e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v5i1.9223

Breast Cancer Classification Using Outlier Detection and Variance Inflation Factor

Budi Juarto

Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
budi.juarto@binus.ac.id

*Correspondence: budi.juarto@binus.ac.id

Abstract – *In terms of malignant tumors, breast cancer is one of the most prevalent. Breast cancer is a form of cancer that develops in the breast tissue when the surrounding, healthy breast tissue is overtaken by the uncontrollably growing cells in the breast tissue. Several features or patient conditions can be used in a machine learning approach to predict breast cancer. Machine learning will be utilized in these situations to determine if the cancer is malignant or benign. The Wisconsin Breast Cancer (Diagnostic) Data Set, which contains 32 characteristics and 569 collected data, was the dataset used in this research. Feature selection in this study is done by eliminating outliers using the upper and lower quartile of each feature then feature selection is also carried out on features that have features that have a high variance inflation factor. The machine learning methods used in this research are Logistic Regression, Random Forest, KNN, SVC, XG Boost, Gradient Boosting, and Ridge Classifier. The selection of this method is based on the target that will be predicted by 2 labels, namely benign cancer, and malignant cancer. The result obtained is that the selection of features using the variance inflation factor increases the accuracy of the previous Logistic Regression and Random Forest methods from 98.25% to 99.12%. The method that has the highest level of accuracy is the Logistic Regression and Random Forest methods which have a value of 99.12%. The next research will be developed by trying other optimization techniques for hyperparameter tuning.*

Keywords: *Breast Cancer Detection; Logistic Regression; Random Fores; Classification; Variance Inflation Factor*

I. INTRODUCTION

Breast Cancer is classified as one of the most common malignant tumors. Breast cancer is a type of cancer that occurs and forms in the breast tissue part when the cells in the breast tissue grow uncontrollably and take over the healthy and surrounding breast tissue. [1, 5]. Breast cancer itself is one of the degenerative effects of cells in the tissue mechanism in the breast that divide and grow into cancer. These growths are neoplasms that have an aggressive nature with abnormal growths in excessive amounts, it causes cell tissue in the breast to be damaged. Breast cancer is still a disease with a high mortality rate in women. Based on data from the WHO (World Health Organization), breast cancer has a mortality rate of 42.5% in the world in 2018 with an average number of deaths each year of 9.3 women[3]. For detecting breast cancer previous research has been done by some researchers using Machine Learning and Biosensors. Yash et.al has made some comparative analysis of breast cancer detection using machine learning and biosensors [7]. Many different studies and related articles were reviewed and analyzed systematically that has been reviewed shown that Biosensors and ML both have the potential to detect breast cancer quickly and effectively.

For the literature review, numerous breast cancer studies proposing diverse strategies have already been conducted. S. Nanglia et al. presented the Stacking technique for breast cancer classification using the K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), and ensemble learning algorithms [1]. Using Chi-square, the following characteristics are utilized: Glucose, Resistin, HOMA, Insulin, and BMI. At K =20, the ensemble learning achieves 78% accuracy with a log loss of 0.56.

The research of Huan-Jung Chiu et.al, proposed the principal component analysis method (PCA) and multilayer perceptron network (MLP) method. Experiments carried out using 10-fold cross-validation have an accuracy of 86.97% which shows effective results for checking indications of breast cancer [2].

The research of Amrane et.al, proposed the Breast Cancer Classification using Naïve Bayes and K nearest neighbor (KNN) for predicting cancer with binary label benign cancer or malign cancer [6]. The result shows that KNN gives the highest accuracy with a 97.51% score and the naïve Bayes classifier with a score of 96.19%

Yash Amethiya et al. introduced the AdaBoostM1 Classifier with median absolute deviation (MAD) utilizing nine characteristics, including age, BMI, glucose, insulin, HOMA, leptin, adiponectin, and MCP-1 [4]. MAD normalization gave a classification accuracy of 75% in the identification of breast cancer, while k-means clustering (KMC)-based feature weighting paired with MAD achieved an accuracy of 91.37.

Breast Cancer detection is also carried out with the image dataset by Merinda [8]. The data used in this dataset is histopathological as much as 277,524 data. The method used in this research is Convolutional Neural Network (CNN). The results evaluated in this study were using an accuracy of 80%. In addition, other methods used in detecting breast cancer using images were also carried out by Poonam and Snehal using Mammogram image data [9]. The method used in this study is to use Random Forest with an accuracy rate of 95%.

Research conducted by Yadavendra and Satish was carried out on the breast Histopathology images dataset using 60% training, 20% for validation, and 20% for testing [10]. The method used in this study is to use logistic regression, bagging, voting classifier, and the Xception model. The Xception model has the highest score with precision, recall, and F1 score with a value of 90%.

Mumammed Fatih conducted study comparing various machine learning techniques to breast cancer datasets from the University of Wisconsin Hospital [11]. Methods such as logistic regression, k-nearest neighbors, support vector machine, naive Bayes, decision tree, random forest, and rotation forest are compared. Comparing many machine learning methods, the logistic regression algorithm has the highest classification accuracy at 98.1%.

Gupta, P., & Garg, S conducted research on comparisons using machine learning and also deep learning on breast cancer datasets [12]. The research was conducted using KNN, logistic regression, decision tree, SVM, and random forest. For deep learning algorithm using Adam Gradient Descent Learning. The results obtained highest results are using deep learning using Adam Gradient Descent Learning with an accuracy value of 98.24%.

Most of the previous studies used machine learning and deep learning algorithms in their research for breast cancer datasets. This study will use the Wisconsin Breast cancer (diagnostic) data set [13]. This study will use breast cancer data from Kaggle to predict whether a breast cancer can be categorized as benign or malignant. The study and research was carried out because there were very many cases of breast cancer patients. The purpose of this study, several machine learning methods were carried out by eliminating outliers in the data and also selecting important features using the variance inflation factor (VIF) and evaluated with accuracy and confusion matrix.

II. METHODS

This study began with the collection of breast cancer data sets called Breast Cancer Wisconsin (Diagnostic). Breast Cancer Wisconsin (Diagnostic) Data Set is a well-known dataset for machine learning and data analysis in the field of medical diagnosis. It is frequently used in research and academic settings for developing and testing algorithms for the diagnosis of breast cancer. The dataset was created by Dr. William H. Wolberg from the University of Wisconsin Hospitals, Madison and was made available for public use in the 1990s.

The dataset contains 569 samples of biopsy images of breast masses. Each sample is characterized by 30 features, including the mean radius, texture, perimeter, area, and smoothness of the mass. These features were computed from digitized images of biopsy specimens, and they provide information about the morphological properties of the masses. The goal of using this dataset is to classify the samples as benign or malignant based on these features.

The dataset has been widely used in the development and evaluation of machine learning algorithms for breast cancer diagnosis. It is considered a benchmark for comparing the performance of different algorithms, as it provides a well-defined task with clear evaluation criteria. The dataset has been used to train and test algorithms based on a variety of techniques, including decision trees, neural networks, support vector machines, and ensemble methods.

The Breast Cancer Wisconsin (Diagnostic) Data provides a rich source of data for exploring and developing new techniques for diagnosing breast cancer, and it helps advance the state of the art in this important area. The dataset continues to be widely used and cited in the literature, and it remains a popular choice for developing and evaluating machine learning algorithms for medical diagnosis. The data is obtained from a Kaggle-collected shared dataset. The data comprises of ten characteristics extracted from the cell nucleus, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave spots, symmetry, and fractal dimensions. This study utilized data with two labels: 357 designated benign and 212 classified malignant. This study will focus on the pre-processing of data prior to its incorporation into machine learning algorithms. Python version 3.7.14, which runs on Google Colab, is the programming language utilized in this study. Figure 1 illustrates the study's methodology.

The first process is to collect data obtained from a public dataset, namely Kaggle. The dataset is uploaded on Google Drive and linked to Google Colab. The data reading process is assisted by the pandas library. In addition, pandas is also used to select the necessary features and data and remove data that is not needed in this study. The column that is discarded in this study is id first because it is not the data needed in the feature selection process and is also the primary key identifier of the data for each column.

Next, we will try to identify outliers for each feature using 3 types of visualization, including histograms, violins, and box plots. These three types of visualization are included in the type of univariate visualizations to understand the distribution of a variable by looking at the characteristics of the data owned by one feature [14]. Machine learning models will produce good prediction results if the predictor variable has a smaller value than the response variable. Besides that, visualization makes it easier for us to see outliers by using visualizations from histograms, violins, and box plots.

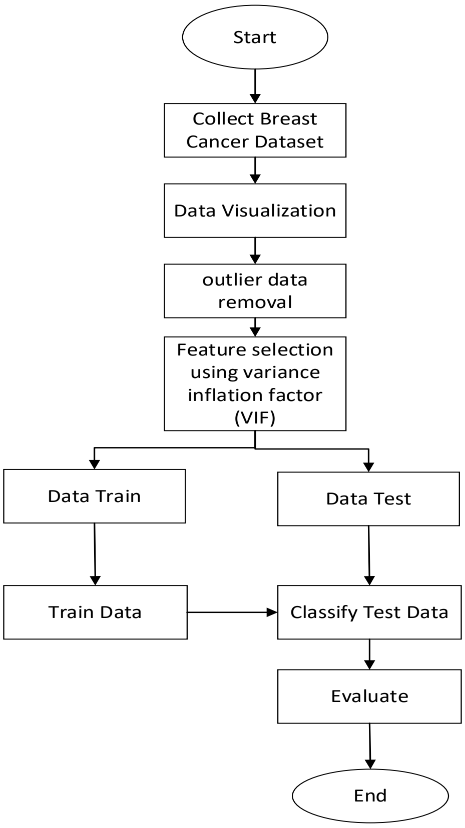

Figure 1. The flowchart of the proposed method.

The variance inflation factor is then used for feature selection after the outliers have been removed from the data. It is crucial to consider the probability of multicollinearity among the features that will be included in the analysis before beginning the regression process. Using a method known as the variable inflation factor, or VIF, this study assesses the data's multicollinearity. The R2 produced by completing a regression of it predictor on the other predictors is known as Ri2 [15]. Iterative steps were taken to remove variables after obtaining the variable inflation factor. Since the trend of that variable is largely reflected by other variables, the process starts with the variable that has the highest value for the variable inflation factor. It was found that improving the largest variable inflation factor values for the features that still existed will improve the system as a whole. Formula 1 contains the formula for the variable inflation factor. Features with VIF values of more than 10,000 are eliminated in the feature selection process.

$$vifi = \frac{1}{1-R_i^2} \tag{1}$$

After selecting the next feature, machine learning capital is used to carry out the classification model that will be made. This study uses seven types of models from machine learning that will be used, including:

2.1 Logistic Regression

The Linear Regression algorithm, which has the same fundamental idea as other regression models, is what gave rise to the machine learning technique known as Logistic Regression [16]. This algorithm determines the relationship between the outcome variable and several independent factors. In contrast to linear regression, logistic regression uses outcome variables that are dichotomous or binary. Models with a single variable or several variables can be handled by logistic regression. One advantage of logistic regression is this. Formula 2 illustrates the logistic regression model's formula.

$$ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2}$$

where the likelihood of the event is represented by the value of π. The slope parameters are βs, the Y-intercept is, and the set of predictors is Xs.

2.2 Random Forest

Breiman developed the supervised method known as random forest in 2001 [17]. Classification, regression, and other issues that are frequently encountered in machine learning can all be solved using random forests. There are numerous justifications for using random names in random forests, including:

- Random training data are utilized to create bootstrap samples of the trees in the random forest.
- When creating a decision tree, the best nodes are chosen by selecting a sample of m variables from the original data set at each split node.

This approach is known as a random forest algorithm because it combines numerous decision trees, each of which depends on random vector values derived from randomly chosen samples that are distributed freely and equally across all trees in the forest. Formula 3 [18] describes the formula for the random forest where I is the indicator function and *hn* is the nth tree of RF.

$$l(y) = argmax_c\left(\sum_{n=1}^{N} I_{h_n(y)=c}\right) \tag{3}$$

2.3 K-Nearest Neighbour

The K-Nearest Neighbor (K-NN) algorithm is a classification method that classifies new data by using the shortest distance as a basis for determining unknown label results. The K-Nearest Neighbor algorithm is classified as a supervised learning algorithm, which functions to classify using training data that has a known class or previous information, then takes the value of k based on the closest distance [19].

2.4 Support Vector Machine

Support vector machine is a learning algorithm that uses a hyperplane which is the dividing field of feature space. The best hyperplane between two classes can be found by measuring the hyperplane margins and finding the maximum point. The margin is the distance between the hyperplane and the closest data from each class, while the data closest to the hyperplane is called a support vector.

2.5 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a technique in machine learning for regression analysis and classification based on the Gradient Boosting Decision Tree (GBDT) [21]. The (XGBoost) method was first introduced by Friedman in 2001, in his research Friedman linked boosting and optimization in building a Gradient Boosting Machine (GBM). Building a new model to predict errors from the previous model is used in the boosting method. The addition of new models is carried out until there are no more error corrections that can be made.

2.6 Gradient Boosting

Gradient Boosting is a special type of algorithm used for the GBT classification task capable of building a decision tree [21] based on an increase in the tree structure on weak learning to correct tree faults and prevent potential overfitting.

In building a decision tree, you can add a very conservative number of iterations which can produce and improve better model performance. GBT can solve the problem by adjusting weak learning to the negative gradient of the loss function and increasing the trees with parameters representing the split variables assigned to each terminal node of the tree.

2.7 Ridge Classifier

Ridge classification is a strategy for analyzing linear discriminant models. Regularization technique that penalizes model coefficients to prevent overfitting. Overfitting is a prevalent problem in machine learning that happens when a model is overly complex and catches noise in the data rather than the underlying signal. This can result in ineffective generalization of new data. Ridge classification tackles this issue by incorporating a penalty term that discourages complexity into the cost function.

Ridge classification works by incorporating a penalty term that discourages complexity into the cost function. Typically, the penalty term is the sum of the squared coefficients of the model’s features. This keeps the coefficients minimal, hence preventing overfitting. Controlling the amount of regularization is possible by adjusting the penalty term. A greater penalty leads to increased regularization and diminished coefficient values. This can be advantageous when a few training data are available. However, if the punishment time is too long, underfitting can occur.

III. RESULTS AND DISCUSSION

In the correlation analysis, it can be seen in Table 1 that it can be seen the correlation of the relationship between the diagnosis and all features. The darker purple the color, the higher the resulting correlation. Correlation values range from 0 to 1. Several features have a high correlation value, including the worst concave point with a correlation value of 0.79, worst radius with a correlation value of 0.78, worst perimeter with a correlation value of 0.78, and concave point means with a correlation value of 0.78.

Table I. Table of correlation values of diagnosis results with its features

No	Features	VIF
1.	concave point worst	0.79
2.	concave point means	0.78
3.	radius_worst	0.78
4.	perimeter_worst	0.78

After performing correlation analysis, the feature data that determines the type of breast cancer is visualized using distribution, violin, and box plots. Figure 4 shows one of the features, namely the concave points mean with the resulting visualization.

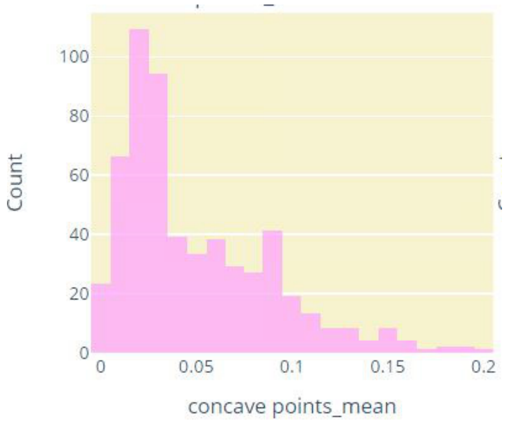

Figure 1. Concave Points Mean Distribution

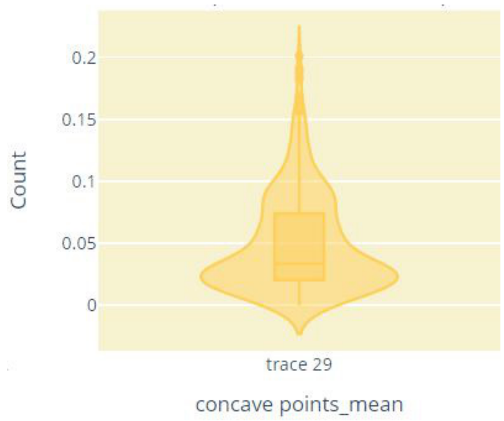

Figure 2. Concave Points Mean Violin Plot

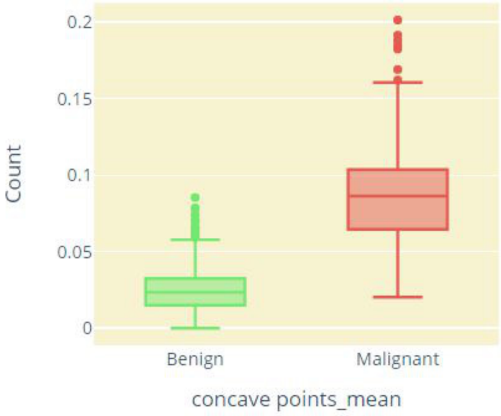

Figure 3. Concave Points Mean Box Plot

As can be seen in figure 4, the concave points means feature has a distribution that tends to have a lot of data when it has a size below 0.05 by using distribution, violin, and box plot visualizations.

Before entering into the modeling process using machine learning, this study removes outliers and data that have a high level of multicollinearities. Eliminating outliers is done by removing data that crosses the upper quartile limit (Q3) and also the lower quartile limit (Q1) as shown in Figure 4 where some data crosses the upper quartile limit. After that to reduce the features that have a high level of multicollinearities is to use the variance inflation factor. Table 1 shows the features with the variance inflation factor.

Table II. VIF of Each Feature

No	Features	VIF
1.	radius_mean	32981.93
2.	perimeter_mean	30694.59
3.	radius_worst	7964.57
4.	perimeter_worst	3898.53
5.	area_mean	1397.10

In Table II, it is possible to observe that the feature values for the mean radius and the mean perimeter both have values that are significantly higher than those of the other features, specifically 32981.93 and 30694.59. As a result, these two characteristics have a variance inflation factor value that is excessively high, and the researchers decided to exclude them from the analysis because they can also contribute to an increase in multicollinearities.

The data that has been generated is then ready to be entered into the machine learning model once outliers have been eliminated and feature selection has been carried out. Logistic Regression, Random Forest, K-Nearest Neighbours, support vector machine, XGBoost, Gradient Boosting, and Ridge Classifier are the names of the machine learning models that were used in this research.

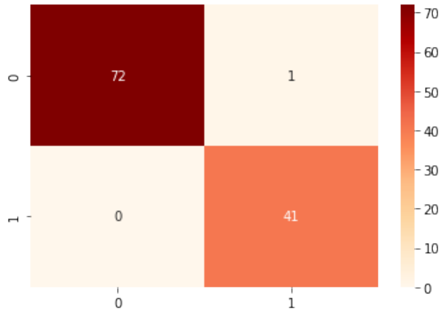

Figure 4. Logistic Regression Confusion Matrix

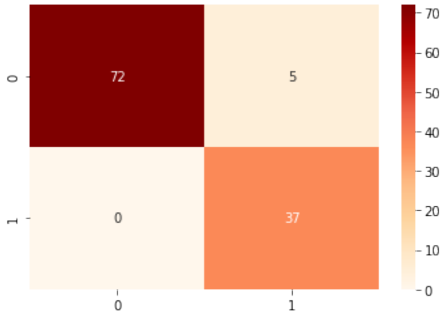

Figure 5. KNN Confusion Matrix

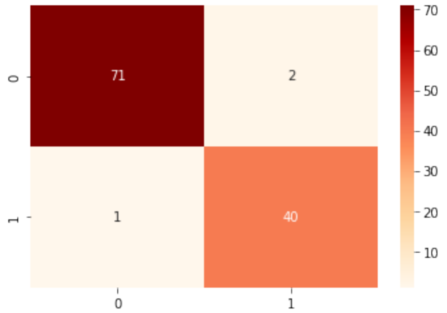

Figure 6. Support Vector Machine Confusion Matrix

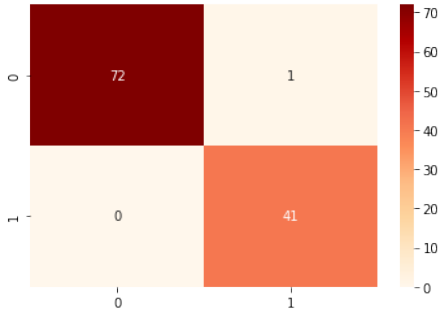

Figure 7. Random Forest Confusion Matrix

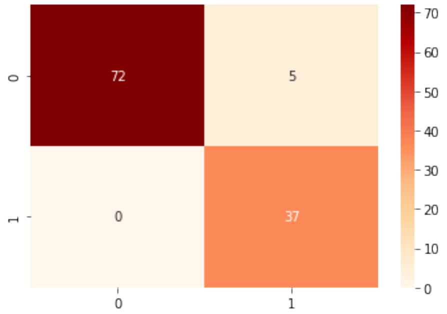

Figure 8. Ridge Classifier Confusion Matrix

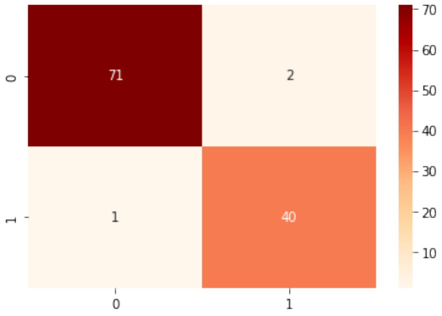

Figure 9. Gradient Boosting Confusion Matrix

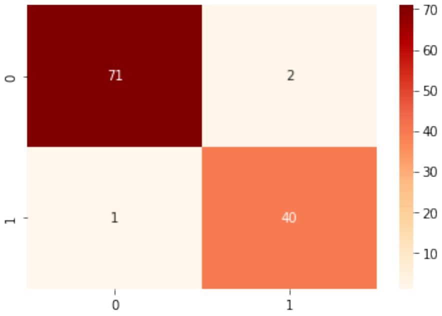

Figure 10. XGBoost Confusion Matrix

Figure 5 to Figure 11 is the Confusion Matrix of each machine learning model. It can be seen that the algorithm that has the best Confusion Matrix is the logistic regression and random forest algorithm with only one data error where the data that should be included in the benign category is included in the malignant category. Figure 12, Figure 13, and Table 2 also compare the accuracy of each machine learning model where the highest accuracy results are in the logistic regression and random forest models with an accuracy value of 99.12%. Obtaining high accuracy can be achieved by eliminating outliers that cross the upper and lower quartile limits and also selecting features using the variance inflation factor. This value increases if you do not use the feature selection variance inflation factor with an accuracy value of only 98.25%.

Table III. Student Distribution Frequency

No	Predictor	Accuracy
1.	Logistic Regression	99.12%
2.	Random Forest	99.12%
3.	Support Vector Machine	97.37%
4.	Gradient Boosting	97.37%
5.	XGBoost	97.37%
6.	KNN	95.61%
7.	Ridge Classifier	95.61%

IV. CONCLUSION

In this study, research was conducted to predict breast cancer with 10 features and two labels, namely 357 labeled benign and 212 malignant for the type of cancer to be predicted. Several previous studies have used several machine learning models and in this study, several machine learning models were selected and evaluated using accuracy metrics. The machine learning models used are logistic regression, random forest, support vector machine, KNN, XGBoost, Gradient Boosting, and Ridge Classifier. Before entering the machine learning model, the data to be used is done by removing outliers and selecting features first. Elimination of outliers is done by removing the upper quartile and lower quartile limits while feature selection is done using the variance inflation factor. The best machine learning models are random forest and logistic regression with an accuracy value of 99.12% which increases without using the variance inflation factor with an accuracy value of 98.25%. The next research will be developed by trying other optimization techniques for hyperparameter tuning.

REFERENCES

[1] S. Nanglia, Muneer Ahmad, Fawad Ali Khan, and N.Z. Jhanjhi, An enhanced Predictive heterogeneous ensemble model for breast cancer prediction, Science Direct, 2022.

[2] Huan-Jung Chiu, Tzuu-Hseng S. Li, (Member, IEEE), and Ping-Huan Kuo, Breast Cancer-Detection System Using PCA, Multilayer Perceptron, Transfer Learning, and Support Vector Machine, IEEE, 2020.

[3] Ilham Mubarog , Arief Setyanto, Heri Sismoro, Sistem Klasifikasi pada Penyakit Breast Cancer dengan Menggunakan Metode Naïve Bayes, Citec Journal, 2019.

[4] Kemal Polat and Ümit Şentürk, A Novel ML Approach to Prediction of Breast Cancer: Combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier, IEEE, 2018.

[5] Momenimovahed, Z., & Salehiniya, H. (2019). Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer: Targets and Therapy, 11, 151.

[6] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018, April). Breast cancer classification using machine learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.

[7] Yash Amethiya, Prince Pipariya, Shlok Patel, and Mana n Shah, Comparative analysis of breast cancer detection using machine learning and biosensors, Science Direct, 2021.

[8] Lestandy, M. (2022). Deteksi Dini Kanker Payudara

Menggunakan Metode Convolution Neural Network (CNN). Inspiration: Jurnal Teknologi Informasi dan Komunikasi, 12(1), 65-72..

[9] Poonam Kathale and Snehal Thorat, Breast Cancer Detection and Classification, IEEE, 2020.

[10] Chand, S. (2020). A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method. Machine Vision and Applications, 31(6), 1-10.

[11] Ak, M. F. (2020, April). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In Healthcare (Vol. 8, No. 2, p. 111). MDPI.

[12] Wang Zhiqiong, Li Mo, Wang Huaxia, Jiang Hanyu, Yao Yudong, Zhang Hao, and Xin Junchang, Breast Cancer Detection Using Extreme Learning Machine Based on Feature Fusion With CNN Deep Features, IEEE, 2019.

[13] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository {http://archive.ics.uci.edu/ml/} (1992)

[14] Cristea, A. I., & Troussas, C. (Eds.). (2021). Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings (Vol. 12677). Springer Nature.

[15] Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. Educational and psychological measurement, 79(5), 874-882.

[16] Peng, C. Y. J., & So, T. S. H. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(1), 31-70.

[17] Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*.

[18] Liparas, D., HaCohen-Kerner, Y., Moumtzidou, A., Vrochidis, S., & Kompatsiaris, I. (2014, November). News articles classification using random forests and weighted multimodal features. In *Information Retrieval Facility Conference* (pp. 63-75). Springer, Cham

[19] Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. CESS (Journal of Computer Engineering, System and Science), 4(1), 78-82.

[20] Supriyatna, A., & Mustika, W. P. (2018). Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil. J-SAKTI (Jurnal Sains Komputer dan Informatika), 2(2), 152-161.

[21] Wade, C. (2020). Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd.