

# Breast Cancer Classification based on Ultrasound Images using the Support Vector Machine (SVM) Algorithm

<sup>1</sup>Nurazmi Aprilia, <sup>2</sup>Rumini\*, <sup>3</sup>Tri Susanto

<sup>1,2,3</sup>Informatics, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta  
Jl. Padjajaran, Ring Road Utara, Condongcatur, Depok, Sleman, Yogyakarta, Indonesia 55283

\*e-mail: [rumini@amikom.ac.id](mailto:rumini@amikom.ac.id)

(received: 25 April 2024, revised: 27 April 2024, accepted: 21 June 2024)

## Abstract

According to statistics from the Global Burden of Cancer Study (Globocon) of the World Health Organization (WHO), cancer, particularly breast cancer, is a severe health issue in Indonesia with 68,858 new cases and 22,000 deaths recorded in 2020. Ultrasonography (USG) technology is acknowledged as one of the potentials to support early detection, which is vital in reducing mortality from breast cancer. This study focuses on classifying ultrasound images using the Support Vector Machine (SVM) algorithm, GLCM feature extraction, Min-Max normalization, and Mutual Information with SelectKBest Feature Selection. From several experiments using the SVM algorithm with various combinations of parameter values that have been set and different Tests, namely using a Train/Test Split with a proportion of 80/20 and K-Fold Cross Validation, it shows that the SVM algorithm is capable of classifying ultrasound images of breast cancer. into two categories (Benign Tumor and Malignant Tumor) with the same maximum accuracy of 79% after applying the SMOTE Balancing Data technique or without using the Balancing Data technique. As a result, the Support Vector Machine (SVM) algorithm has the potential to be an effective model for identifying breast cancer ultrasound images, both on data from the original set that has not been balanced and data from the set that has been balanced.

**Keywords:** maximum breast cancer, ultrasound images, classification, support vector machine.

## 1 Introduction

Cancer is a significant health issue in Indonesia and is ranked second in causes of death after cardiovascular disease[1]. One of the many cancer causes of death is breast cancer which is generally experienced by women[2]. Statistical data from the World Health Organization (WHO) Global Burden of Cancer Study (Globocon), recorded 68,858 new cases of breast cancer or 16.6% of the total 392,914 new cases of cancer reported in Indonesia with a death toll of more than 22 thousand people in 2020 [1][3].

Early detection of breast cancer is an important step to reduce death rates that involves identifying early signs before the cancer spreads, with the aim of increasing the chances of cure and reducing the risk of complications[4]. One method that can be used to support early detection of breast cancer is ultrasound technology. Breast ultrasound (USG) uses sound waves to create images of the breasts. This helps detect changes such as cysts or lumps that are difficult to see on a mammogram. Although not a primary screening, ultrasound is effective, more affordable, and safe without radiation[5].

Digital image classification is the process of grouping images based on visual characteristics. This is used in fields such as medicine, where an example is recognizing breast cancer in ultrasound images with machine learning algorithms that process image features as input to build classification models. Several previous studies [6][7][8], has applied one of the Machine Learning algorithms such as the Support Vector Machine (SVM) algorithm to classify digital images in different cases and obtained quite good accuracy values. So from the background explained previously, this research will apply the Support Vector Machine (SVM) algorithm to determine the level of accuracy and whether this algorithm is good enough to be applied in classifying digital images, namely ultrasound images in breast cancer..

## 2 Literature Review

There is several literature from previous research that has classified breast cancer using various machine learning algorithms.

Research [9], implementing Machine Learning methods, namely Support Vector Machine (SVM) and Decision Tree (DT), to detect breast cancer in women. The data used in the study came from the Gynecology Department of the University Hospital Center of Coimbra (CHUC). The research results show that the SVM algorithm with feature selection provides the best classification results, with an accuracy of 87.5%, sensitivity of 90%, and specificity of 85%.

Research [10], focuses on improving the performance of the Computer-Aided Diagnosis (CAD) system in classifying benign and malignant tumors in breast cancer mammogram images. The method developed uses feature extraction with Gray Level Co-Occurrence Matrix (GLCM) and classification using Support Vector Machine (SVM). The trial was carried out using the DDSM database which contains 256 abnormal images, with accuracy results of 83.59%, sensitivity of 87.58%, and specificity of 76.84%. Apart from that, the AUC (Area Under the Curve) value obtained reached 0.98%.

Research [11], focuses on an approach by extracting features using Discrete Cosine Transformation (DCT) directly from breast cancer ultrasound images, then compared with 4 classification methods, namely Linear Regression, Random Forest, Decision Tree, and Support Vector Machine. The results of this research show that this approach has the best accuracy using the Random Forest algorithm with an accuracy of around 84% in classifying BUSI ultrasound images (Breast Ultrasound Images).

Research [12], Classifying benign and malignant tumors in breast cancer mammography images based on texture characteristics using histograms and Gray Level Co-Occurrence Matrix (GLCM) as well as the Naive Bayes method. This research resulted in an accuracy of 80%, sensitivity of 90%, and specificity of 70%.

### **Gray Level Co-Occurrence Matrix (GLCM)**

According to Rao et al. [13], *Gray Level Co-Occurrence Matrix* (GLCM) is a matrix representation used to identify the extent to which certain pairs of pixels appear at certain distances and angles in an image. GLCM is used to compute various features of this matrix. This approach has many applications including image classification, texture pattern recognition, image segmentation, object identification, and color analysis in images.

There are many texture characteristics that can be extracted from the co-occurrence matrix according to the concept proposed by Haralick [13]. However, this research focuses on six main attributes in GLCM that are often used in analysis, that is ASM (*Angular Second Moment*), *Contrast*, *Correlation*, *Dissimilarity*, *Energy*, and *Homogeneity*.

### **Synthetic Minority Over-sampling Technique (SMOTE)**

SMOTE is an oversampling technique that creates synthetic data in a minority class by taking examples from that class, finding the nearest neighbors with k-nearest neighbors, and combining the differences between examples and neighbors to avoid excessive overfitting[14]. The algorithm process in SMOTE begins by calculating the difference between the feature vector in the minority class and the nearest neighbor value of the minority class. Then, this difference value is multiplied by a random number in the range between 0 and 1. The results of this calculation are then added back to the original feature vector, producing a new feature vector [15].

### **Mutual Information**

Mutual Information (MI) in Feature Selection it is used to measure how much information a feature has, helping to assess the feature's contribution to the model's ability to carry out accurate classification [16]. The Mutual Information (MI) value between two random variables is always positive, indicating the level of dependence. An MI of zero indicates independence, while a high MI indicates a strong level of dependence between the two variables [17].

### **Support Vector Machine (SVM)**

*Support Vector Machine* (SVM) is the algorithm used for classification. This method looks for the largest margin in the form of a hyperplane (vector line) as the dividing boundary between two data classes in the case of binary classification. Although originally designed for two-class problems, SVMs have evolved to handle multi-class classification problems by combining multiple binary classifiers [18][19]. The SVM algorithm searches for the optimal hyperplane by measuring the

distance and finding the maximum point. The accuracy of the SVM model depends on the kernel and parameters. SVM is divided into linear SVM which separates data linearly and non-linear SVM which uses kernel tricks in high-dimensional space [20].

Several types of kernels commonly used in the SVM algorithm are as follows:

1. Linear Kernels

Linear kernels are a very basic type of kernel function. Linear kernels are used when data can be separated linearly. More suitable for data with many features, as mapping to a higher dimensional space does not always help performance, especially in classification. SVM uses a linear kernel by default, where the data is separated by a hyperplane [21].

2. Polynomial Kernels

Polynomial kernels are a more general form of linear kernels and are used to measure the similarity between training sample vectors in a feature space. This kernel is suitable for normalized datasets.

$$K(x_i, x_j) = ((x_i x_j) + c)^d \tag{1}$$

The parameter value d in equation (1) affects the degree of the hyperplane curve in the polynomial function. The higher the d value, the more curved the hyperplane line becomes, which can make the accuracy unstable [22].

3. Radial Basic Function Kernels (RBF)

The Gaussian Radial Basis Function (RBF) kernel is used to classify data that cannot be separated linearly. With the right parameters, RBF can provide good performance and reduce errors in model training. RBF uses a feature space with an unlimited number of dimensions that can be adjusted by parameters, so it can handle complex data, especially when the data does not have a clear linear pattern [23].

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{2}$$

### 3 Research Method

This research uses quantitative research methods, which is an approach that utilizes data in the form of numbers to analyze and compile scientific information based on numerical data [24].

The research flow is used as a basic step for researchers in building a breast cancer classification model based on ultrasound images using the Support Vector Machine (SVM) algorithm. A general overview of the flow of this research can be seen in Figure 1 below.

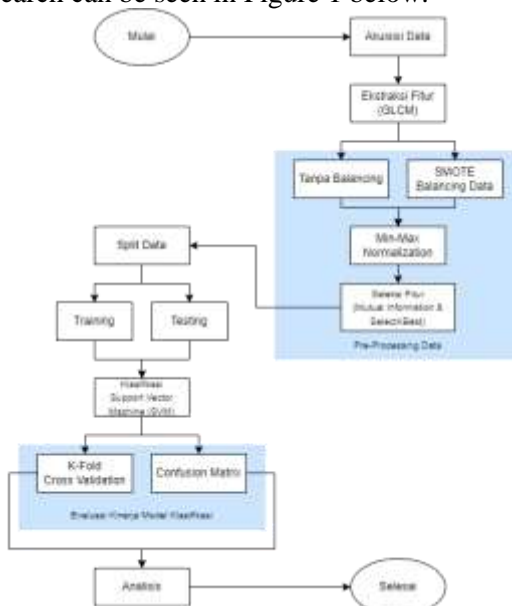


Figure 1. Research flow

1. Data Acquisition

Data collection is the first step in data analysis and machine learning, involving extracting information from various sources relevant to the research objectives. In this study, the dataset

used is BUSI (Breast Ultrasound Images) which consists of 647 ultrasound images, available at <https://www.kaggle.com/>. This dataset was published by Fahmi, et al. in 2020 in an international journal.

2. Feature Extraction (GLCM)

The results of this feature extraction can be used in the classification process. In the feature extraction stage in this research, the method used was GLCM (Gray Level Co-Occurrence Matrix) calculation.

3. Pre-Processing Data

In the pre-processing stage of this research, there are two experiments that will be carried out. The first experiment uses a dataset without Balancing Data, immediately proceeds to the normalization stage with Min-Max Normalization, and Feature Selection using Mutual Information & SelectKBest. The second experiment uses a dataset that has been "balanced" first with the SMOTE method, before proceeding to the next pre-processing stage. The aim of this pre-processing is to prepare a dataset with better quality for further analysis and classification processes.

4. Split Data

At this stage, the dataset is divided into training data (80%) and testing data (20%) at this stage. Training data is used to train the model, so that it can understand the patterns and relationships of features in the dataset. Test data is used to test the performance of a trained model, measuring the model's ability to predict data that has never been seen before.

5. Building a Classification Model (SVM)

In the development stage of the classification model, the Support Vector Machine (SVM) method is used to predict cancer classes: Benign (class 0) and Malignant (class 1). Grid Search is used to find the best hyperparameters that maximize the performance of the SVM model including the C parameter (error penalty), kernel type (kernel function), and gamma (kernel coefficient).

6. Evaluation of Classification Model Performance

The algorithm evaluation stage involves measuring the performance of the classification model. K-Fold Cross Validation and Confusion Matrix are used to generate evaluation metrics, including accuracy, precision and sensitivity. The results of these metrics are then compared to determine the best evaluation.

7. Analysis

The evaluation results analysis stage involves understanding and interpreting the evaluation metrics from the model or experiment. It involves analyzing and comparing metrics such as accuracy, precision, and sensitivity to gain an understanding of the model performance or results obtained.

## 4 Results and Analysis

### Data Acquisition

The data used in this research is the Breast Ultrasound Image (BUSI) Dataset. This data was taken from Kaggle with the dataset URL <https://www.kaggle.com/datasets/aryashah2k/breast-ultrasound-images-dataset>. The dataset taken consists of a total of 647 ultrasound images, with 437 images representing benign tumors (Benign) and 210 images representing malignant tumors (Malignant). A visual representation of this dataset can be observed in Figure 2.




Figure 2. Examples of benign images

### Feature Extraction Using GLCM

The feature extraction stage uses the GLCM (Gray Level Co-Occurrence Matrix) method to produce texture features such as dissimilarity, correlation, homogeneity, contrast, ASM, and energy. GLCM takes into account the relationship between two neighboring pixels that have gray intensity, taking into account distance and angle. There are four corners used : 0°, 45°, 90°, dan 135°. The results of this feature extraction are exemplified in Table 1.

**Table 1. Example of benign image extraction results**

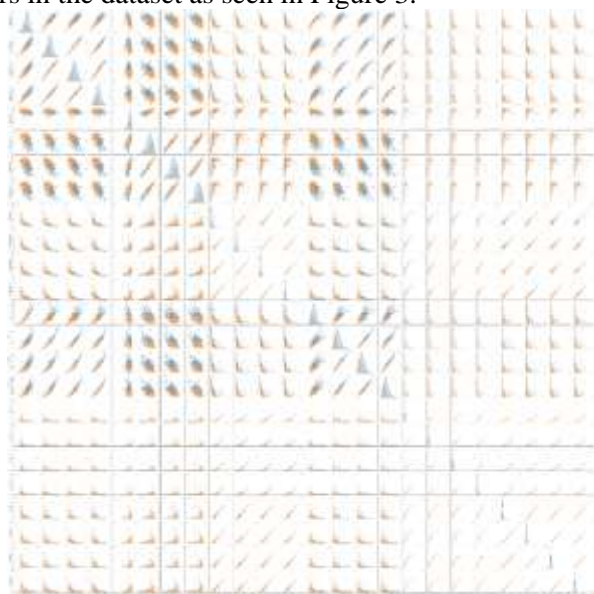
Citra	Fitur			
 <p><i>Benign</i></p>	<b>Disimilaritas 0°, 45°, 90°, 135°</b>			
	18.04710458	26.81366366	29.50699505	29.19144144
	<b>Korelasi 0°, 45°, 90°, 135°</b>			
	0.81931475	0.5720866	0.507454874	0.528226284
	<b>Homogenitas 0°, 45°, 90°, 135°</b>			
	0.062115488	0.042354769	0.034186278	0.034966956
	<b>Kontras 0°, 45°, 90°, 135°</b>			
	581.1205704	1325.650601	1493.950306	1455.512763
	<b>ASM 0°, 45°, 90°, 135°</b>			
	0.00017996	0.000152946	0.000144212	0.000146768
<b>Energi 0°, 45°, 90°, 135°</b>				
0.013414903	0.012367123	0.012008815	0.0121148,1	

After that, the GLCM calculation results are saved in .csv file format. The researcher changed the labels manually using Excel to produce a division of class types into 2 classes for each image which can be seen in Table 2.

**Table 2. Labeling results**

Image Type	Label	Data Amount
<i>Benign</i>	0	437
<i>Malignant</i>	1	210
Total number		647

After doing manual labeling via Excel, the .csv file is uploaded to Google Colaboratory for the Pre-Processing stage. However, before continuing, visualization is carried out to understand the relationship between variables or features in the dataset and to gain insight into the correlation, distribution and patterns of the data as a whole. This visualization includes the patterns of all combinations of feature pairs in the dataset as seen in Figure 3.



**Figure 3. Overall data visualization**



Figure 3 displays the relationship between feature combinations and target variables in the BUSI dataset after feature extraction using GLCM, showing the various data distributions and patterns.

### Data Balancing Using SMOTE

Table 2 indicates an imbalance between classes in the breast cancer dataset, which can impact the quality of classification. Therefore, the Synthetic Minority Over-Sampling Technique (SMOTE) method is used to create synthetic data in the minority class so that the number is equivalent to the majority class. The goal is to improve classification quality by ensuring the model is not biased towards the majority class. After applying the SMOTE technique, the amount of data in each class is now the same as can be seen in Table 3.

Table 3. SMOTE results

Image Type	Label	Data Amount
Benign	0	437
Malignant	1	437
Total number		874

From Table 3, it can be concluded that to balance the minority class with the majority class, 227 synthetic data were added to the Malignant class.

### Data Normalization Using Min-Max

The results of balancing the dataset still show differences in varying attribute scales. This difference can affect the performance of the Machine Learning model in carrying out optimal classification. Therefore, it is important to standardize, ensuring that attributes have a uniform scale when building Machine Learning models. In this context, the Min-Max normalization technique is used to change the attributes so that they have values in the range 0 to 1. The results of this standardization process are shown in Figure 4.

```
[12] print(X_normalized)

[[0.68694487 0.56711034 0.62096162 ... 0.00269413 0.00244153 0.0023263 ]
 [0.44975959 0.29531779 0.3310575 ... 0.01507405 0.01419402 0.01381513]
 [0.5101868 0.44209144 0.44740474 ... 0.00172912 0.00185164 0.00191195]
 ...
 [0.25842643 0.18040252 0.17860196 ... 0.11532015 0.11812077 0.11779661]
 [0.65305051 0.60811572 0.62595034 ... 0.00436323 0.00428014 0.00440467]
 [0.4662313 0.30325793 0.29700447 ... 0.01007888 0.00995762 0.01018585]]
```

Figure 4. Balanced min-max normalization data results

### Feature Selection Using Mutual Information with SelectKBest

At the feature selection stage, the Mutual Information method with SelectKBest was used to improve the quality and efficiency of data analysis and modeling. This method calculates the Mutual Information score between each feature variable and the target variable (label) in the dataset. The results of calculating the Mutual Information score can be seen in Figure 5.

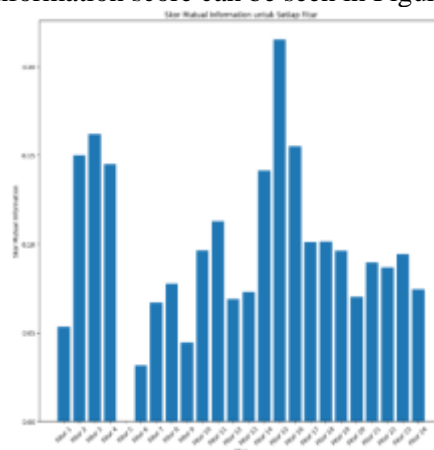


Figure 5. Balanced data mutual information score

Features with higher Mutual Information scores are considered more informative and important in building a good model or performing accurate analysis. In this research, SelectKBest was used to select k features, and the researcher selected k=6 features with the highest Mutual Information score. In this way, only the most informative subset of features is retained. From Figure 5, the six features with the highest Mutual Information score are Feature 15 (contrast 90°), Feature 3 (dissimilarity 90°), Feature 16 (contrast 135°), Feature 2 (dissimilarity 45°), Feature 4 (dissimilarity 135°), and Feature 14 (contrast 45°).

### Train/Test Split Data

In this step, the dataset is divided into training data (Training) and test data (Testing) with a ratio of 80/20. The aim is to test the model's ability to predict 130 or 175 test data. An overview of the data distribution can be found in Table 4.

**Tabel 4. Train/test split**

Techniq	Training Data (80%)	Testing Data (20%)	Amount (100%)
No Balancing	517	130	647
SMOTE Balancing	699	175	874

### Building Classification Models Using SVM

This research implements Grid Search in the Support Vector Machine (SVM) algorithm to find the optimal combination of C, kernel and gamma parameters to improve the performance of the SVM model. This approach makes it easier for researchers to experiment efficiently without having to test one by one. The results from various tests with grid parameters determined by researchers can be seen in the form of a classification report in Table 5.

**Table 5. Comparison of classification model evaluation results**

Method	Parameter Grid	Best Parameters	Accuracy	Precision	Recall
SVM without Balancing	'kernel':['linear', 'rbf', 'poly']	'C': 100,	0.78	0.77	0.78
	'C': [0.01, 0.1, 1, 10, 100]	'kernel': 'rbf'			
	Kernel= Linear	'C': 100	0.76	0.75	0.76
	'C': [0.01, 0.1, 1, 10, 100]				
	Kernel= RBF	'C': 100,	0.79	0.79	0.79
	'C': [0.01, 0.1, 1, 10, 100]	'gamma': 10			
	'gamma': [0.001, 0.01, 0.1, 1, 10, 100]				
	Kernel= Polynomial	'C': 100,	0.74	0.73	0.74
	'degree': [2, 3, 4]	'degree': 2			
	'C': [0.01, 0.1, 1, 10, 100]				
SVM with SMOTE Balancing	'kernel':['linear', 'rbf', 'poly']	'C': 100,	0.77	0.77	0.77
	'C': [0.01, 0.1, 1, 10, 100]	'kernel': 'rbf'			
	Kernel= Linear	'C': 100	0.73	0.73	0.73
	'C': [0.01, 0.1, 1, 10, 100]				
	Kernel= RBF	'C': 10,	0.78	0.78	0.78
	'C': [0.01, 0.1, 1, 10, 100]	'gamma': 100			
	'gamma': [0.001, 0.01, 0.1, 1, 10, 100]				
	Kernel= Polynomial	'C': 100,	0.77	0.77	0.77
	'degree': [2, 3, 4]	'degree': 4			
	'C': [0.01, 0.1, 1, 10, 100]				

From Table 5, it can be concluded that the test results with various parameter values show quite varied results. The results of 2 different experiments with four tests each, namely using original data without SMOTE and using data that has been carried out by SMOTE, show that the Support Vector Machine algorithm with the 'RBF' kernel, parameters C=100, and Gamma=10 on the dataset without

SMOTE provides the best evaluation, namely achieving an accuracy of 79%, precision of 79%, and sensitivity of 79% which is better than other parameters.

### Model Performance Evaluation Using K-Fold Cross-Validation

The next stage involves applying K-Fold Cross Validation to compare the performance of previously tested models. In K-Fold Cross Validation, the dataset is divided into k subsets which in this study uses 5 folds. The same experiments and tests as before will be applied. The performance evaluation results of the SVM model using K-Fold Cross Validation can be observed in Table 6.

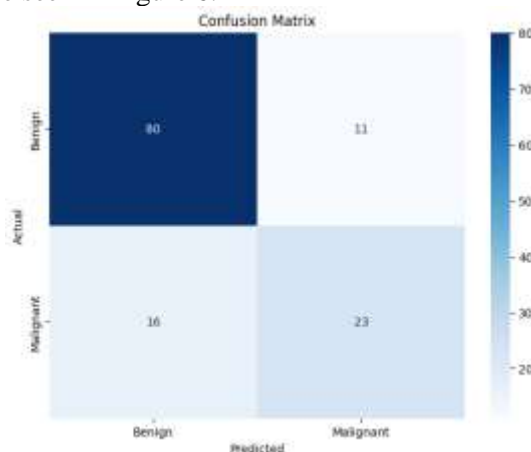
**Table 6. Comparison of evaluation results using k-fold cross validation**

Method	Parameter	Accuracy	Precision	Recall
SVM No <i>Balancing</i> <i>Kfold=5</i>	'C': 100, 'kernel': 'rbf'	0.77	0.77	0.77
	Kernel= Linear 'C': 100	0.76	0.76	0.76
	Kernel= RBF 'C': 100, 'gamma': 10	0.78	0.77	0.78
	Kernel= Polynomial 'C': 100, 'degree': 2	0.75	0.74	0.75
	SVM with <i>SMOTE</i> <i>Balancing</i> <i>Kfold=5</i>	'C': 100, 'kernel': 'rbf'	0.76	0.76
	Kernel= Linear 'C': 100	0.74	0.75	0.74
	Kernel= RBF 'C': 10, 'gamma': 100	0.79	0.79	0.79
	Kernel= Polynomial 'C': 100, 'degree': 4	0.76	0.78	0.76

From Table 6, it can be seen that the test results with various parameter values provide varying evaluation results. Test results using K-Fold Cross Validation show that the Support Vector Machine algorithm with the 'RBF' kernel and parameters C=10 and Gamma=100 on the dataset that has been carried out by SMOTE Balancing Data, provides the best evaluation. This model achieves an accuracy of 79%, precision of 79%, and recall of 79%, which is superior to other parameters.

### Model Performance Evaluation Using Confusion Matrix

Based on a series of experiments and tests to classify breast cancer using the Train/Test Split 80/20 method, the best evaluation results were found using the Support Vector Machine kernel 'RBF' algorithm, parameters C=100, and Gamma=10. The best performance results from all previous classification stages using the Train/Test Split method are displayed in the form of a Confusion Matrix evaluation which can be seen in Figure 6.



**Figure 6. Confusion matrix result**



Figure 4.6 shows the Confusion Matrix results from the Support Vector Machine algorithm experiment using the 'RBF' kernel with parameter values C='100' and Gamma='10' on a dataset where SMOTE Data Balancing was not carried out. From these results it can be seen that :

1. A total of 80 class data in the Benign class (Benign Tumors) and 23 data in the Malignant class (Malignant Tumors) were predicted to be correct in total.
2. A total of 11 Benign class data were predicted incorrectly, namely as the Malignant class (Malignant Tumor).
3. A total of 16 Malignant class data were predicted incorrectly, namely Benign class).

### Analysis

From the results of two different types of testing, it can be concluded that in both tests using Train/Test Split and K-Fold Cross Validation, the highest evaluation value achieved is the same, but the application of the method and parameter values are quite different. A comparison of the best evaluation values of the two methods can be seen in Table 7.

**Table 7. Comparison of test types**

<b>Evaluation</b>	<b>Method</b>	<b>Parameter</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
<i>Train/Test Split</i>	SVM	Kernel= RBF	0.79	0.79	0.79
<i>80/20</i>	<i>tanpa Balancing</i>	'C': 100, 'gamma': 10			
<i>K-Fold Cross Validation</i>	SVM	Kernel= RBF	0.79	0.79	0.79
<i>KFold=5</i>	dengan <i>SMOTE Balancing</i>	'C': 10, 'gamma': 100			

Based on the comparison in Table 7 above, it can be seen and concluded that the type of testing either uses the Train/Test Split 80/20 method using a dataset without Balancing using an SVM model with the kernel 'RBF', parameters C=100, and Gamma=10 or K-Fold Cross Validation using a dataset carried out by SMOTE Balancing Data using an SVM model with kernel 'RBF' parameters C='10', and Gamma='100' has the same highest model evaluation value, namely with accuracy, precision and sensitivity of 79 %.

Furthermore, this study was compared with previous research which used a similar dataset, namely the Breast Ultrasound Image (BUSI) dataset with two classes, Benign (benign tumors) and Malignant (malignant tumors), totaling 210 data for each class with details that can be observed in table 8.

**Table 8. Research comparison**

<b>Researcher Name</b>	<b>Dataset</b>	<b>Method</b>	<b>Accuracy Results</b>
Heru Arwoko	<i>Breast Ultrasound Image (BUSI) Dataset</i>	LR, RF, DT, SVM with DCT feature extraction	LR = 74% RF = 84% DT = 73% <b>SVM = 77%</b>
Proposed research	<i>Breast Ultrasound Image (BUSI) Dataset</i>	SVM with GLCM feature extraction and Mutual Information & SelectKBest feature selection	<b>SVM = 79%</b>

From the comparison of results with previous research using the BUSI dataset. In previous research, SVM with DCT feature extraction achieved 77% accuracy in breast cancer classification. In this study, with a larger dataset (647 and 874 data), after several experiments and parameter testing, SVM achieved a maximum accuracy of 79%. These results show an increase in accuracy of 2% by

applying the SVM algorithm as well as the GLCM feature extraction technique, and the Mutual Information and SelectKBest feature selection methods.

## 5 Conclusion

Based on the results of the analysis that has been carried out previously, it can be concluded that the Support Vector Machine algorithm implemented using GLCM feature extraction, Min-Max Normalization, as well as Mutual Information and SelectKBest feature selection produces quite good scores in classifying two classes (Benign and Malignant) of images. ultrasound in breast cancer. This can be seen from several experiments and tests that have been carried out, the maximum accuracy value is the same as 79% by adding the SMOTE Balancing Data technique or without adding the Balancing Data technique. Furthermore, by comparing previous research which used the same type of dataset as this research, namely the Breast Ultrasound Images (BUSI) Dataset, it was found that the accuracy value increased by 2% by applying the Support Vector Machine Algorithm using the 'RBF' Kernel, GLCM feature extraction technique, Min-Max Normalization, as well as selection of Mutual Information and SelectKBest features.

## Reference

- [1] C. M. Annur, "Kanker Payudara, Penyakit Kanker Paling Banyak Dialami Masyarakat Indonesia," 2022. <https://databoks.katadata.co.id/datapublish/2022/10/11/kanker-payudara-penyakit-kanker-paling-banyak-dialami-masyarakat-indonesia>
- [2] M. A. Rohman, P. Mudjirahardjo, and M. A. Muslim, "Implementasi *Filter Gray Level Co-Occurance Matriks* Terhadap Sistem Klasifikasi Kanker Payudara Dengan Metode Convolutional Neural Network," *Transmisi*, vol. 23, no. 4, pp. 160–168, 2021, doi: 10.14710/transmisi.23.4.160-168.
- [3] "Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan," *Kementerian Kesehatan RI*, 2022. [https://www.kemkes.go.id/article/view/22020400002/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html#:~:text=Data Globocan tahun 2020%2C jumlah,dari 22 ribu jiwa kasus](https://www.kemkes.go.id/article/view/22020400002/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.html#:~:text=Data%20Globocan%20tahun%202020%2C%20jumlah,dari%2022%20ribu%20jiwa%20kasus)
- [4] "American Cancer Society Recommendations for Prostate Cancer Early Detection," *American Cancer Society*, 2021. <https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging/acs-recommendations.html>
- [5] American Cancer Society, "Breast Cancer Early Detection and Diagnosis American Cancer Society Recommendations for the Early Detection of Breast Cancer," *Am. Cancer Soc.*, pp. 1–55, 2016, [Online]. Available: <https://www.cancer.org/content/dam/CRC/PDF/Public/8579.00.pdf>
- [6] R. Suhendra, I. Juliwardi, and S. Sanusi, "Identifikasi dan Klasifikasi Penyakit Daun Jagung Menggunakan *Support Vector Machine*," *J. Teknol. Inf.*, vol. 1, no. 1, pp. 29–35, 2022, doi: 10.35308/v1i1.5520.
- [7] Y. Amrozi, D. Yuliati, A. Susilo, N. Novianto, and R. Ramadhan, "Klasifikasi Jenis Buah Pisang Berdasarkan Citra Warna dengan Metode SVM," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 11, no. 3, pp. 394–399, 2022, doi: 10.32736/sisfokom.v11i3.1502.
- [8] A. G. Sooi, P. A. Nani, N. M. R. Mamulak, C. O. Sianturi, S. C. Sianturi, and A. H. Mondolang, "Klasifikasi Citra Daun Anggur Menggunakan SVM Kernel Linear," *JOINTECS (Journal Inf. Technol. Comput. Sci.)*, vol. 8, no. 1, p. 19, 2023, doi: 10.31328/jointecs.v8i1.4496.
- [9] H. Imaduddin, B. A. Hermansyah, and F. A. Salsabilla B, "Comparison of Support Vector Machine and Decision Tree Methods in the Classification of Breast Cancer," *Cybersp. J. Pendidik. Teknol. Inf.*, vol. 5, no. 1, p. 22, 2021, doi: 10.22373/cj.v5i1.8805.
- [10] L. M. Wisudawati, "Klasifikasi Tumor Jinak Dan Tumor Ganas Pada Citra Mammogram Menggunakan *Gray Level Co-Occurrence Matrix (GlcM)* Dan *Support Vector Machine (Svm)*," *J. Ilm. Inform. Komput.*, vol. 26, no. 2, pp. 176–186, 2021, doi: 10.35760/ik.2021.v26i2.4897.
- [11] H. Arwoko, "Klasifikasi Kanker Payudara pada Citra Ultrasound Menggunakan Fitur <http://sistemasi.ftik.unisi.ac.id>

- Koefisien Discrete Cosine Transform (DCT),” *Pros. HUBISINTEK*, vol. 2, no. 1, p. 451, 2022.
- [12] A. D. Achmad, “Klasifikasi Breast Cancer Menggunakan Metode Logistic Regression,” *Jtriste*, vol. 9, no. 1, pp. 143–148, 2022.
- [13] G. T. Situmorang, A. W. Widodo, and M. A. Rahman, “Penerapan Metode *Gray Level Co-Occurrence Matrix* ( GLCM ) untuk ekstraksi ciri pada telapak tangan,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 5, pp. 4710–4716, 2019.
- [14] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
- [15] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, “Improving accuracy of students’ final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique,” *Decis. Anal.*, vol. 2, no. 1, p. 1, 2015, doi: 10.1186/s40165-014-0010-2.
- [16] A. Hanafi, A. Adiwijaya, and W. Astuti, “Klasifikasi Multi Label pada Hadis Bukhari Terjemahan Bahasa Indonesia Menggunakan Mutual Information dan k-Nearest Neighbor,” *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, pp. 357–364, 2020, doi: 10.32736/sisfokom.v9i3.980.
- [17] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, “Part Mutual Information for Quantifying Direct Associations In Networks,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 18, pp. 5130–5135, 2016, doi: 10.1073/pnas.1522586113.
- [18] J. Z. Liang, “SVM multi-classifier and web document classification,” *Proc. 2004 Int. Conf. Mach. Learn. Cybern.*, vol. 3, pp. 1347–1351, 2004, doi: 10.1109/icmlc.2004.1381982.
- [19] M. N. Rakhmasari, “Implementasi Metode *Support Vector Machine* (SVM) pada Klasifikasi dan Karakterisasi Tingkat Kedalaman Kemiskinan Provinsi Jawa Timur,” *Univ. Islam Negeri Malang*, pp. 1–71, 2022.
- [20] “Scikit-Learn Documentation - SVM,” *Scikit-Learn*. <https://scikit-learn.org/stable/modules/svm.html>
- [21] A. Zeputra and F. Utaminigrum, “Perbandingan Akurasi untuk Deteksi Pintu berbasis HOG dengan Klasifikasi SVM menggunakan Kernel Linear , Radial Basis Function dan Polinomial pada Raspberry Pi,” ... *Teknol. Inf. dan Ilmu Komput. e ...*, vol. 5, no. 11, pp. 4746–4757, 2021, [Online]. Available: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10090%0Ahttp://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/10090/4483>
- [22] Z. Liu and H. Xu, “Kernel parameter selection for *Support Vector Machine* classification,” *J. Algorithms Comput. Technol.*, vol. 8, no. 2, pp. 163–177, 2014, doi: 10.1260/1748-3018.8.2.163.
- [23] A. N. Khobragade, M. M. Raghuvanshi, and L. Malik, “Evaluating Kernel Effect on Performance of SVM Classification using Satellite Images,” *Int. J. Sci. Eng. Res.*, vol. 7, no. 3, pp. 742–748, 2016.
- [24] Kusnawi, M. A. F. E. Putra, and J. Ipmawati, “Prediksi harga bahan pokok dengan menggunakan metode forecasting ARIMA melalui open data Kabupaten Sumedang,” *J. Sist. Inf.*, vol. 12, no. 2, pp. 293–307, 2023.