ENHANCING BREAST CANCER PREDICTION THROUGH HYPERPARAMETER OPTIMIZATION IN SUPPORT VECTOR MACHINE

VERONICA LANDREA DARNELLA OSWARI[1], RAYMOND SUNARDI OETAMA[2]

[1] Information System, Universitas Multimedia Nusantara, Tangerang, Indonesia

[2] Information System, Universitas Multimedia Nusantara, Tangerang, Indonesia

E-mail: [1]veronica.darnella@student.umn.ac.id, [2]raymond@umn.ac.id

ABSTRACT

Breast cancer is a prevalent and serious disease in the United States, ranking as the third leading cause of death worldwide. The number of cancer-related deaths has risen from 6.2 million in 2000 to 10 million in 2020. Early detection is vital for saving lives and improving treatment outcomes. While the accuracy rate in breast cancer prediction has reached around 97%, other research in different healthcare fields has achieved even higher accuracy rates ranging from 98.062% to 100%. To advance the field and enhance the accuracy of breast cancer detection methods, this study investigates and optimizes previously unexplored parameters. One approach involves utilizing data mining techniques, specifically by combining the Support Vector Machine (SVM) algorithm with Linear Kernel, RBF Kernel, and tuning hyperparameters. The Linear SVM model exhibited exceptional performance, accurately predicting most Malignant and Benign instances with only two incorrect predictions. The SVM model with the RBF kernel demonstrated comparable performance, with minimal errors. By tuning the hyperparameters and utilizing the RBF kernel, the SVM model achieved perfect predictions for Benign cases and high accuracy for Malignant cases. Both the Linear SVM and H-SVM models achieved the highest accuracy of 98.83%, with the RBF SVM model close behind at 98.24%.

Keywords: *Breast Cancer, Data Mining, Early Detection, SVM algorithm.*

1. INTRODUCTION

Breast cancer is the most common cancer in the United States. As can be seen in Figure 1, in 2020, there were 253.465 new cases of breast cancer in the United States, and had the third-highest number of deaths with 42,617 deaths after lung cancer and pancreatic cancer [1].

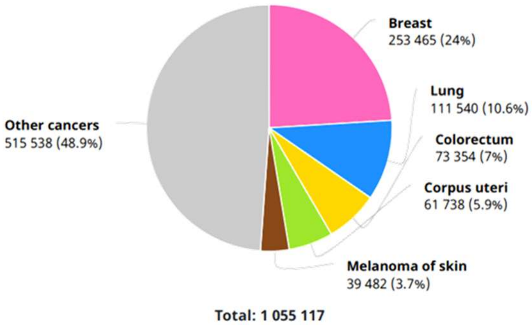


Figure 1: Cancer Types that Affect Women in the USA, 2020 [1]

Although there are several ways for patients to undergo treatment for breast cancer if diagnosed, according to the journal 'Advances in Breast Cancer Radiation Therapy,' the disease can be cured with various treatments. One of the treatments that can be done is surgical resection. Surgical resection involves the surgical removal of part or all the tumors. Surgical resection is usually followed by a program of chemotherapy, radiation therapy, and endocrine therapy, which can be tailored based on the severity of breast cancer [2]. Projections indicate that the number of individuals diagnosed with cancer will continue to rise in the coming years, with a nearly 50% increase expected by 2040 compared to 2020. Additionally, cancer-related deaths have also increased from 6.2 million in 2000 to 10 million in 2020 [3].

While it is possible for individuals diagnosed with breast cancer to potentially achieve remission through treatment, it should be acknowledged that the disease can progress and pose a risk of mortality. Therefore, it is crucial to minimize the risk of breast cancer by focusing on early detection through predictive methods utilizing advanced technology. Numerous previous studies have employed data mining techniques to classify and predict breast cancer data. The utilization of data mining can significantly aid in the early diagnosis of breast cancer, enabling timely intervention and treatment before cancer progresses to an advanced stage [4].

Numerous studies have employed machine learning algorithms, specifically Support Vector Machines (SVM), to predict and diagnose breast cancer. These studies consistently reveal the impressive accuracy of SVM, ranging from 97% to 97.2%. A previous study utilized algorithmic modeling for the early detection of breast cancer using Breast Cancer Wisconsin Diagnostic. The study employed five algorithms, including SVM, Random Forest, Logistic Regression, k-Nearest Neighbor, and Decision Tree. Results showed that SVM achieved the highest accuracy of 97.2% [3]. The next study also predicted breast cancer using the same dataset. This study utilized two algorithms, namely SVM and Random Forest. The SVM algorithm achieved the highest accuracy in this study, with an accuracy rate of 97% [5]. Another study conducted in 2020, also predicted breast cancer. The highest accuracy was achieved by the SVM algorithm, with a rate of 97.2% [6]. However, these results still fall below the achievements reported in other healthcare fields. The research applied SVM on Covid-19 prediction has successfully obtained higher accuracy (98.062%) [7]. Another study found that the Linear Kernel achieved an even better accuracy of 100% [8]. They have achieved such exceptional results by attaining the utmost precision through their meticulous fine-tuning of hyperparameters within the SVM algorithm. This realization presents a promising opportunity to enhance breast cancer prediction accuracy and surpass previous achievements in the field. So, this study aims to improve the prediction and diagnosis of breast cancer by exploring an unexplored area, namely hyperparameter tuning in the SVM algorithm. In real life, the remarkable accuracy of machine learning algorithms in breast cancer prediction and diagnosis holds the potential to have a significant impact on patient outcomes, quality of life, and healthcare resource management. It represents a significant advancement in the field, offering hope for improved detection and personalized treatment strategies for individuals affected by breast cancer.

2. LITERATURE REVIEW

2.1 Support Vector Machine

SVM is a commonly used algorithm for classification and regression tasks. The main idea behind SVM is to find the best possible line or plane (hyperplane) that separates different classes of data [9]. The goal of SVM is to find an optimal hyperplane that is positioned as far as possible from the support vectors [10]. The training dataset in SVM is mathematically formulated as follows in Formula (1): [11].

$$(x_1, y_1), \dots, (x_n, y_n), x_i \epsilon R^d, y_i \epsilon (-1, +1) \quad (1)$$

where x_i represents the feature vector and y_i represents the class label, both in positive and negative values. The optimal hyperplane can be formulated in Formula (2).

$$wx^T + b = 0 \quad (2)$$

where w represents the weight vector, x represents the input feature vector, and b represents the bias value. The values of w and b can satisfy formulas (3) and (4) [11]. The training of this SVM model aims to find the values of w and b to maximize the margin, as shown in Figure 2.

$$w{x_i}^T + b \geq +1 \ if \ y_i = 1 \quad (3)$$

$$w{x_i}^T + b \leq -1 \ if \ y_i = -1 \quad (4)$$

In addition, SVM also has a Kernel function to map classification data. The use of SVM with Kernels is done according to relevant parameters such as the penalty parameter (C) and the Gamma parameter (γ), which are useful for managing the learning stages in the SVM algorithm and can impact its accuracy [12]. In this SVM algorithm, there are several kernels. The Linear Kernel is the simplest kernel that analyzes linearly separable data [13]. The RBF (Radial Basis Function) Kernel is used for non-linearly separable data [14]. Linear Kernel SVM applies formula (5) while RBF Kernel applies formula (6) [11].

$$K(x_i, x_j) = 1 + \ x_i^T x_j \quad (5)$$

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \quad (6)$$

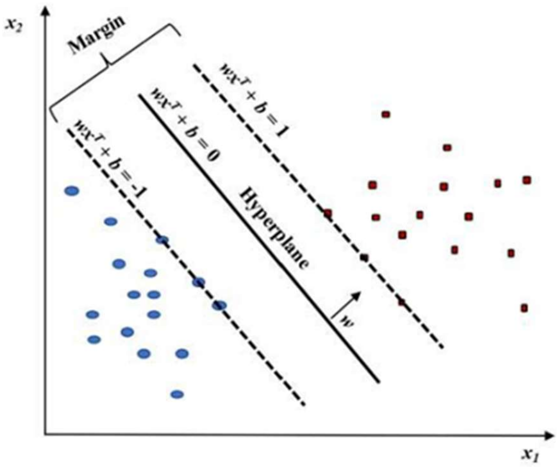


Figure 2: SVM Hyperplane [11]

2.2 Hyperparameter

Hyperparameter is a machine learning classification method for determining and improving performance [15]. Hyperparameters play a role in influencing learning models, for example in machine learning. Hyperparameters also play a role in determining construction and determining the appropriate and best evaluation value in conducting the classification process in machine learning. In addition, hyperparameters also play a role in finding the best parameters in classifying machine learning so that they can help provide the best possible accuracy value [16]. Hyperparameters in SVM are tunable parameters that are set before the training process begins and significantly impact the performance and behavior of the SVM model [17]. Common hyperparameters include C, which controls the trade-off between margin maximization and training error minimization; the choice of kernel, determining the mapping of input data into a higher-dimensional feature space with options like linear or radial basis function (RBF); and gamma (γ), which influences the shape of the decision boundary, where a smaller value leads to a smoother boundary, and a larger value results in a more complex boundary closely fitting the training data [18].

2.3 Classification Report.

The confusion matrix is commonly used in machine learning, especially in supervised classification or determining classification models. The structure of a confusion matrix is represented in the form of rows and columns, where rows typically contain actual classes, and columns contain predicted classes. In the case of binary classification, the confusion matrix is usually displayed as a 2 x 2 matrix. The confusion matrix includes four measures: true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [19]. The confusion matrix in machine learning provides important measures for evaluating classification models. It consists of four elements: true positive, true negative, false positive, and false negative. True positive represents the number of positive instances correctly classified by the system, while true negative represents the number of negative instances correctly classified. False positive refers to the number of negative instances incorrectly classified as positive, and false negative represents the number of positive instances incorrectly classified as negative. [20].

A classification report is a performance evaluation matrix in machine learning which usually contains a report consisting of accuracy, precision, recall, and f1-score values. Accuracy is the output value of a model, which is usually used to represent a good model that is predicted in a true positive or false negative value compared to the entire data. The accuracy formula is shown in formula (7) [21].

$$Accuracy = \frac{TP+T}{TP+TN+FP+FN} \tag{7}$$

Precision is a metric that provides a ratio of the true positive values to the data that is predicted to be positive. It measures the accuracy of positive predictions made by a classification model. To calculate precision, formula (8) can be used that considers the true positive (TP) and false positive (FP) values. The formula for precision is [21]:

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

The recall represents the proportion of actual positive instances correctly identified by a classification model. It measures the model's ability to capture all positive instances in the dataset. To calculate recall, formula (9) is used which considers the true positive (TP) and false negative (FN) values. The formula for the recall is [21]:

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

The f1-score value is a value that represents a combination of precision and recall values. The f1-score usually contains the average value of the two values, so it is necessary to make observations between false positives and false negatives. The f1-score is computed using formula (10) [21].

$$f1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision+\ Recall} \qquad (10)$$

3. RESEARCH METHODS

3.1 Research Object and Data Collection

The data used in this research is collected from the UCI Machine Learning Repository, specifically from a dataset called "Breast Cancer Wisconsin." The dataset was created by Dr. William H. Wolberg from the General Surgery Department and W. Nick Street and Olvi L. Mangasarian from the Computer Sciences Department, all associated with the University of Wisconsin [22].

3.2 Research Flow

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a framework that simplifies the data mining process into six easy-to-understand phases. These phases include Business Understanding, where project goals are defined; Data Understanding, where the data is explored and analyzed; Data Preparation, where the data is cleaned and transformed; Modeling, where various techniques are applied to create models; Evaluation, where the models are assessed for performance; and Deployment, where the chosen model is implemented in a real-world setting [23]. However, this research was not conducted in the deployment stage because it was limited to study purposes only and was not implemented in companies.

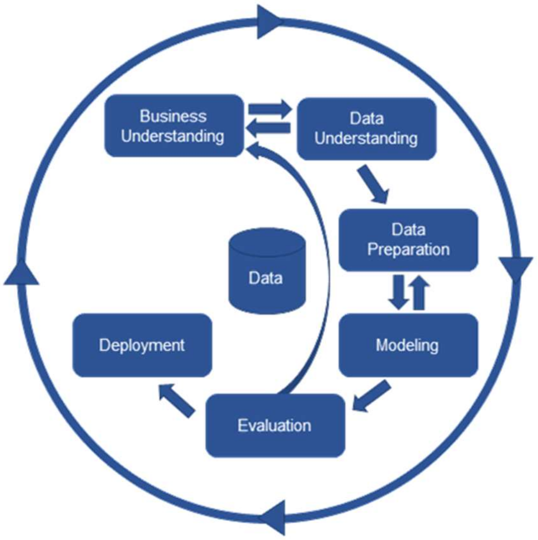


Figure 3: CRISP-DM Stages

Table 1: Breast Cancer Wisconsin Data Structure

No.	Attributes	Descriptions	Data Types
1	id	Id number	numeric
2	diagnosis	Diagnosis (M = malignant, B = benign)	factors
3	radius_mean	mean radius	numeric
4	texture_mean	mean texture	numeric
5	perimeter_mean	mean perimeter	numeric
6	area_mean	mean area	numeric
7	smoothness_mean	mean smoothness	numeric
8	compactness_mean	mean compactness	numeric
9	concavity_mean	mean concavity	numeric
10	concave points_mean	mean concave points	numeric
11	symmetry_mean	mean symmetry	numeric
12	fractal_dimension_mean	mean fractal_dimension	numeric
13	radius_se	standard error radius	numeric
14	texture_se	standard error texture	numeric
15	perimeter_se	standard error perimeter	numeric
16	area_se	standard error area	numeric
17	smoothness_se	standard error smoothness	numeric
18	compactness_se	standard error compactness	numeric
19	concavity_se	standard error concavity	numeric
20	concave points_se	standard error concave points	numeric
21	symmetry_se	standard error symmetry	numeric
22	fractal_dimension_se	standard error fractal dimension	numeric
23	radius_worst	worst radius	numeric
24	texture_worst	worst texture	numeric
25	perimeter_worst	worst perimeter	numeric
26	area_worst	worst area	numeric
27	smoothness_worst	worst smoothness	numeric
28	compactness_worst	worst compactness	numeric
29	concavity_worst	worst concavity	numeric
30	concave points_worst	worst concave points	numeric
31	symmetry_worst	worst symmetry	numeric
32	fractal_dimension_worst	worst fractal_dimension	numeric

During the Business Understanding stage, the goal is to understand the problems and find solutions. In the Data Understanding stage, we obtained data from the UCI Machine Learning Repository, specifically the Breast Cancer

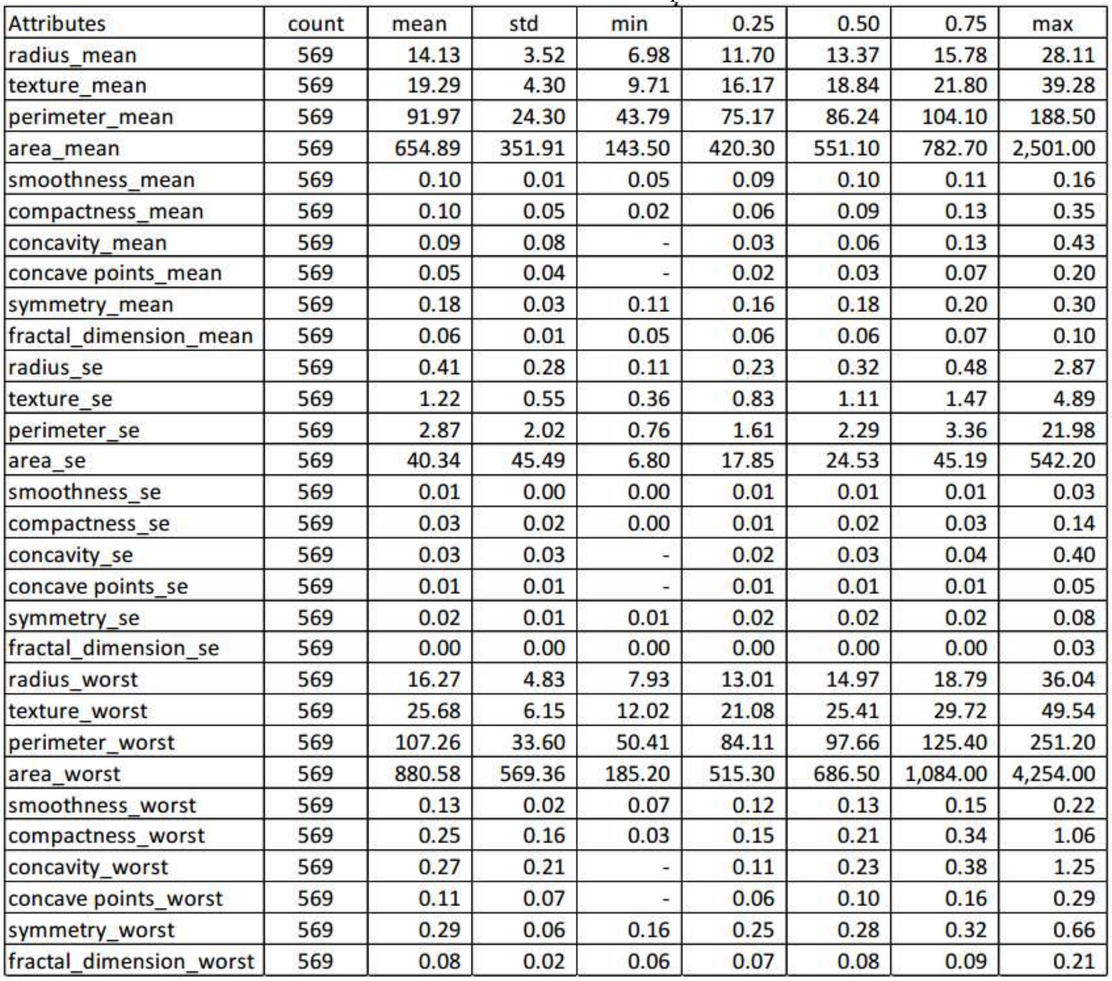
Wisconsin dataset, which has thirty-two attributes. The next step is Data Preparation, where we preprocess the data by filtering out NULL values and removing unused attributes. We also convert categorical data to integers using label encoding and normalize the data. Then, we split the data into training and testing sets, with a 70:30 ratio. In the Modelling stage, we use Jupyter Notebook with Python to create models using the SVM algorithm. We apply both Linear Kernel and RBF Kernel, along with hyperparameters. To evaluate the models, we use a confusion matrix and classification report to assess their performance. Finally, we proceed to the evaluation stage, where we analyze the results of the modeling process. At the evaluation stage, the performance evaluation process is conducted to determine how the prediction results with the modeling built to complete the Business Understanding. The evaluation stage will be conducted by displaying the accuracy value, classification report, confusion matrix, and the results of assessing the data mining model with testing data that has been previously divided following previous research.

4. RESULTS AND DISCUSSION

4.1 Business Understanding

Breast cancer is the most common type of cancer that affects women. It begins in the cells of the breast and can typically be detected through a noticeable lump or X-ray scans. Breast cancer can be categorized as either benign (non-cancerous) or malignant (cancerous) based on how the tumor grows and its ability to spread to other areas of the body through the bloodstream or lymphatic system [24]. Breast cancer is a significant threat to women's

Table 2: Data Statistic Descriptions



Attributes	count	mean	std	min	0.25	0.50	0.75	max
radius_mean	569	14.13	3.52	6.98	11.70	13.37	15.78	28.11
texture_mean	569	19.29	4.30	9.71	16.17	18.84	21.80	39.28
perimeter_mean	569	91.97	24.30	43.79	75.17	86.24	104.10	188.50
area_mean	569	654.89	351.91	143.50	420.30	551.10	782.70	2,501.00
smoothness_mean	569	0.10	0.01	0.05	0.09	0.10	0.11	0.16
compactness_mean	569	0.10	0.05	0.02	0.06	0.09	0.13	0.35
concavity_mean	569	0.09	0.08	-	0.03	0.06	0.13	0.43
concave points_mean	569	0.05	0.04	-	0.02	0.03	0.07	0.20
symmetry_mean	569	0.18	0.03	0.11	0.16	0.18	0.20	0.30
fractal_dimension_mean	569	0.06	0.01	0.05	0.06	0.06	0.07	0.10
radius_se	569	0.41	0.28	0.11	0.23	0.32	0.48	2.87
texture_se	569	1.22	0.55	0.36	0.83	1.11	1.47	4.89
perimeter_se	569	2.87	2.02	0.76	1.61	2.29	3.36	21.98
area_se	569	40.34	45.49	6.80	17.85	24.53	45.19	542.20
smoothness_se	569	0.01	0.00	0.00	0.01	0.01	0.01	0.03
compactness_se	569	0.03	0.02	0.00	0.01	0.02	0.03	0.14
concavity_se	569	0.03	0.03	-	0.02	0.03	0.04	0.40
concave points_se	569	0.01	0.01	-	0.01	0.01	0.01	0.05
symmetry_se	569	0.02	0.01	0.01	0.02	0.02	0.02	0.08
fractal_dimension_se	569	0.00	0.00	0.00	0.00	0.00	0.00	0.03
radius_worst	569	16.27	4.83	7.93	13.01	14.97	18.79	36.04
texture_worst	569	25.68	6.15	12.02	21.08	25.41	29.72	49.54
perimeter_worst	569	107.26	33.60	50.41	84.11	97.66	125.40	251.20
area_worst	569	880.58	569.36	185.20	515.30	686.50	1,084.00	4,254.00
smoothness_worst	569	0.13	0.02	0.07	0.12	0.13	0.15	0.22
compactness_worst	569	0.25	0.16	0.03	0.15	0.21	0.34	1.06
concavity_worst	569	0.27	0.21	-	0.11	0.23	0.38	1.25
concave points_worst	569	0.11	0.07	-	0.06	0.10	0.16	0.29
symmetry_worst	569	0.29	0.06	0.16	0.25	0.28	0.32	0.66
fractal_dimension_worst	569	0.08	0.02	0.06	0.07	0.08	0.09	0.21

physical and mental health worldwide. It is crucial to detect and treat breast cancer early, as patients diagnosed in the preliminary stages have higher survival rates compared to those diagnosed later. To aid in the identification of breast cancer, various imaging techniques are available, including mammography (MG), ultrasonography (US), magnetic resonance imaging (MRI), positron emission computed tomography (PET), computed tomography (CT), and single-photon emission computed tomography (SPECT). These imaging techniques offer quick and accurate detection of breast cancer. However, it is important to consider that these methods can be costly and may pose risks due to the potential for high radiation exposure. [25]. Another alternative for detecting breast cancer is through a biopsy procedure. During a biopsy, a small sample of tissue is obtained either through a needle or a small incision. This tissue sample is then examined under a microscope to determine if cancer cells are present. Biopsy provides a direct and accurate method for diagnosing breast cancer and is often performed when suspicious findings are detected during imaging tests or physical examinations. By analyzing the tissue sample, healthcare professionals can confirm the presence of cancer and gather additional information about its characteristics, such as its type and grade, which helps guide treatment decisions. [26].

Data mining technology, with its classification and prediction capabilities, facilitates breast cancer early detection. In this study, data analysis was performed using the CRISP-DM data processing technique on the Wisconsin Breast Cancer dataset. By leveraging data analysis and machine learning, breast cancer prediction can assess the likelihood of breast cancer based on patient information. The target variable for prediction is the diagnosis attribute, initially categorized as "Malignant" or "Benign." Enhancing early detection is crucial in addressing the impact of breast cancer mortality.

4.2 Data Understanding

In this study, a dataset in .csv format was utilized, comprising a total of 569 data points with thirty-two attributes. At the initial stage, a thorough examination was conducted to ensure the content and completeness of each attribute in the dataset. After retrieving the data, a shape check was performed to determine its dimensions. The dataset consists of thirty-two rows and 569 data points. To conduct the analysis, the study focused on independent variables that encompassed all the attributes listed in Table 1. Furthermore, the statistical description of the variables presented in Table 2 aids in understanding the distribution and characteristics of the dataset. It provides valuable information about the central tendency (mean), variability (standard deviation), and range (minimum and maximum) of each attribute. Additionally, the quartiles (first, second, and third) illustrate the data's distribution across different percentiles. These statistical measures offer researchers a comprehensive overview of the dataset's numerical properties, facilitating further analysis and interpretation.

4.3 Data Preparation

In this stage, data processing is conducted. First, the unused column will be dropped, namely the "id" and "unnamed: 32" attributes. Attribute "id" is not used because it only contains the id number, and its value does not affect this prediction process. While the attribute "unnamed: 32" is not used because it has 569 NULL values.

The dataset undergoes a filtering or cleaning process. This involves identifying data with missing values or NULL values to be promptly addressed by either excluding or removing the rows containing missing values. Additionally, unnecessary attributes are also eliminated during this stage. Encoding is performed using label encoding to convert categorical datasets into integer values, and the data is normalized to prepare it for further processing. Furthermore, data encoding is performed on the diagnosis attribute. The encoding process will be conducted using an encoding label by changing the diagnosis attribute value which initially contains the values "Malignant" and "Benign" and will be converted to an integer with a value of "0" for "Benign" and "1" for "Malignant". Figure 4 shows the Diagnosis Attribute Bar Chart, where this attribute consists of 212 records of Malignant (M) and 357 records of Benign (B).

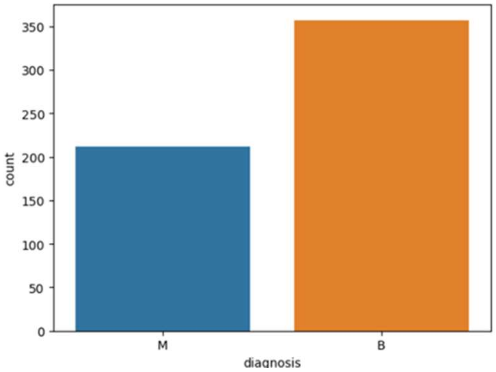


Figure 4: Diagnosis Attributes

Next, the data normalization process will be conducted. Normalization is done to obtain dataset values with the same range [27]. The normalization process uses the MinMaxScaler() function from sklearn [28]. Once the initial data processing steps, such as removing unused attributes, encoding the data, and normalizing it, are completed, the next crucial step is to divide the dataset. This division process involves creating two separate sets: the training data and the testing data. The dataset is divided into a 70:30 proportion, meaning that 70% of the data is allocated for training the model, while the remaining 30% is reserved for evaluating the model's performance.

4.4 Modeling

At the modeling stage, the SVM algorithm with Linear Kernel, RBF, and Hyperparameters is employed. First, the modeling process involves the utilization of Linear SVM. The Linear SVM model is created using the sklearn and seaborn libraries. By utilizing the train and test data that has been previously split, the quality of the model is evaluated using the test data. Figure 5 illustrates the results of the Linear SVM model, demonstrating its remarkable performance with 108 correctly predicted positive instances (Malignant) and sixty-one correctly predicted negative instances (Benign). Moreover, the model only produces 2 False Negatives, indicating its high performance in identifying Malignant cases.

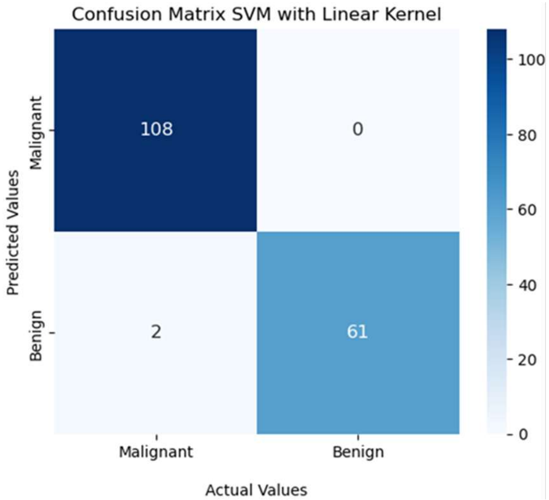


Figure 5: Linear SVM Confusion matrix

Furthermore, the SVM Kernel RBF modeling is conducted. This model also results in extremely high prediction performance, like the Linear SVM. But it makes an exceedingly small mistake. There is one instance incorrectly classified as positive as shown in Figure 6.

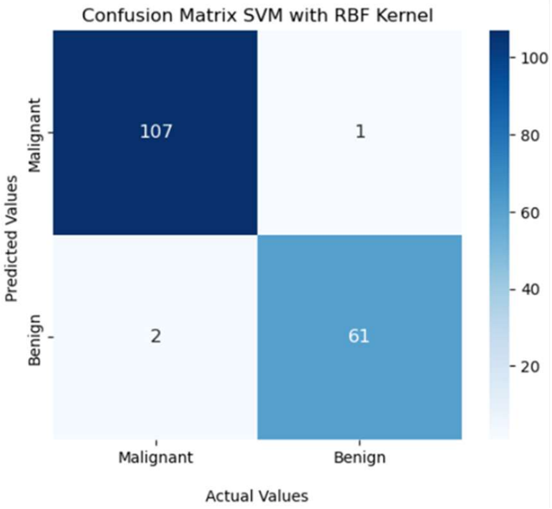


Figure 6: RBF SVM Confusion Matrix

Afterward, the SVM modeling with hyperparameters is conducted. Values for the parameter C are {0.1, 1, 10, 100, dan 1000} and for Gamma parameter is {1, 0.1, 0.01, 0.001, dan 0.0001}, and the kernel is {Linear, RBF}. The best result is found at C=10, Gamma=0.01, Kernel=RBF. As shown in Figure 7, the true positive is 108, the true negative is sixty-one, the false negative is two, and the false positive is zero. This condition is the same as the predicted result of Linear SVM.

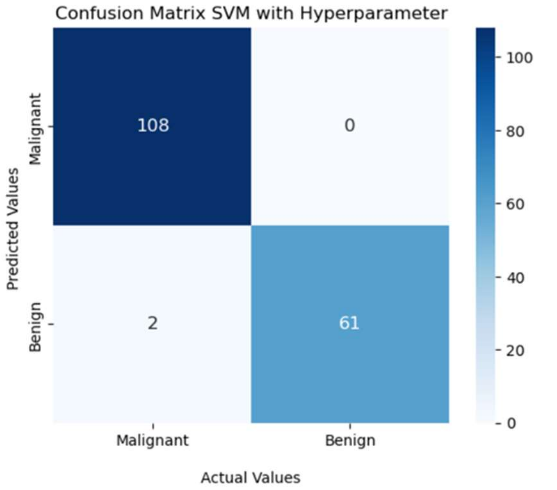


Figure 7: SVM with Hyperparameter Confusion Matrix

4.5 Evaluation

After the modeling is done and the results of the classification report are seen, the value of the output of each model will be compared. Table 3 will show the comparison of Linear SVM, RBF SVM, and SVM with hyperparameter or H-SVM. Judging from the accuracy value, both the Linear SVM model and SVM with Hyperparameter model have the highest accuracy value of 98.83%. Even so, the RBF SVM accuracy is also extremely high (98.24%). Similar situations occur for F1-score and Precision. Meanwhile, the results are the same for recall across all three models.

Table 3: Comparison of All Models

Models	Accuracy	F1-score	Precision	Recall
Linear SVM	0.9883	0.9838	1.0000	0.9682
RBF SVM	0.9824	0.9760	0.9838	0.9682
H-SVM	0.9883	0.9838	1.0000	0.9682

In addition, Table 4 compares the results obtained in this study with those of previous studies. The findings demonstrate an impressive performance in accurately predicting and diagnosing breast cancer using the machine learning algorithms employed. Both [3] and [6] achieved an accuracy of 97.2%, which is slightly lower than the results of this study. [5], on the other hand, reported a lower accuracy of 97%, even lower than [3] and [6]. Overall, the reported accuracies in all the studies indicate a remarkable performance in breast cancer prediction and diagnosis. This is a positive outcome as it suggests that machine learning algorithms can effectively contribute to accurate predictions in this field.

Table 4: Comparison of Accuracy with Previous Research Results

Literature	Dataset	Accuracy
This Study	Breast Cancer Wisconsin Diagnostic	98.83%
Machine learning Algorithms for Breast Cancer Prediction and Diagnosis [3]	Breast Cancer Wisconsin Diagnostic	97.2%
Predicting the Possibility of Cancer with Supervised Learning Algorithms [6]	Breast Cancer Wisconsin Diagnostic	97.2%
Diagnosis of Breast Cancer Based on Support Vector Machine and Random Forest Method [5]	Breast Cancer Wisconsin Diagnostic	97%

These studies share similarities in terms of recommending Support Vector Machines (SVM) as the preferred algorithm and utilizing the Breast Cancer Wisconsin Diagnostic dataset from UCI for breast cancer prediction. However, the distinguishing factor in this study lies in the inclusion of hyperparameter tuning. Consequently, the initial SVM accuracy of approximately 97% was further improved to reach an impressive 98.83%. While some studies may solely focus on comparing algorithms, it is crucial to emphasize the improvement in quality to ensure that the compared results represent the optimal performance achievable by those algorithms.

5. CONCLUSION

In general, the performance of the Linear SVM model was excellent, correctly predicting 108 instances of Malignant as positive and sixty-one instances of Benign as negative. It only made two incorrect predictions as False Negatives. Additionally, the SVM model using the RBF kernel was also implemented. This model demonstrated a remarkably important level of prediction performance, comparable to Linear SVM. However, it had minimal error, with only one instance incorrectly classified as positive. Furthermore, SVM modeling was performed with different hyperparameter values. The best result was achieved with C=10, Gamma=0.01, and the RBF kernel. The SVM model with the specified hyperparameters achieved a perfect prediction for the negative class (Benign) with a true negative count of sixty-one and no false positives. Additionally, it achieved a high number of true positives (108) and a low count of false negatives (2) for the positive class (Malignant). These evaluation metrics closely matched the predicted results obtained from the Linear SVM model, indicating the effectiveness and consistency of the SVM modeling approach. Afterward, the output values of each model are compared. In terms of accuracy, both the Linear SVM and H-SVM models achieved the highest accuracy value of 98.83%. The RBF SVM model also had a high accuracy of 98.24%. The F1-score and Precision values showed similar patterns. Moreover, the recall values were identical across all three models. In comparison to previous studies, this study achieved the highest accuracy of 98.83% in accurately predicting and diagnosing breast cancer using machine learning algorithms. Other studies reported accuracies ranging from 97% to 97.2%, which were slightly lower. Therefore, the aim of this research to enhance the accuracy of breast cancer prediction has been achieved. The implementation of hyperparameters is highly recommended to achieve optimal or even maximum results. In comparative algorithm research, it is also advisable to use enhancement techniques such as hyperparameter tuning to ensure fair comparisons among different algorithms, thus obtaining reliable and unbiased results for each algorithm.

For future research, possible weaknesses of the models include the need for evaluation of more external data to improve generalizability. Addressing these weaknesses can improve the models' performance and reliability in breast cancer prediction.

ACKNOWLEDGEMNT:

Thank you to Universitas Multimedia Nusantara for their support in terms of funding and facilities for this research.

REFERENCES:

[1] IARC Inc., “United States of America Fact Sheet 2020,” *WHO*, 2020. https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf

[2] M. Siedow and V. Grignol, “Advances in Breast Cancer Radiation Therapy,” *Curr. Breast Cancer Rep.*, vol. 13, no. 1, pp. 49–55, 2021, doi: 10.1007/s12609-020-00401-z.

[3] 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: 10.1016/j.procs.2021.07.062.

[4] M. Kaya Keleş, “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.

[5] Y. Wu, “Diagnosis of breast cancer based on support vector machine and random forest methods,” *Proc. - 2020 Int. Conf. Comput. Data Sci. CDS 2020*, pp. 147–151, 2020, doi: 10.1109/CDS49703.2020.00036.

[6] B. G. Pillai, I. Jeena Jecob, J. A. Madhurya, and A. K. Saritha, “Predicting the Possibility of Cancer with Supervised learning Algorithms,” *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 9, pp. 5177–5179, 2020, doi: 10.30534/ijeter/2020/47892020.

[7] T. B. Chandra, K. Verma, B. K. Singh, D. Jain, and S. S. Netam, “Coronavirus disease (COVID-19) detection in Chest X-Ray images using majority voting based classifier ensemble,” *Expert Syst. Appl.*, vol. 165, no. August 2020, p. 113909, 2021, doi: 10.1016/j.eswa.2020.113909.

[8] H. M. Afify and M. S. Zanaty, “Computational predictions for protein sequences of COVID-19 virus via machine learning algorithms,” *Med. Biol. Eng. Comput.*, vol. 59, no. 9, pp. 1723–1734, 2021, doi: 10.1007/s11517-021-02412-z.

[9] D. Karmiani, R. Kazi, A. Nambisan, A. Shah, and V. Kamble, “Comparison of Predictive Algorithms: Backpropagation, SVM, LSTM and Kalman Filter for Stock Market,” in *Proceedings - 2019 Amity International Conference on Artificial Intelligence, AICAI 2019*, 2019, pp. 228–234. doi: 10.1109/AICAI.2019.8701258.

[10] C. Destitus, W. Wella, and S. Suryasari, “Support Vector Machine VS Information Gain: Analisis Sentimen Cyberbullying di Twitter Indonesia,” *Ultim. InfoSys J. Ilmu Sist. Inf.*, vol. 11, no. 2, pp. 107–111, 2020, doi: 10.31937/si.v11i2.1740.

[11] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, “Applications of support vector machine (SVM) learning in cancer genomics,” *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[12] I. K. Nti, O. Nyarko-Boateng, F. A. Adekoya, and B. A. Weyori, “An empirical assessment of different kernel functions on the performance of support vector machines,” *Bull. Electr. Eng. Informatics*, vol. 10, no. 6, pp. 3403–3411, 2021, doi: 10.11591/eei.v10i6.3046.

[13] A. Zeputra and F. Utaminingrum, “Perbandingan Akurasi untuk Deteksi Pintu berbasis HOG dengan Klasifikasi SVM menggunakan Kernel Linear , Radial Basis Function dan Polinomial pada Raspberry Pi,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 5, no. 11, pp. 4746–4757, 2021.

[14] F. Peter and S. John, “SUPPORT VECTOR MACHINE WITH RADIAL BASIS FUNCTION FOR FACIAL EMOTION VALENCE RECOGNITION,” *Eur. J. Mol. Clin. Med.*, vol. 9, no. 4, pp. 1960–1969, 2022.

[15] R. S. Oetama, Y. Heryadi, Lukas, and W. Suparta, “Improving Candle Direction Classification in Forex Market Using Support Vector Machine with Hyperparameters Tuning,” in *2022 7th International Conference on Informatics and Computing, ICIC 2022*, 2022, pp. 1–6. doi: 10.1109/ICIC56845.2022.10006993.

[16] F. Khan, S. Kanwal, S. Alamri, and B. Mumtaz, "Hyper-parameter optimization of classifiers, using an artificial immune network and its application to software bug prediction," *IEEE Access*, vol. 8, pp. 20954–20964, 2020, doi: 10.1109/ACCESS.2020.2968362.

[17] L. R. Halim and A. Suryadibrata, "Cyberbullying Sentiment Analysis with Word2Vec and One-Against-All Support Vector Machine," *IJNMT (International J. New Media Technol.*, vol. 8, no. 1, pp. 57–64, 2021, doi: 10.31937/ijnmt.v8i1.2047.

[18] J. Wainer and P. Fonseca, "How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms," *Artif. Intell. Rev.*, vol. 54, no. 6, pp. 4771–4797, 2021, doi: 10.1007/s10462-021-10011-5.

[19] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[20] C. S. Hong and T. G. Oh, "TPR-TNR plot for confusion matrix," *Commun. Stat. Appl. Methods*, vol. 28, no. 2, pp. 161–169, 2021, doi: 10.29220/CSAM.2021.28.2.161.

[21] J. Shaikh and R. Patil, "Fake news detection using machine learning," *Proc. - 2020 IEEE Int. Symp. Sustain. Energy, Signal Process. Cyber Secur. iSSSC 2020*, vol. 2020, 2020, doi: 10.1109/iSSSC50941.2020.9358890.

[22] Dr. WIlliam H. Wolberg, "Breast Cancer Wisconsin (Original) Data Set," *UCI Machine Learning Repository*, 2016. https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

[23] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

[24] AMS, "What is breast cancer ? What causes breast cancer ?," *Am. Cancer Soc.*, pp. 1–13, 2015, [Online]. Available: American Cancer Society, "Breast Cancer What is breast cancer ?," Am. Cancer Soc. Cancer Facts Fig. Atlanta, Ga Am. Cancerhttp://www.cancer.org/cancer/breast-cancer/about/what-is-breast-cancer.html.

[25] P. Yadav and C. C. Mandal, "Bioimaging: Usefulness in Modern Day Research," in *Practical Approach to Mammalian Cell and Organ Culture*, Springer, 2023, pp. 1–26. doi: 10.1007/978-981-19-1731-8_23-1.

[26] W. Zhang, S. Chen, F. Cao, L. Chen, and H. Chen, "A multicenter, randomized, controlled study of the breast biopsy and circumferential excision system for breast lesions," *Clin. Breast Cancer*, 2023.

[27] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[28] M. A. Wani, P. Garg, and K. K. Roy, "Machine learning-enabled predictive modeling to precisely identify the antimicrobial peptides," *Med. Biol. Eng. Comput.*, vol. 59, no. 11–12, pp. 2397–2408, 2021, doi: 10.1007/s11517-021-02443-6.