accuracy Range	Quality Classifier
1	0.90 – 1.00	Excellent
2	0.80 – 0.90	Good
3	0.70 – 0.80	Fair
4	0.60 – 0.70	Poor
5	0.50 – 0.60	Failure

3 Methodology

The authors used a dataset collected by the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg which has been uploaded to the source dataset (https://atapdata.ai/) which contains data on patients who are identified as having breast cancer or not having

breast cancer. From the data set presented, there are five relevant features considered with a total of 569 data with the distribution of diagnosed as many as 357 samples and undiagnosed 212 samples. This dataset is the result of the diagnosis of breast cancer on x-ray when an abnormal lump is found. The following is a display of the data collected (see table 3).

Table 3. Dataset Description

No	Name	Type
1	mean_radius (mean of distances from center to points on the perimeter)	Integer
2	mean_texture (standard deviation of gray-scale values)	Real
3	mean_perimeter (core tumor's size)	Real
4	mean_area (inside the boundary of core tumor)	Real
5	mean_smoothness (local variation in radius lengths)	Real

Figure 1 shows the details of the experimental work carried out by the researcher.

Fig. 1. Propose work for Research

3.1 Tools and Design

In this study, the researcher used a support tool called RapidMiner which was registered with the student version. Researchers use RapidMiner because it has a variety of descriptive and predictive techniques to provide users with insights so they can make the best decisions it also provides a data mining process consisting of operators that can be nested, described in XML and built with a GUI. RapidMiner has approximately 500 data mining operators, including input, output, data preprocessing, and visualization operators [37]. The following is the design of the experiment carried out (see Figure 2).

Fig. 2. Design on RapidMiner

3.2 Preprocessing

In this section, the researcher identifies data to be measured so that it can produce optimal classification performance for each data mining algorithm used, including labeling and data type. For labeling, the researcher performs labeling on the attributes of the dataset used so that the system can identify and determine the class of labels that will be used as the size target and for data type validation, after the data is labeled, it is necessary to pay attention to the use of the data type used because it can affect the algorithm used. at this stage the researcher determines the target validation and the type of data used so that the measurement can be maximized while the attributes that have been changed are the "diagnosis" attribute (see Figure 3 and Figuire 4).
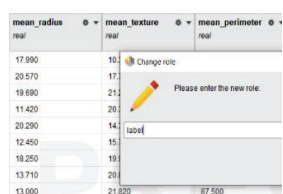
Fig. 3. Validate class for labeling
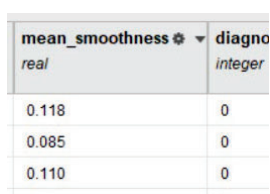
Fig. 4. Validate data type

3.3 Process

Implementation of breast cancer classification using seven data mining algorithms on RapidMiner 9.10 using seven operators that have been provided in RapidMiner, The operator is select attribute, remove duplicate, set role, multiply, cross-validation, apply model, and compare ROCs. Select Attribute, this operator provides various types of filters that can be used in the measurement so that we can choose whether the Attribute subset of the ExampleSet (dataset) is used or not used. Remove Duplicates is used to remove data that is in a duplicated dataset by comparing all instances with each other based on the specified attribute so that only one of all duplicate instances is saved. In the set role operator is used to determine the class attribute or label of the data used as a reference. In it, there is an "attribute name" parameter; this parameter is to determine the data attribute that will be used as a class or label; in this breast cancer data, the class or label is the Dataset attribute, namely "Diagnosis". in the "target role" parameter we can choose a label, this label indicates that we specify the attribute is a class label (see figure 5).
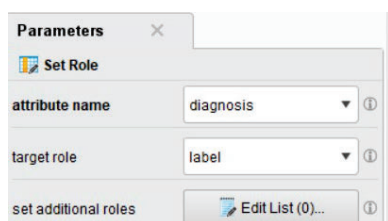

Fig. 5. Setup parameter Set Role

Multiply operator used to connect to several operators used in the design so that they can be processed simultaneously. Cross-Validation operator to maximizes the accurate performance of the algorithm used. In this study, seven different data mining algorithms are used and placed on each cross-validation operator. Not only that, but this operator also contains the apply model operator, which is used to implement the training data model to the testing data on the dataset used, and the performance operator is used to evaluate the algorithm. The performance operator produces data in the form of a confusion matrix and its accuracy. Operator apply model to read the data to be estimated based on the data that has been studied previously and operator compare ROCs to display ROCs curves based on the process performance of each algorithm in the dataset used. In this study, the operator Compare ROC is given 7 algorithm operators whose performance will be measured.

3.4 Result

After designing the scheme on RapidMiner, then data processing is carried out so that the accuracy results of the seven algorithms used for the classification of breast cancer can be known. The results of the accuracy shown in this study after running the Runtime process ten times in each algorithm used. In the deep learning (DL) algorithm, the researcher used H2O version 3.30.0.1. The researcher uses the Tanh, Rectifier, Maxout and Exprectifier activation function which are implemented in RapidMiner. The configuration carried out in this algorithm is to use an activation rectifier because it allows the deep learning model to consider non-linearity and the effects of certain interactions on the dataset used and epoch 10, which is to see that the entire dataset has been lost. Through the training process for 10 rounds. In the process of repeating the process with the same dataset for 10 running times, this algorithm undergoes several changes in accuracy. The highest accuracy was obtained at 94.19% in the model used. However, the researcher took the results below (see table 4) because the Rectifier and Exprectifier models showed the same accuracy results, which was 93.14% in the interval of experiments carried out (see table 4).

Table 4. Confusion metric Deep learning

		True 0 (No)	True 1 (Yes)	Class Precision
Pred (No)	0	190	17	91.79%
Pred (Yes)	1	22	340	93.92%
Class Recall		89.62%	95.24%	

In the Support Vector Machine (SVM) algorithm, the configuration uses the LibSVM function. in the LibSVM function the formula used is CSupport Vector Classifier (C-SVC) with a Radial Basis Function (RBF) kernel with

high accuracy. the accuracy of this algorithm is 88.59% (see table 5).

Table 5. Confusion metric SVM

	True 0 (No)	True 1 (Yes)	Class Precision
Pred 0 (No)	162	15	91.53%
Pred 1 (Yes)	50	342	87.24%
Class Recall	76.42%	95.80%	

Naïve bayes algorithm produces an accuracy of 90.52% (see table 6)

Table 6. Confusion metric Naïve Bayes

	True 0 (No)	True 1 (Yes)	Class Precision
Pred 0 (No)	169	11	93.89%
Pred 1 (Yes)	43	346	88.95%
Class Recall	79.72%	96.92%	

For k-nearest neighbor (KNN) algorithm, the researcher configures using clusters is "k = 5" with mixedmeasures type measure types with the concept for numeric values, Euclidean distance is calculated and at nominal value, distance 0 is taken if both values are the same and distance one is taken otherwise. Accuracy this algorithm is 88.93% (see table 7)

Table 7. Confusion metric KNN

	True 0 (No)	True 1 (Yes)	Class Precision
Pred 0 (No)	167	18	90.27%
Pred 1 (Yes)	45	339	88.28%
Class Recall	78.77%	94.96%	

Random Forest (RF) algorithm, the researcher configures number of tree "100", with the criterion "gain_ratio" so that the variance of information acquisition adjusts the information gain for each attribute so that there is breadth and uniformity of Attribute values. Accuracy this algorithm is 92.26% (see table 8).

Table 8. Confusion metric Random Forest

	True 0 (No)	True 1 (Yes)	Class Precision
Pred 0 (No)	186	18	91.18%
Pred 1 (Yes)	26	339	92.88%
Class Recall	87.74%	94.96%	

Neural Network (NN) algorithm, in this algorithm, the researcher configures using "2" hidden layers with 200 training cycles, 0.01 learning rate and "0.9"

momentum. Accuracy this algorithm is 92.97% (see table 9). the results obtained are as follows:

Table 9. Confusion metric Neural Network

	True 0 (No)	True 1 (Yes)	Class Precision
Pred 0 (No)	189	17	91.75%
Pred 1 (Yes)	23	340	93.66%
Class Recall	89.15%	95.24%	

The researcher conducted several trials by increasing the number of hidden layers from 1 layer with 2-6 neurons to 2 layers with 3-4 neurons each. The results can increase the accuracy of the algorithm's performance but the recall quality decreases so that the classification performance is low.

Decision Tree (DT) algorithm, for this configuration is the same as the configuration in the random forest algorithm with the number of trees "100", criterion "gain_ratio". Accuracy this algorithm is 90.50% (see table 10)

Table 10. Confusion metric Decision Tree

	True 0 (No)	True 1 (Yes)	Class Precision
Pred 0 (No)	187	29	86.57%
Pred 1 (Yes)	25	328	92.92%
Class Recall	88.21%	91.88%	

Table 11 shows the results of measuring accuracy, precision and recall from the seven algorithms mentioned above, namely:

Table 11. Algorithm Measurement Results

Algorithm	Accuracy	Precision	Recall	F-Measure
DL	93,14%	91.79%	89.62%	0.91
SVM	88,59%	91.53%	76.41%	0.83
Bayes	90,52%	93.89%	79.71%	0.86
KNN	88.93%	90.27%	78.77%	0.84
RF	92.26%	91.18%	87.74%	0.89
NN	92,97%	91.75%	89.15%	0.90
DT	90.5%	86.57%	88.21%	0.87

Based on the data presented in table 11. The Deep Learning algorithm has the highest accuracy of the other algorithms tested in this study with an accuracy of 93.14%, followed by the neural network algorithm in second place with an accuracy of 92.97%. The recall size of the deep learning algorithm is the largest, with a height of 89.62%; these results indicate that the ability of this algorithm is the best in classifying to diagnose

patients with breast cancer or not, while the F-Measure value in the deep learning algorithm shows the best with a value of 0.91 where the value is close to 1 which indicates that the classification model in the algorithm has good precision and recall.
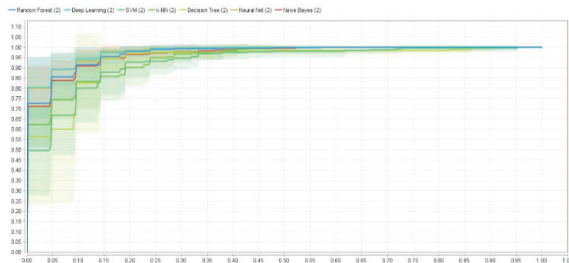

Fig. 6. Curve ROC

If we look at the curve (see figure 6), all the algorithms used touch the number 1.0 on the y-axis, which is the true positive class (TP) for the result. However, in terms of performance improvement from the start of the process the neural network algorithm has a graph that outperforms all of them. This proves that the neural network is an algorithm that has a better classification quality in solving breast cancer classification problems.

4 Conclusion

In this study, seven algorithms were tested to solve the problem of classifying breast cancer or not by using a tool called RapidMiner. This study used a dataset uploaded on the website roofdata.ai sourced from the University of Wisconsin Hospitals, Madison, by Dr. William H. Wolberg, who collected data from patients with identified or not breast cancer. The results of measurements carried out by the deep learning algorithm show the best accuracy and classification model. All the algorithms are used to touch the number 1.0 in the result on the ROC curve. However, in terms of improving performance, neural network algorithms have a graph of improvement that outperforms all of them. This proves that the neural network is an algorithm with better classification quality in solving breast cancer classification problems. Suggestions for further research can be explored for better algorithm performance by using this dataset to produce better accuracy and classification.

References

[1]   Kementrian Kesehatan RI, "Situasi Penyakit Kanker Indonesia," *Pus. Data dan Inf. Kemenkes RI*, no. 2, pp. 31–33, 2015.

[2]   F. Ma'arif and T. Arifin, "Optimasi Fitur Menggunakan Backward Elimination dan Algoritma SVM untuk Klasifikasi Kanker Payudara," *J. Inform.*, vol. 4, no. 1, pp. 46–53, 2017.

[3]   Kementrian Kesehatan RI, "Penyakit Kanker di Indonesia Berada Pada Urutan 8 di Asia Tenggara dan Urutan 23 di Asia," *Januari 31, 2019*, 2019. .

[4]   S. K. Nur Handayani, "Kanker dan Serba-Serbinya (Hari Kanker Sedunia 2022)," *February 4, 2022*, 2022. .

[5]   Anggorowati, "Faktor Risiko Kanker Payudara Wanita," *KESMAS - J. Kesehat. Masy.*, vol. 8, no. 2, pp. 121–126, 2013.

[6]   A. M. Zamani and B. Amaliah, "Implementasi Algoritma Genetika pada Struktur Backpropagation Neural Network untuk Klasifikasi Kanker Payudara," *Tek. ITS*, vol. 1, 2012.

[7]   H. Sharma, P. Singh, and A. Bhardwaj, "Breast Cancer Detection: Comparative Analysis of Machine Learning Classification Techniques," in *2022 International Conference on Emerging Smart Computing and Informatics, ESCI 2022*, 2022.

[8]   Y. A. Badr and A. H. Abou El-Naga, "A Hybrid Metaheuristic Approach for Automatic Clustering of Breast Cancer," in *5th International Conference on Computing and Informatics, ICCI 2022*, 2022, pp. 392 – 399.

[9]   S. Suthagar, C. Snegha, M. Sureka, and S. Velmurugan, "Analysis of Breast Cancer Classification using Various Algorithms," in *Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022*, 2022, pp. 1286 – 1291.

[10]   M. Monirujjaman Khan *et al.*, "Machine Learning Based Comparative Analysis for Breast Cancer Prediction," *J. Healthc. Eng.*, vol. 2022, 2022.

[11]   S. Tounsi, I. F. Kallel, and M. Kallel, "Breast cancer diagnosis using feature selection techniques," in *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology, IRASET 2022*, 2022.

[12]   M. Deva Priya, J. Sangeetha Priya, S. Sam Peter, V. Ravi, S. Karthik, and R. Karthik, "Efficient Machine Learning-Based Diagnosis System for Breast Cancer," in *Lecture Notes in Networks and Systems*, 2022, vol. 341, pp. 695–708.

[13]   A. Priyadarshini and J. Aravinth, "Correlation Based Breast Cancer Detection using Machine Learning," in *2021 6th International Conference on Recent Trends on Electronics, Information, Communication and Technology, RTEICT 2021*, 2021, pp. 499–504.

[14]   S. Bensaoucha, "Breast Cancer Diagnosis Using Optimized Machine Learning Algorithms," in *Proceedings - 2021 IEEE International Conference on Recent Advances in Mathematics and Informatics, ICRAMI 2021*, 2021.

[15]   U. Ilhan, K. Uyar, and E. I. Iseri, "Breast Cancer Classification Using Deep Learning," in *Advances in Intelligent Systems and Computing*, 2021, vol. 1306, pp. 709–714.

[16] P. Mekha and N. Teeyasuksaet, "Deep learning algorithms for predicting breast cancer based on tumor cells," in *ECTI DAMT-NCON 2019 - 4th International Conference on Digital Arts, Media and Technology and 2nd ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering*, 2019, pp. 343–346.

[17] A. Wassim, E. Elarbi, and R. Khadija, "Application of Machine Learning Approaches in Health Care Sector to the Diagnosis of Breast Cancer," in *Journal of Physics: Conference Series*, 2022, vol. 2224, no. 1.

[18] H. Saad and N. Nagarur, "Data Mining Techniques in Predicting Breast Cancer," *J. Appl. Sci.*, vol. 20, no. 3, pp. 124–133, 2020.

[19] G. K. Gupta, *Introduction to Data Mining with Case Studies*, 3th ed. Prentice-Hall Of India Pv, 2014.

[20] A. Ashari, "Performance Comparison between Naïve Bayes,Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.

[21] D. Bazazeh and R. Shubair, "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, 2016, pp. 1–4.

[22] G. Williams, *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery*. 2012.

[23] 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.

[24] Z. Jun, "The Development and Application of Support Vector Machine," *J. Phys. Conf. Ser.*, vol. 1748, no. 5, 2021.

[25] A. Kumar, R. Patra, and A. Ghosh, "Model Selection for Predicting Breast Cancer using Supervised Machine Learning Algorithms," *2020 IEEE Int. Conf. Converg. Eng. ICCE 2020 - Proc.*, pp. 320–324, 2020.

[26] A. Purwanto *et al.*, "Perbandingan Minat Siswa Smu Pada Metode Klasifikasi Menggunakan 5 Algoritma," *IKRAITH-INFORMATIKA*, vol. 2, no. 1, pp. 43–47, 2018.

[27] C. Smith, *Decision Trees and Random Forests: A Visual Introduction for Beginners*. Blue Windmill Media, 2017.

[28] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[29] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*. Springer International Publishing, 2018.

[30] C. C. Aggarwal, *Data Mining: The Textbook*. Springer International Publishing, 2015.

[31] V. Mohan, "Decision Trees : A comparison of various algorithms for building Decision Trees," pp. 1–23, 2013.

[32] B. Santosa and A. Umam, *Data Mining dan Big Data Analytics : Teori dan Implementasi Menggunakan Python \& Apache Spark*. .

[33] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann, 2011.

[34] S. Selvaraj, S. Shajahaan, and M. Chitra, "Application of Data Mining Techniques to Predict Breast Cancer," *Procedia Comput. Sci.*, vol. 163, pp. 11–18, 2019.

[35] T. Patel, D. Patel, and T. Patel, "Breast Cancer Prediction Analysis Using Data Mining Techniques," in *Computer Sciences and Engineering*, vol. Volume 6, 2022, pp. 623–631.

[36] A. P. Ayudhitama and U. Pujianto, "Analisa 4 Algoritma Dalam Klasifikasi Liver Menggunakan Rapidminer," *J. Inform. Polinema*, vol. 6, no. 2, pp. 1–9, 2020.

[37] M. Hofmann and R. Klinkenberg, *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press, 2016.