talenta PUBLISHER

DATA SCIENCE

Journal of Computing and Applied Informatics

Machine Learning Approach in Breast Cancer Prediction Using Support Vector Machine

Nur Tri Ramadhanti Adiningrum [1], Resa Rianti[2], Cahyo Prianto[3]

[1,2,3] Computer Engineering, Vocational College, Universitas Logistik dan Bisnis Internasional, Indonesia

Abstract. Breast cancer is the most common disease found in women and the death rate still ranks second in cancer cases which can affect more than 2.1 million people in 2015. Based on these cases, it can be seen that breast cancer is a type of cancer that is the main cause of death for women , but this death can be reduced by early detection of cancer cells. Therefore, cancer detection plays an important role in the treatment process and helps increase survival rates. Cancer prediction can help patients to consult doctors more quickly. So, with the right prediction of cancer is very important to update the treatment of breast cancer patients. Machine Learning techniques can be used to predict breast cancer because they can capture high-level interactions between data that may produce better predictions to differentiate between benign and malignant. Therefore, this research introduces an effective classification approach based on Support Vector Machine. Support Vector Machine (SVM) is a model used to predict breast cancer. To simplify the prediction process, breast cancer prediction results are implemented in the form of a web base with the Django framework so that doctors can make decisions quickly. The modeling results show that the prediction of breast cancer using SVM has the highest accuracy, namely 98.24%.

Keyword: Breast cancer, Prediction, Machine Learning, Support Vector Machine, Web Base

Received date month year. | Revised date month year | Accepted date month year

1 Introduction

Breast cancer is the most common disease found in women and the mortality rate is still in second place among other types of cancer [1]. A study using databases from GLOBOCAN, CDC, and the World Health Organization (WHO) health repository saw that breast cancer is a deadly disease that claims thousands of lives every year[2]. In 2020, a study conducted by the World Health Organization (WHO) showed that more than 2,000,000 new cases and more than 600,000 deaths were reported due to breast cancer in one year[2].

Based on these cases, it can be seen that breast cancer is a type of cancer that is the main cause of death for women. However, this mortality can be reduced by early detection of cancer

*Corresponding author at: Please enter address of author's affiliation, including faculty, department, university, address, city and country

E-mail address: Please enter the email of corresponding author

Journal Homepage: https://jocai.usu.ac.id

cells [3]. Early diagnosis and prediction of breast cancer can increase the chances of survival because it can help provide timely treatment for patients [2]. Timely prediction of cancer can help patients to consult doctors on time[2]. Every development for the prediction and diagnosis of cancer is an important capital for a healthy life. Thus, high accuracy in cancer prediction is important for updating treatment aspects and survival standards in patients[4].

In making early predictions of breast cancer, the method applied is to use Machine Learning with the Support Vector Machine algorithm [4]. Support Vector Machine can be used to predict and diagnose breast cancer, and has become a major research topic and has proven to be a powerful technique of various types of machine learning algorithms[5]. One of the advantages of using machine learning models over statistical models is the amount of flexibility in capturing high-level interactions between data, which may lead to better predictions[2].

Besides that, the use of the website can help users in doing work because it is easy to use and able to process quickly. Django is a web framework based on the Python programming language designed to create dynamic, feature-rich and secure web applications. Django, which was developed by the Django Software Foundation, continues to get improvements, making this web framework the top choice for many web application developers [6].

The main objective of this study is to predict and diagnose breast cancer using Machine Learning with the model that has the highest accuracy based on previous research[4]. Support Vector Machine (SVM) is a model used to predict breast cancer because it has the best accuracy compared to other Machine Learning models [4]. The results of breast cancer prediction can be used by users such as doctors to make decisions quickly. In addition, the model will be made in the form of a website so that predictions can be easily used. The visualization of the prediction results will be displayed web-based with the Django framework to make it easier to understand and for use by medical personnel.

2. Related Works

Research that has been done previously is used to add to the depth of discussion on current research and to support the current research. Meanwhile for research that discusses predictions as described below.

Various types of research have been conducted related to breast cancer which is the most common type of cancer found in women and the death rate is still in second place among other types of cancer. A study conducted by Azminuddin et al evaluated an efficient Machine Learning approach for breast cancer prediction. In this study, this research aims to improve the accuracy performance of the Machine Learning method through pre-processing processes such as: Missing Value Replacement, Data Transformation, Smoothing Noisy Data, Feature Selection / Attribute Weighting, Data Validation, and Unbalanced Class Reduction which are more efficient for

prediction Breast cancer. The results of this study propose an approach: Wisconsin Breast Cancer Dataset - SVM - F-Score with an accuracy of 99.51%[1].

In another study, Anusha et al evaluated the application of machine learning in medical applications such as the detection of cancer cell types. Breast cancer is a disease that causes a high number of deaths every year, and it is the most common type of cancer and the leading cause of death for women worldwide. Cancer cells are classified as benign (B) or malignant (M). In this project, an analysis was performed using the Support Vector Machine (SVM) on the Wisconsin Breast Cancer dataset. From the methodology used, the SVM technique was identified as a powerful technique for predictive analysis. And based on the findings, it was concluded that SVM using the Gaussian kernel is the most suitable technique for predicting recurrence or non-recurrence of breast cancer[3].

The machine learning approach for early detection of breast cancer by Muawia in this study presents an effective method based on a Support Vector Machine (SVM) with the right feature selection method that only considers features that have a high influence and ignores the others. This study also uses reliable datasets and appropriate validation methods to produce reliable and reliable results. The selection of SVM was taken after conducting real experiments with seven classifiers that are popular in the field of breast cancer diagnosis, the experimental results show that the SVM-based classifier approach is superior to other methods. The experimental results reflect that the SVM-based classifier approach outperforms other classifiers by obtaining the highest accuracy, reaching 97.4%. The contributions of this paper include introducing an efficient SVM approach for predicting breast cancer and presenting comparative studies for seven popular classifiers in this field. Our results have been thoroughly validated to nominate SVM as the best classifier for breast cancer detection[5].

Another research was conducted by making a web-based 372 Coffee Sales Prediction application. The framework used in the development of this application is Django with a MySQL database. Sales predictions are made by applying a multiple linear regression algorithm using 372 copies of transaction data, daily weather for the city of Bandung, national holidays and public holidays in 2018-2019. The prediction model built is evaluated with three types of model evaluation, namely MAE, RMSE, and R2. Based on the results of the evaluation of the prediction model, it shows that Gajua Kopi has the best prediction model, with MAE = 104,259,000, RSME = 114,408,000, and R2 = 0.57581. Meanwhile, Kopi 372 Kolmas has the worst prediction model, with MAE = 5,249,620,000, RMSE = 1,465,270,000, R2 = 0.37809 [7].

3. Methodology

Figure 1. *Research Methodology*

3.1 Formation of Dataset

A. Data Collection

In this study, data from around the world was used to train and test all classifier methods used in the study. The data used in this study is data on breast cancer patients obtained through a site called Kaggle in 2022 which can be visited via the link listed below.: https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset.

The process of using the dataset for this study began with collecting breast cancer patient data from the Kaggle website. The data is labeled and has 32 columns and 569 records. This dataset contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. It consists of 32 columns, with the first column showing the ID number, the second column showing the result of the diagnosis (benign or malignant), followed by the mean, standard deviation and mean of the worst measurements of the ten features. There are no blank values. These features were obtained from digital images from a fine needle aspiration biopsy of the tumor. at this stage, data attributes are selected based on the level of correlation with high values between attributes. The selected attributes are as many as 21 attributes. The following are the attributes and their meanings[8] that have been selected based on high attribute correlation values:

Table 1. Dataset attribute and their meaning

Attribute	Meaning
id	ID integer
diagnosis	M = malignant = 1B = benign = 0
Radius_mean	Mean of distances from the center to points on the perimeter cell
Perimeter_mean	Perimeter of cell
Area_mean	Area of cell
Compactness_mean	Perimeter ^ 2/area—1.0
Concavity_mean	The severity of concave portions of the contour
Concave ponts_mean	Number of concave portions of the contour
Radius_se	-
Texture_se	-
Perimeter_se	-
Area_se	-
Compactness_se	-
Concavity_se	-
Concave points_se	-
Fractal_dimension_se	-
Radius_worst	-
Texture_worst	-
Perimeter_worst	-
Area_worst	-
Compactness_worst	-
Concavity_worst	-
Concave points_worst	-

3.2 Model Building SVM

A. Training

This research begins by separating training and testing data using the "train_test_split" function imported from the python sci-kit learn library in the model selection section. Most of the training data is 80% and testing is as much as 20% or 455 data is used as training data and the remaining 114 data is used for data testing. The test data is then used in the modeling process without data labels to ensure that the model works with the type of data to be tested. Then do the fitting process between x_train and y_train.

B. SVM Model

After training and producing a classification model, the next step is to test the classification model with data testing to determine the accuracy of the classification model. After the model is generated from the classification, it is tested from data other than the dataset. The test uses data from welding results with different products[9].

SVM is a machine learning algorithm used for supervised classification and regression [10]. SVM is an algorithm that works using nonlinear mapping to change the original training data to a higher dimension. In this case the new dimension, will find the hyperplane to separate linearly and with proper nonlinear mapping to the higher dimension, data from the two classes can always be separated by the hyperplane. SVM finds this using support vectors and margins. In this technique, an attempt must be made to find the optimal separator function (classifier) that can separate two different classes. This technique seeks to find the best separator function (hyperplane) between functions that are not limited in number to separate two kinds of objects. The best hyperplane is the one that lies midway between two sets of objects of two classes. This can be formulated in the SVM optimization problem for linear classification, as shown below[9].

$$\min_{\rightarrow} \ \tau(w) = \frac{1}{2} \ ||\overrightarrow{w}||^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\frac{1}{2}||\overrightarrow{w}||^2 = \frac{1}{2}(w_1^2 + w_2^2) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

With the provision of:

$$y_i(x_i.w + b) \geq 1, i = 1,2,3, \ldots, n \ldots (3)$$

$$y_i(x_1.w_1 + x_1.w_1 + b) \geq 1 \ldots\ldots\ldots\ldots\ldots (4)$$

Where x_i is the input data y_i is the output of the data x_i, w, b are the parameters we are looking for the values for. In the above formula, you want to minimize the objective function $\frac{1}{2}||\overrightarrow{w}||^2$ or maximize the quantity $||\overrightarrow{w}||^2$ taking into account the delimiter $y_i(x_i.w + b) \geq 1$. If the output data y_1= +1, then the delimiter becomes $(x\ .w + b) \geq 1$.

3.3 Evaluation

The evaluation phase aims to evaluate the quality of the model built. The method used at this stage includes accuracy and confusion matrices to evaluate the results of the model. Cancer cells can be classified as benign (B) or malignant (M). The Support Vector Machine (SVM) method was used on the breast cancer dataset and from the methodology used, the SVM technique was identified as a powerful technique for predictive analysis. Based on the findings, it was concluded that the SVM method is the most suitable technique for predicting recurrence or non-recurrence of breast cancer. This technique uses proper feature selection, considering only the features that have high influence and ignoring the others. This study uses reliable datasets and proper validation methods to produce reliable and trustworthy results. The selection of SVM was based on the results

of experiments with seven popular classification methods in the diagnosis of breast cancer, the results showed that the SVM-based classifier approach was superior to other methods. The experimental results show that the SVM-based classifier approach obtains an accuracy of 98.24%.

	precision	recall	f1-score	support
Tumor Ganas (Terdeteksi Kanker)	1.00	0.95	0.98	42
Tumor Jinak (Tidak Terdeteksi Kanker)	0.97	1.00	0.99	72
accuracy			0.98	114
macro avg	0.99	0.98	0.98	114
weighted avg	0.98	0.98	0.98	114

Figure 2. Confussion Matrix

4. Result and Discussion

In this study, several common classification methods for breast cancer were determined and the one that had the best performance in terms of classification accuracy was selected compared to the others. After applying various factors and techniques to improve the classification accuracy The selection of this method is based on the best method in the classification domain and especially in the diagnosis of breast cancer. The experimental results show that the Support Vector Machine (SVM) method obtains the highest accuracy, namely 98.24% of the training data as much as 80% and 20% is used for testing data. Then a fitting process is carried out between x_train and y_train to evaluate the model's performance on unknown data.

Use case diagrams are a type of UML (Unified Modeling Language) diagram used to describe system interactions with the actors involved in the system. This diagram can be a good picture to explain the context of a system so that the boundaries of the system are

clearly visible [11]. This diagram focuses on the function of the system and the way actors are involved with the system.

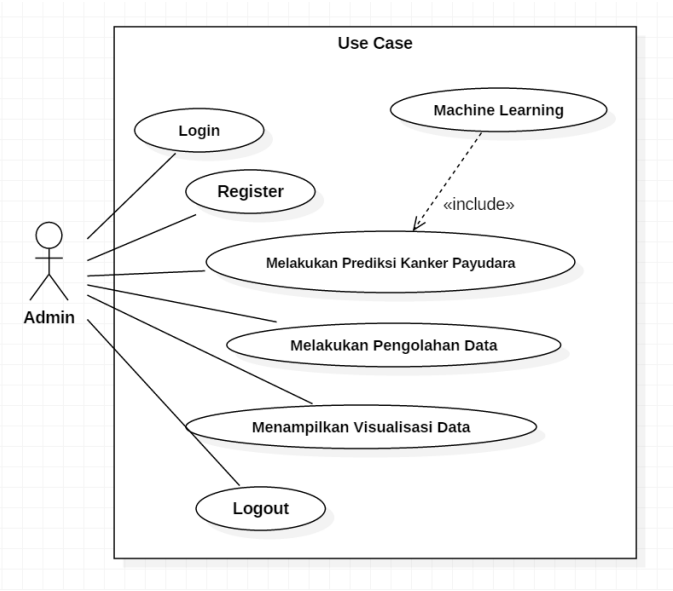

Figure 3. Use Case Diagram

In the prediction website implementation section with the Django framework, the UI results can be seen in the following figure.

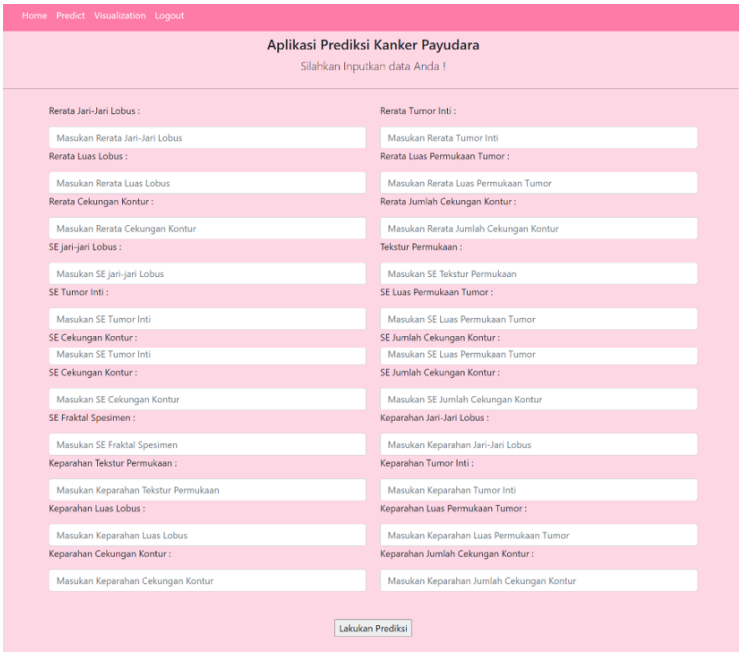

Figure 4. UI Prediction Form

Figure 5. UI Prediction Result

5. Conclusion

Based on the results of the analysis and discussion that has been carried out, several points can be concluded, namely:

1. Based on the model made, the accuracy value is 98.24%. This accuracy is a good accuracy value, so it can be said that machine learning models can perform well to predict breast cancer.

2. Data visualization from the results of the breast cancer prediction model can be used to form a web-based application using the Django framework. With this application, admins can predict breast cancer easily and quickly.

References

[1] A. I. S. Azis, I. Surya Kumala Idris, B. Santoso, and Y. Aril Mustofa, "Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara," *Jurnal Rekayas a Sistem dan T eknol ogi Informasi* , vol. 3, no. 3, pp. 458–469, 2019, Accessed: Oct. 22, 2022. [Online]. Available: http://jurnal.iaii.or.id/index.php/RESTI/article/view/1347/180

[2] S. Raj Gupta, "Prediction Time Of Breast Cancer Tumor Recurrence Using Machine Learning," *Cancer Treat Res Commun*, vol. 32, pp. 2–9, 2022, doi: https://doi.org/10.1016/j.ctarc.2022.100602.

[3] A. Bharat, N. Pooja, and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," *IEEE Third International Conference on Circuits, Control, Communication and Computing*, 2018, doi: https://doi.org/10.1109/CIMCA.2018.8739696.

[4] M. A. Naji, S. el Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis," *Procedia Comput Sci*, vol. 191, pp. 487–492, 2021, doi: https://doi.org/10.1016/j.procs.2021.07.062.

[5] Muawia A. Elsadig, "A Machine Learning Approach For Breast Cancer Early Detection," *J Theor Appl Inf Technol*, vol. 99, no. 5, pp. 1044–1053, 2021, Accessed: Jan. 17, 2023. [Online]. Available: http://www.jatit.org/volumes/Vol99No5/4Vol99No5.pdf

[6] D. Saputra and R. Fathoni Aji, "Analisis Perbandingan Performa Web Service Rest Menggunakan Framework Laravel, Django Dan Ruby On Rails Untuk Akses Data Dengan Aplikasi Mobile (Studi Kasus: Portal E-Kampus STT Indonesia Tanjungpinang)," *Bangkit Indonesia*, vol. 2, no. 2, pp. 17–22, 2018, doi: 10.52771/bangkitindonesia.v7i2.90.

[7] M. Aditia Farhan, "Pengembangan Aplikasi Prediksi Penjualan di 372 Kopi Menggunakan Algoritma Multiple Linear Regression," Bandung, 2021.

[8] M. H. Memon and Z. Wang, "Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection," *Hindawi: Wireless Communications and Mobile Computing*, vol. 2019, no. 5176705, pp. 1–19, 2019, doi: 10.1155/2019/5176705.

[9] A. S. Ritonga and E. S. Purwaningsih, "Penerapan Metode Support Vector Machine (SVM) Dalam Klasifikasi Kualitas Pengelasan SMAW (Shield Metal Arc Welding)," *Jurnal Ilmiah Edutic*, vol. 5, no. 1, pp. 17–25, 2018.

[10] Samsudiney, "Penjelasan Sederhana tentang Apa Itu SVM?," 2019. https://medium.com/@samsudiney/penjelasan-sederhana-tentang-apa-itu-svm-149fec72bd02 (accessed Jan. 11, 2023).

[11] Tri A. Kurniawan, "Pemodelan Use Case (UML): Evaluasi Terhadap Beberapa Kesalahan Dalam Praktik," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 5, no. 1, pp. 77–86, 2018, doi: 10.25126/jtiik.201851610.