Breast Cancer Classification Procedure Using Machine Learning Techniques

Jerry Dwi Trijoyo Purnomo[1*] and *Dea Restika Augustina* Pratiwi[1]

[1]Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Abstract. Breast cancer is a malignant tumor that attacks breast tissue. This disease can be treated and managed properly if diagnosed at an early stage. An appropriate, fast and effective cancer stage detection algorithm is required so that patients can be treated precisely. In this study, the classification of breast cancer stages will be carried out using several machine learning methods. The number of patients in each stage is unequal or unbalanced as well. Therefore, the oversampling method with SMOTE is applied. The selection of the best parameters is done using 10-fold cross validation on the training data. Next, modeling was carried out using the Neural Network method, and K-Nearest Neighbor on training and testing data which had been oversampled with SMOTE. It was found that the neural network had a higher AUC value than k-Nearest Neighbor, namely 82.3% while k-NN was 80.8%.

1 Introduction

As women age, the risk of breast cancer increases. However, breast cancer in young women tends to be more aggressive and has a higher stage than in older women [1]. A family history of breast cancer is an important risk factor. Women with a family history of breast cancer are significantly more likely to be diagnosed with stage III breast cancer than women without a family history of breast cancer [2]. Research regarding the identification of breast cancer characteristics and epidemiological risk factors based on age and menopause status showed that breast cancer cases for women with premenopausal status were significantly more likely to have stages II, III, or IV compared to women with postmenopausal status. Women who have a history of breast disease, especially those who have had breast cancer before, have a higher risk of suffering from breast cancer a second time. However, second breast cancer detected in the asymptomatic phase (without clinical symptoms) has a better stage than second breast cancer in the symptomatic phase, because it has a smaller tumor size and fewer metastases [3].

Examination to determine the subtype of breast cancer is carried out using immunohistochemistry (IHK), namely to see the presence of estrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptor 2 (HER2). CPI examination plays a role in helping determine predictions of systemic therapy response and prognosis [4]. Dunnwald, Rossing, and Li [5] conducted research with the results that women

* Corresponding author : jerrypurnomo@gmail.com

with ER- / PR+ and ER- / PR- receptor status had larger tumors and were diagnosed with advanced breast cancer when compared with women whose receptor status was ER+ / PR+ and ER+ / PR-. Meanwhile, in research conducted by Seshadri, et al. [6] on 1056 patients with stage I-III breast cancer showed that HER2 receptor status had a significant relationship with breast cancer stage.

Machine learning, which is currently known as an alternative to modern classification methods that has good classification results, has also been applied in several studies with classification accuracy results that are generally better than logistic regression models. Research on the classification of breast cancer malignancies was carried out on Wisconsin City breast cancer patients by Kurniawan & Ivandri [7] by comparing the k-NN and Decision Tree methods, obtaining higher accuracy for the k-NN method, namely 94.71%. This research was conducted to compare the results of breast cancer stage classification using the Neural Network and k-Nearest Neighbor (k-NN) methods in patients at the Surabaya Oncology Hospital. Modeling was carried out using factors thought to influence the stage of breast cancer in patients at the Surabaya Oncology Hospital. The machine learning classification method in data mining works by using past data or training data to form an algorithm which can later be used as a reference for classification of subsequent data. Several data mining classification techniques have proven to be good and produce high accuracy [7]. Neural Network and k-NN classification methods in data mining have several advantages, including the formation of training algorithms that are simple, fast and effective. It is hoped that the best model obtained by this research in classifying breast cancer stages can be used as a reference for medical personnel in detecting or diagnosing the severity of breast cancer in patients quickly and effectively.

2 Methods

In this section, we briefly illustrate the literature review and the proposed method used in this paper.

2.1 Synthetic Minority Oversampling Technique (SMOTE)

One method that can be used to overcome the problem of imbalanced data or data that has unequal proportions between target class categories is by oversampling. Addition or replication of data or what is usually called oversampling can be done in several ways, one of which is Synthetic Minority Oversampling Technique (SMOTE). This method was discovered by Chawla et. al [8] where the number of samples was increased in the minor class to make it equal to the major class by generating synthetic data based on k-nearest neighbors which is included in the group of non-parametric statistical methods. Nearest neighbors are selected based on the euclidean distance between the two data. Determining the number of replications is adjusted to the amount of data in the major class and the number of k in the nearest neighbor. If the number of replications is n, then the number of k is (n-1). For instance, there are two data structures with dimension p, namely $x^T = \left[x_1, x_2, ..., x_p \right]$ and $y^T = \left[y_1, y_2, ..., y_p \right]$, then the Euclidean distance $d\left(x, y\right)$ between the two data can be calculated with Equation 1.

$$d\left(x, y\right) = \sqrt{\left(x_1 - y_1\right)^2 + \left(x_2 - y_2\right)^2 + ... + \left(x_p - y_p\right)^2} \quad (1)$$

Further, data replication is carried out using Equation 2.

$$x_{syn} = x_i + \left(x_{knn} - x_i\right) \times \tau \tag{2}$$

where x_{syn} be the data from replication results, x_i be the data to be replicated, x_{knn} be the data that has the closest distance from the data to be replicated, and τ be the random number between 0 and 1.

2.2 Neural Network

Artificial Neural Networks or Neural Networks is an information processing system with characteristic capabilities that resemble biological networks in humans. Three types of components of biological networks that are similar to Neural Networks components are dendrites, axons, and soma. Dendrites receive signals from other neurons, where signals are electrical impulses emitted in the synaptic gap when chemical processes take place. The activity of chemical transmitters that change incoming signals is similar to the activity of weights in Neural Networks. Soma or body cells (body cells) calculate incoming signals. When input is received, the body cells will release or provide stimulation, then the stimulation sends signals via axons to other body cells [9]. The shape of human biological tissue is shown in Figure 1.

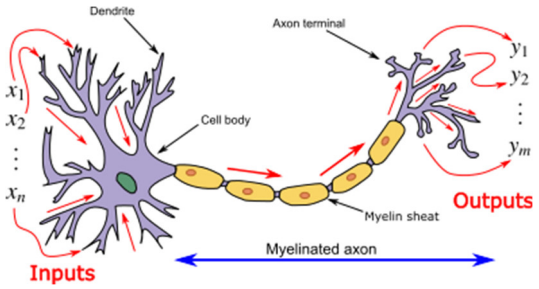

Fig. 1. Biological Neurons in the Human Nervous System

Neural networks were developed as a generalization of mathematical models of biological neural networks in the human brain with the following assumptions [9]:

1. Information processing occurs in several simple elements called neurons.
2. Signals are received by neurons via connections.
3. Each link has its own weight which is the weight of the signal that passes through that link.
4. Each neuron uses an activation function on its net input (sum of weight single input) to determine the output signal.

Meanwhile, the characteristics of neural networks include the following.

1. The pattern of connections between neurons is called architecture.
2. Methods for determining weights on connections are called training, learning, and algorithms.
3. Has activation function.

Neural networks consist of many simple processing elements called neurons, units, cells, or nodes. Each neuron is connected to other neurons via a direct link, each of which has a weight. Weight describes the information used by the network to solve a problem. Each neuron has an internal state called activation or activity level which is a function of the input received. Specifically, a neuron sends its activation as a signal to some other neurons. For example, a neuron illustrated in Figure 2 receives input from neurons Z_1, Z_2, and Z_3. The activation signals from these neurons are z_1, z_2, and z_3, respectively. The associated weights of Z_1, Z_2, and Z_3 on neuron Y are w_1, w_2, and w_3.

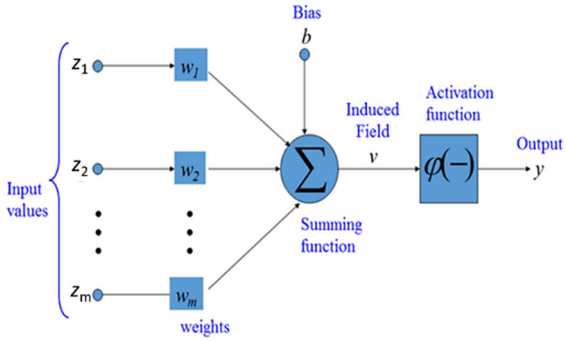

Fig. 2. Simple Neural Network Process Flow

The input to neuron Y is expressed as net input, while the summing function is the weighted sum of signals from neurons Z_1, Z_2, and Z_3 with the formula in Equation (3).

$$\sum = w_1 z_1 + w_2 z_2 + w_3 z_3 \tag{3}$$

The activation signal y of neuron Y is obtained from the net function of its input (Equation (4)),

$$y = f\left(\sum\right) \tag{4}$$

2.3 k-Nearest Neighbour (k-NN)

k-NN is an approach that is simple to implement and is an old method used in classification. Research conducted by Hamamoto et. al. [10] and Alpaydin [11] resulted in the conclusion that k-NN has a high level of efficiency and in some cases provides a high level of accuracy in terms of classification. In other terms, k-NN is one of the methods used in multiclass classification. The working principle of k-NN is to carry out classification based on the proximity of the location (distance) of a data to other data [12]. How close or far a location is (distance) can be calculated using one of the predetermined distance quantities, namely Euclidean distance, Minkowski distance and Mahalanobis distance. However, in its application the Euclidean distance is often used because it has a high level of accuracy and productivity [13]. This Euclidean distance concept treats all variables as independent (uncorrelated). The standard transformation carried out means eliminating the influence of data diversity or in other words all variables will make the same contribution to distance. Euclidean distance is the distance between a straight line that connects objects. The Euclidean distance formula is shown in Equation 5 [14].

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} \left(x_{ir} - x_{jr}\right)^2} \tag{5}$$

x_{ir} be the ith testing data on the rth predictor variable, x_{jr} be the jth training data on the rth predictor variable, d be the Euclidean Distance, and p be the number of predictor variable.

2.4 Cross Validation

There are several methods for evaluating the performance of a model in making predictions through testing data and training data, including holdout and cross-validation [15]. The holdout method uses two-thirds of the data to be used as training data and uses the remainder

as testing data. There is a possibility that the sample taken is not representative, because there is a chance that each class in the data is not represented. To check whether the sample taken is representative or not, namely by balancing the proportions of each class for testing data and training data. If there is a class that is not represented in the training data, the classifier cannot be formed properly to carry out classification in the testing data. Random sampling should pay attention to and ensure that the samples taken are sufficiently representative of each existing class. The cross validation method divides data into *k*-folds or partitions that have an equal number. Each partition acts as training data as well as testing data in turn. The cross-validation method uses one data partition as testing data and the remaining *k*-1 as training data. This procedure continues to repeat until all data partitions have become testing data. This method or procedure is known as *k*-fold cross validation. In this research, the cross validation method was run 10 times ($k = 10$) with different training data, where each data set has the same number as the other data sets. Tests have been carried out using different data and different learning techniques, the conclusion is that 10 folds is the best number of folds to get the best error. The 10 folds cross validation method has become a standard method in machine learning and data mining [15].

2.5 Area Under Curve

One of the basic measures used to measure and evaluate classification performance is sensitivity and specificity. A binary classification model has a pair of sensitivity and specificity. If several classification models are used in a classification case, problems will arise in terms of selecting the best model and method. This is because there are several pairs of sensitivity and specificity of the classification model used. This problem can be overcome by using the ROC (Receiving Operating Characteristic) curve. The ROC curve is a graphical representation of the relationship between sensitivity and specificity [16].

The ROC curve is often used to evaluate classification methods because it has comprehensive and quite good capabilities [17]. In the ROC curve, sensitivity (true positive rate) is plotted as a function of 1 specificity (false positive rate) for different cut-off points. Each point on the ROC curve represents a pair of sensitivity and specificity that corresponds to a particular decision boundary. A test with perfect discrimination has a plot that passes through the upper left corner of the ROC curve (100% sensitivity and 100% specificity. The closer the ROC plot is to the upper left corner, the higher the accuracy of the overall test [18]. The method commonly used to calculate classification performance is to calculate the area under the ROC curve. The area under the ROC curve is usually called the Area Under the ROC Curve (AUC). The AUC value is between 0 and 1. If the AUC value is closer to 1, the classification model formed will be more accurate. A good ROC curve is at the top of the diagonal lines (0,0) and (1,1), so there is no AUC value smaller than 0.5. AUC calculations are carried out using several approaches, the most widely used is the trapezoidal method. The approach is based on a geometric method based on linear interpolation between each point on the ROC curve [19].

$$AUC = \frac{1}{l}\sum_{i=1}^{l} R_i \tag{6}$$

The R_i value in Equation 6 can be obtained from Equation 7.

$$R_i = \frac{n_{ii}}{\sum_{i=1}^{l} n_{il}} \tag{7}$$

The constituent elements in Equation 7 are confusion matrix elements from the classification of four stage categories for breast cancer patients which are illustrated in Table 1.

Table 1. Confusion Matrix of Classification of Breast Cancer Stages

Observation class	Prediction Class				Total
	1	2	3	4	
1	n_{11}	n_{12}	n_{13}	n_{14}	n_1
2	n_{21}	n_{22}	n_{23}	n_{24}	n_2
3	n_{31}	n_{32}	n_{33}	n_{34}	n_3
4	n_{41}	n_{42}	n_{43}	n_{44}	n_4

3 Data Example

The data used in this study is secondary data originating from medical records of breast cancer patients at the Surabaya Oncology Hospital January 2011 – December 2019. The information from the medical records used are factors that are thought to influence the stage of breast cancer in Surabaya Oncology Hospital patients. The total number of medical record data is 382 data from breast cancer patients who are or have undergone treatment at the Surabaya Oncology Hospital. In this study, only data that had complete medical information was used, namely medical records that did not have missing values for all the variables that would be used in this study, so that medical records with incomplete information were eliminated and 294 patient data were obtained. Training and testing data partitioning was carried out using the k-fold cross validation method with k =10. The research variables used in this research include response variables (Y) and predictor variables (X). The response variable in this study is the stage of breast cancer, while the predictor variables include factors that are thought to influence the stage of breast cancer. Table 2 shows a description of the research variables used.

Table 2. Research Variable

Variable	Name	Definition
Y	Stage	: Stage I : Stage II : Stage III : Stage IV
X_1	Age	Age at diagnosis of breast cancer
X_2	Family history of breast cancer	1 : No 2 : Yes
X_3	Hormonal status	1 : Pre-monopouse 2 : Post-monopouse
X_4	History of breast disease	1 : No 2 : Yes
X_5	ER receptor status	1 : Negative 2 : Positive
X_6	ER receptor status	1 : Negative 2 : Positive
X_7	HER2 Status	1 : Negative 2 : Positive
X_8	Long-suffering	Time in months, namely between the patient being diagnosed with breast cancer and the patient's first examination at the Surabaya Oncology Hospital.

Before carrying out oversampling using the SMOTE method, standardization is first carried out on numerical data and one hot encoding on categorical data. Standardization is

carried out by reducing each data on a numerical variable by the average of the data and then dividing it by the standard deviation of the variable. The numerical variables in this study are age and length of suffering. Meanwhile, for other variables of the categorical type, one hot encoding is carried out so that there are as many dummy variables as there are categories in the variable containing the numbers 0 or 1. Standardization and one hot encoding are carried out to prepare the data so that classification can be optimal using both the *k*-NN and neural network methods.

The distribution of breast cancer patients at the Surabaya Oncology Hospital in January 2011 – December 2019 was not balanced in the four stage categories. Stage I, stage II, stage III, and stage 4 had patient proportions of 19%, 44.2%, 35%, and 1.7%, respectively. The number of breast cancer patients diagnosed with stage II and stage III is not too different, however, stage I and stage IV have quite a big difference compared to stages II and III. Stage II has the largest number of patients so it is called the major class. Meanwhile stages I and IV are called minor classes and data synthesis will be carried out using Equation 2 based on nearest neighbors using Euclidean distance in Equation 1. The stage I minor class has a total of 56 patients, so to balance the number of members with the major class, which is 130 people, it is necessary to replicate once. The number of nearest neighbors for each stage I patient data is 0 because it is only replicated once, so the coordinate points of the replicated data will be the same as the coordinate points of the replicated data. Meanwhile, in the minor stage IV class there were only 5 patients. So, to balance with the major class, replication needs to be carried out 25 times and the number of nearest neighbors for each data is 24 data. The coordinate points of the replicated data will be different from the coordinate points of the replicated data. Comparison of the amount and proportion of data before and after replication with SMOTE can be seen in Table 3. Meanwhile, the number of stage III breast cancer patients will

Table 3. Proportion of Stage Categories of Breast Cancer Patients at Surabaya Oncology Hospital Before and After SMOTE

before SMOTE		**after SMOTE**		**Number of Replication**
Major	**Minor**	**Major**	**Minor**	
Stage II 44% (130)	Stage I 19% (56)	Stage II 27% (130)	Stage I 24% (112)	1
	Stage IV 2% (5)		Stage IV 27% (130)	25

remain at 103 patients. This is because the number of patients in stage III is not much different from the major class, namely stage II. Thus, the proportion of patients in each class in the response variable is balanced. The research data, which originally consisted of 294 samples, has now become 475 samples. This data comes from synthetic data generated based on replication using SMOTE to balance the amount of data between major and minor classes. Apart from increasing the amount of data on the response variable, the amount of data on the predictor variable will also increase following the amount of data on the response variable. By balancing the number of patients at each stage, it is hoped that there will be no cases of underfitting or overfitting, resulting in a good level of model fit. After preprocessing, the next step taken to classify breast cancer stages in Surabaya Oncology Hospital patients was to divide the data into training and testing with proportions of 80% and 20% respectively. The division is carried out with stratification so that the number of patients in each stage remains balanced based on training data and testing data. Training data will be used for parameter optimization using stratified cross validation with k = 5, so that the training data will be divided into train set and validation set. This step was taken with the hope that the best model would be formed not only for testing data, but also for patient data that would be predicted in the future. The selection of the best parameters is done by looking at the goodness of the

cross validation validation set. After the best parameters are found, these parameters are applied to the overall training data and testing data and then evaluated again for the goodness of the testing data to determine the best model. Evaluation is carried out by selecting the largest AUC using the formula in Equation 6.

3.1 Classification Using Neural Network

There are several parameters in building a neural network model. In this research, the ReLu (Rectified Linear Unit) activation function was used, the Adam solver, alpha was 0.05, the number of hidden layers was 1, and the number of neurons in the hidden layer was compared between 2 and 10 to get the highest AUC value. 10-cross validation is used to evaluate the model for each number of hidden layers. Based on Table 4, it can be seen that

Table 4. Average AUC of Neural Network Models on Training Data with Cross Validation Before and After SMOTE

Number of Node	before SMOTE		after SMOTE	
	Training Set	Validation Set	Training Set	Validation Set
2	59.7%	58.2%	77.6%	75.9%
3	65.4%	61.2%	81.2%	78.1%
4	62.1%	60.4%	82.8%	80.0%
5	70.5%	67.2%	83.8%	82.4%
6	68.6%	63.9%	84.2%	80.2%
7	67.1%	65.0%	84.3%	82.0%
8	69.4%	60.7%	86.9%	81.7%
9	71.0%	64.2%	86.9%	81.9%
10	70.2%	63.2%	89.6%	79.8%

data that has been SMOTE has a higher AUC than before SMOTE. Apart from that, most of the AUC values in the data before SMOTE were <70% in both the training set and validation set, so they had not yet reached the acceptable category. The data before SMOTE had the highest AUC in the validation set at 67.4%, namely at a number of neurons of 5. The AUC in the training set had reached the acceptable category, namely 70.5%. Similar to the data before SMOTE, the data after SMOTE also has the highest AUC when the number of neurons is 5, with values in the training set and validation set of 83.8% and 82.4% respectively. Then, with the same parameters, namely the ReLu (Rectified Linear Unit) activation function, Adam solver, alpha of 0.05, and the number of hidden layers of 1, the training data is modeled and evaluated using AUC. Table 5 shows that the AUC for training data.

Table 5. AUC Neural Network Model on Training Data Before and After SMOTE with the Best Parameters

AUC Training	
before SMOTE	**after SMOTE**
68.6%	85.1%

modeling with the best neural network parameters on data after SMOTE preprocessing has a higher value than data before SM0TE, namely 85.1%. Meanwhile modeling with data before SMOTE produced an AUC that had not yet reached the acceptable category, namely 68.6%. The AUC value for each stage class of breast cancer patients in training data modeling with the best neural network parameters is shown in Table 6. Each stage class had a higher AUC in the data after SMOTE. In data before SMOTE, the AUC value in stages II to IV had not reached the acceptable category (<70%). Table 4.6 also shows that the stage IV class has the

highest AUC value compared to other classes, namely 98.5% in the data after SMOTE. Then followed by stage I with a value of 84.6%. Stages II and III have AUC values of 77.9% and 76.3%, respectively. The best parameters are also modeled on testing data and evaluated using AUC values. Table 7 shows the AUC on data before and after SMOTE. Based

Table 6. AUC Neural Network Model for Each Stadium Category in Training Data Modeling

Class	AUC Training	
	before SMOTE	after SMOTE
Stage I	71.1%	84.6%
Stage II	68.2%	77.9%
Stage III	68.9%	76.3%
Stage IV	43.2%	98.5%

Table 7. AUC Neural Network Model on Testing Data Before and After SMOTE with the Best Parameters

AUC Testing	
before SMOTE	after SMOTE
49.9%	82.3%

on Table 7, the AUC in testing data modeling with the best neural network parameters on data after SMOTE preprocessing has a higher value than data before SM0TE, namely 82.3%. Meanwhile modeling with data before SMOTE produces an AUC of 49.9%. The AUC value in the data before SMOTE is <70%, which indicates that it has not reached the acceptable category.

3.2 Classification Using *k*-NN

k-Nearest Neighbor has a parameter k which indicates the number of nearest neighbors. In this research, the best *k* value will be selected using experimental k values of 2, 3, 5, and 10 with Euclidean distance and uniform weighting. Evaluation was carried out using stratified 10-fold cross validation by paying attention to the average AUC on the validation set. Based on Table 8, the fit of the model in the data before SMOTE is

Table 8. Average AUC of the k-NN Model on Training Data with Cross Validation Before and After SMOTE

k-NN	before SMOTE		after SMOTE	
	Training Set	Validation Set	Training Set	Validation Set
3-NN	86.3%	67.8%	94.6%	80.6%
5-NN	81.0%	55.5%	91.7%	82.1%
6-NN	79.1%	60.1%	90.7%	82.4%
10-NN	71.9%	55.8%	87.6%	80.2%

not good enough because it does not reach 70% or has not reached the acceptable category. Apart from that, overfitting also occurred which was indicated by the average AUC value which was quite far apart between the training set and validation set at all *k* values tested. The highest AUC value in the data before SMOTE is at $k = 3$, where the average AUC for the training set is 86.3% while for the testing set it is 67.8%. After SMOTE was carried out, the average AUC value was higher and not much different between the training set and testing set. The largest average AUC value on the validation set is at $k = 6$, namely 82.4% with an average AUC value on the training set of 90.7%. In the data before SMOTE, $k = 3$ is used, while in the data after SMOTE, $k = 6$ is used. Table 10 shows that the AUC for training data modeling with the best parameters on data after SMOTE preprocessing has a higher

value than data before SM0TE, namely 91.1%. Meanwhile modeling with data before SMOTE produces an AUC of 84.6%. The AUC value for each stage class of breast cancer patients in training data modeling with the best parameters is shown in Table 10. Each stage class had a higher AUC in the data after SMOTE. Table 10 also shows

Table 9. AUC of *k*-NN Model on Training Data Before and After SMOTE with the Best Parameters

AUC Training	
before SMOTE	**after SMOTE**
84.6%	91.1%

Table 10. AUC of *k*-NN Model for Each Stadium Category in Training Data Modeling

Class	**AUC Training**	
	before SMOTE	**after SMOTE**
Stage I	91.2%	91.4%
Stage II	84.0%	88.2%
Stage III	79.2%	84.5%
Stage IV	98.6%	99.2%

that the stage IV class has the highest AUC value compared to other classes, namely 99.2% in the data after SMOTE and 98.6% in the data before SMOTE. Then followed by stage II with the AUC between the data before SMOTE and after SMOTE not being much different, namely only a difference of 0.2%. The lowest AUC value was obtained in the stage III class, namely 79.2% in data before SMOTE and 84.5% in data after SMOTE. The best parameters are also modeled on testing data and evaluated using AUC values. Table 11 shows the AUC on data before and after SMOTE. According to Table 11, the AUC in testing data modeling with the best parameters

Table 11. AUC of the k-NN Model on Testing Data Before and After SMOTE with the Best Parameters

AUC Testing	
before SMOTE	**after SMOTE**
53.3%	80.8%

on data after SMOTE preprocessing has a higher value than data before SM0TE, namely 80.8%. Meanwhile modeling with data before SMOTE produces an AUC of 53.3%. The AUC value in the data before SMOTE is <70%, which indicates that it has not reached the acceptable category. Modeling for breast cancer stage classification at Surabaya Oncology Hospital has been carried out using k-NN and neural networks with the selection of the best parameters. The next step is to compare the AUC values for the two models to obtain the best method. Based on Table 12, it can be seen that the two models on the data before SMOTE have AUCs that tend to

Table 12. Comparison of AUC of k-NN Models and Neural Networks Before and After SMOTE

Model	**before SMOTE**		**after SMOTE**	
	Training	**Testing**	**Training**	**Testing**
k-NN	84.6%	53.3%	91.1%	80.8%
Neural Network	68.6%	49.9%	85.1%	82.3%

be overfitting and cannot be categorized as acceptable because the AUC is <70%. Meanwhile, in the data after SMOTE, the neural network model had a higher AUC on the testing data, namely 82.3%. In the testing data, the k-NN model has an AUC of 80.8%. However, the k-NN model training had a higher AUC value, namely 91.1%, while the neural

network was 85.1%. The criteria for selecting the best model are determined by testing data. Apart from that, in the k-NN model there is quite a significant difference between the AUC of training and testing. So, it can be said that neural networks are better at classifying breast cancer stages in Surabaya Oncology Hospital patients.

4 Conclusion

The number of breast cancer patients at the Surabaya Oncology Hospital in January 2011 – December 2019 was 294 patients, of which 19% were diagnosed with stage I, 44% were diagnosed with stage II, 35% were diagnosed with stage III, and another 2% were diagnosed with stage IV. After pre-processing using the SMOTE method, 475 data were obtained with a percentage of 24% for stage I, 27% for stage II, 22% for stage III, and 27% for stage IV. Furthermore, it is known that the optimal number of nearest neighbors (k) for classifying the breast cancer stage of Surabaya Oncology Hospital patients with k-NN is 3 in the data before SMOTE and 6 in the data after SMOTE. The optimal number of neurons for classifying breast cancer stages in Surabaya Oncology Hospital patients using a neural network is 5 neurons both in data before and after SMOTE. Data classification before SMOTE with neural network or k-NN occurred overfitting and the AUC value did not reach the acceptable category (<70%). The results of the comparison between the neural network and k-NN models showed that the neural network model after SMOTE preprocessing was better at classifying the stage of breast cancer patients at the Surabaya Oncology Hospital compared to the k-NN model with an AUC value on testing data of 82.3%, while k- NN has an AUC testing of 80.8%. For further research, other oversampling methods can be used to overcome imbalanced data such as SMOTE-NC, ADASYN, etc. Meanwhile, other methods such as feature selection can be used to overcome overfitting. Apart from that, selecting the best parameters from a neural network can be combined with a genetic algorithm so that classification accuracy becomes more optimal.

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under the project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2023.

References

1. Pourzand, A., Fakhree, M. B., Hashemzadeh, S., Halimi, M., & Daryani, A. (2011). Hormone Receptor Status in Breast Cancer and its Relation to Age and Other Prognostic Factors. *Breast Cancer: Basic and Clinical Research*, 87-92.
2. Verkooijen, H. M., Chappuis, P. O., Rapiti, E., Vlastos, G., Fioretta, G., Sarp, S., Sappino, A.P., Schubert, H., and Bouchardy, C. (2006). Impact of Familial Risk Factors on Management and Survival of Early-Onset Breast Cancer: a Population-Based Study. *British Journal of Cancer*, 231-238.
3. Houssami, N., Ciatto, S., Martinelli, F., Bonardi, R., & Duffy, S. W. (2009). Early Detection of Second Breast Cancers Improves Prognosis in Breast Cancer Survivors. *Annals of Oncology*, 1505-1510.
4. Muhartono, Ramanisa, S., Mutiara, H., & Riduan, R. J. (2016). Hubungan Antara Status Reseptor Estrogen, Reseptor Progesteron dan Human Epidermal Growth Factor Receptor 2 dengan Derajat Keganasan Karsinoma Payudara Invasif. *Majalah Kedokteran Andalas*, 65-72.

5. Dunnwald, L.K., Rossing, M.A., and Li, C.I. (2007). Hormone Receptor Status, Tumor Characteristics, and Prognosis: a Prospective Cohort of Breast Cancer Patients. *Breast Cancer Res.* 9(1), 1-10.
6. Seshadri, R., Firgaira, F.A., Horsfall, D.J., McCaul, K., Setlur, V., and Kitchen, P. (1993). Clinical Significance of HER-2/Neu Oncogene Amplification in Primary Breast Cancer. The South Australian Breast Cancer Group. *J. Clin. Oncol.* 11(10), 1936-42.
7. Kurniawan, M. F., & Ivandri. (2017). Komparasi Algoritma Data Mining untuk Klasifikasi Penyakit Kanker Payudara. *IC-Tech*, 1-8.
8. Chawla, N.V., Lazarevic, A., Hall, L.O., and Bowyer, K.W. (2003). SMOTEBoost: Improving Prediction of The Minority Class in Boosting. *European Conference on Principles of Data Mining and Knowledge Discovery*, 107-119.
9. Fausett, L. (1994). *Fundamentals of Neural Netwroks Architectures, Algorithms, and Applications.* London: Prentice Hall, Inc.
10. Hamamoto, Y., Uchimura, S., Tomita, S. (1997). A Bootstrap Technique for Nearest Neighbor Clasifier Design. *IEEE* Transactions On Pattern Analysis And Machine Intelligence, vol. 19, no. 1, pp. 73-79.
11. Alpaydin, E. (1997). Voting over multiple condensed nearest neighbors, *Artificial Intelligence Review*, 11, pp. 115-132.
12. Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan MATLAB.* Yogyakarta: Andi Yogyakarta.
13. Lopes, N., & Ribeiro, B. (2015). On the Impact of Distance Metrics in Instance-Based Learning Algorithms. *Iberian Conference on Pattern Recognition and Image Analysis* (pp. 48-56). Springer.
14. Bobrowski, L., & Topczewska, M. (2004). Improving the K-NN Classification with the Euclidean Distance Through Linear Data Transformations. *Industrial Conference on Data Mining* (pp. 23-32). Springer.
15. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques (3rd ed).* USA: Elsevier.
16. Erke, A. R., & Pattynama, P. M. (1998). Receiver Operating Characteristic (ROC) Analysis: Basic Principles and Aplications in Radiology. *European Journal of Radiology*, 88-94.
17. Chou, S., Shan, J., Guo, Y., & Zhang, L. (2010). Automated Breast Cancer Detection and Classification Using Ultrasound Image: A Survey. *Pattern Recognition*, 299-317.
18. Zweig, M. H., & Campbell, G. (1993). Receiver Operating Characteristic (ROC) Plots : A Fundamental Evaluation Clinical Medicine. *Clinical Chemistry*, 561-577.
19. Bekkar, M., Djemaa, H., & Alitouche, T. (2013). Evaluation Measures for Models Assesment Over mbalanced Data Sets. *Journal of Information Engginering and Application*, 3(10), 1-13.