Ovarian cancer classification using K-Nearest Neighbor and Support Vector Machine

V V P Wibowo¹, Z Rustam¹, S Hartini¹, F Maulidina¹, I Wirasati¹ and W Sadewo²

¹ Department of Mathematics, University of Indonesia, Depok 16424, Indonesia
² Department of Neurosurgery, Faculty of Medicine, University of Indonesia, Dr. Cipto Mangunkusumo National General Hospital, Indonesia

E-mail: veleryvirgina@sci.ui.ac.id

Abstract. Ovarian cancer is one of the common malignancies in women and a known cause of death. This condition occurs when a tumor appears from the growth of abnormal cells in the ovary. It causes about 140,000 deaths out of 225,000 cases annually. Most women with ovarian cancer do not have distinctive signs and symptoms even at the late stage. Therefore, diagnosis at an early stage is necessary because it has a significant impact on the survival rate. Machine learning with various methods can be used in the medical field to classify diseases. Among the many methods, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) were used and analyzed in this study to classify ovarian cancer. The data used were from Al Islam Bandung Hospital consisting of 203 instances with 130 labeled ovarian cancer and 73 as non-ovarian. The results showed that the KNN produced higher results than SVM with 90.47% of accuracy and 94.11% of F1-score, while SVM produced accuracy and F1-score values of 90.47% and 92.30% respectively.

1. Introduction
Cancer is a deadly disease and a major cause of global mortality among women in middle and high income countries. The number of women population in the world is about 49.5% [1]. In fact, ovarian cancer is one of their most common malignancies and the deadliest of all gynecologic cancers [2]. The approximated annual occurrence is 225,000 cases with 140,000 deaths worldwide [3]. In addition, over 70% of this condition have an underlying illness at the time of diagnosis, which is one of the causes of the high mortality rate [4].

Ovarian cancer occurs when cells grow abnormally in the ovary and become a tumor. Meanwhile, the ovary is a reproductive gland in women that produces egg cell or ovum. This cell is extruded into the fallopian tubes where it is fertilized. After fertilization, the ovum is propelled towards the uterus and develops into a fetus. In addition, the ovum produces progesterone and estrogen hormone [5].

The risk factors of this condition are genetic, nulliparous, interference of the endocrine system such as hormone replacement therapy after menopause, and lifestyle such as smoking and diet [6]. According to American Cancer Society (2015), other risk factors are obesity, fertility treatment, history of breast cancer, and more. In addition, most women with this condition are those who are above 63 years old. Therefore, the older women are more at risk of developing the disease [5].
The symptoms of ovarian cancer are non-specific such as weight loss, indigestion, changing bowel habit, feeling full early, pelvic and abdominal pain, bloating, and massive swelling of the abdomen. However, most of the patients do not show those symptoms even at a late stage. Therefore, early diagnosis is difficult [7,8].

The condition can be diagnosed through a blood test to detect the cancer antigen 125 protein. This tumor marker is a glycoprotein with high molecular weight which is increased at an advanced stage. The antigen level is a standard component of patients with advanced ovarian cancer and has an inverse relationship with survival rate. Also, an increased level represents tumor relapse and low survival rate, while a decreased level means a positive condition [9]. There are no reliable screening methods for diagnosis until this time [5].

More than 70% of women with ovarian cancer are diagnosed at an advanced stage. Furthermore, less than 40% have a survival rate of 5 years [10]. Since timely detection is related to increased survival rate, early diagnosis is very important because it has a significant impact on its fatality level [11].

In the medical field, machine learning can be implemented for analyzing patient data, and also for diagnosing diseases. The machine takes the important features of the patient as input and produces an accurate diagnosis as output [12]. Previous studies have implemented machine learning in the classification of ovarian cancer with various methods such as the Logistic Regression method [13], IOTA Logistic Regression Model LR2 [14], and Bayesian Logistic Regression [5]. The results are 93% of sensitivity and 76% of specificity for Logistic Regression, while 97% of sensitivity and 69% of specificity for IOTA Logistic Regression Model LR2. Meanwhile, Bayesian Logistic Regression produced accuracy, precision, recall, and F1-score values of 77.33%, 78.76%, 88.46%, and 83.33% respectively.

There are many supervised learning methods of machine learning that are used in classification. Among those methods, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are frequently used due to their good performance. Several studies implemented SVM in classification problems such as schizophrenia with an accuracy of 90.1% [15], acute sinusitis with an accuracy of 90% [16], breast cancer and lung cancer with an accuracy of 96.4286% and 99.9999% respectively [17]. SVM also applied for face recognition with the implementation of the kernels, which are Polynomial Kernel and Radial Basis Function (RBF) Kernel. The results showed that SVM with RBF kernel produced higher accuracies [18]. Meanwhile, The KNN method has applied in the classification of thalassemia with an accuracy of 97.14% [19]. Therefore, this study proposed KNN and SVM with RBF kernel methods to classify ovarian cancer and find which one has better results.

2. Materials and methods

2.1. Dataset

This study used dataset from Al Islam Bandung Hospital, West Java. It consisted of 5 attributes and 203 instances, which are 130 labeled ovarian cancer and 73 non-ovarian. The five attributes are CA125 (U/mL), Hemoglobin (g/dL), Leukocytes (cells/µL), Hematocrit (%), and Platelets (cells/µL).

The normal value of CA125 is lower than 35 U/mL [9]. Meanwhile, the value ranges from 13-18 g/dL for Hemoglobin, 4000-10000 cells/µL for Leukocytes, 40-54% for Hematocrit, and 150000-450000 cells/µL for Platelets [20]. The dataset is shown in Table 1 below.

| CA125 | Hemoglobin | Leukocytes | Hematocrit | Platelets | Outcome |
|-------|------------|------------|------------|-----------|---------|
| 445   | 10         | 26300      | 31         | 254000    | 1       |
| 62    | 11         | 7600       | 39         | 298000    | 1       |
| 40    | 14         | 6900       | 42         | 437000    | 1       |
| 8     | 12         | 11300      | 36         | 222000    | 0       |
Table 1. Ovarian cancer dataset.

|    |    |    |    |    |    |
|----|----|----|----|----|----|
| 69 | 12 | 10400 | 35 | 239000 | 1  |
| 14 | 10 | 8300  | 37 | 536000 | 0  |
| 25 | 11 | 5500  | 29 | 176000 | 0  |
| 24 | 13 | 9000  | 38 | 218000 | 0  |
| 102| 13 | 7100  | 40 | 382000 | 1  |
| 76 | 9  | 8000  | 29 | 547000 | 1  |

There is outcome or a target attribute in the dataset, which means the diagnosis or label, 1 refers to instance with ovarian cancer and 0 refers to normal condition.

2.2. Supervised learning
This study used supervised learning as method for the classification of ovarian cancer data. It involves learning the relationship between attributes in the training dataset and using the result to build a model that predicts unlabeled data [21].

2.3. K-Nearest Neighbor
K-Nearest Neighbor (KNN) is commonly used in many studies, especially when there are only few explanation of the data distribution [22]. This algorithm is an instance-based classifier with basic and simple principle which can be extensively implemented. The idea of KNN is to determine the class label of new samples based on the instances. Therefore, this method classifies new sample based on its similarity to all samples in the labeled instance [23,24].

The steps below are based on KNN algorithm:
1. Set the number of nearest neighbors or k as the parameter.
2. Calculate the distance of a new sample to all samples in the training data.
3. Select k samples or k neighbors in the training data which have the nearest distance to a new sample.
4. Use the most votes of class label from k nearest neighbors as the label of a new sample.

In Figure 1, there are three class labels represented by various colors which are $\omega_1, \omega_2$ and $\omega_3$. Furthermore, there is a sample $x_u$ with five nearest neighbors ($k = 5$), which is classified by KNN as part of $\omega_1$. This is due to the majority class from the nearest neighbors of $x_u$ is $\omega_1$. Therefore, $\omega_1$ is a label of sample $x_u$.

Figure 1. The illustration of KNN [25].
The class label of the samples is determined based on their similarities, and the measured distance is needed for the calculation. Also, the samples will have less similarity as long as the distance between them is large [23]. Therefore, this study implemented the Euclidean Distance as the measurement method to calculate the nearest distance between the samples. This can be obtained by the formula below [26]:

\[ D(x, y) = \left( \sum_{k=1}^{n} (x_k - y_k)^2 \right)^{1/2} \]  \hspace{1cm} (1)

2.4. Support Vector Machine

In 1995, Support Vector Machine (SVM) was first proposed by Copnes and Vapnik. The idea of SVM in a non-linear problem is to map the input vectors or samples into a higher dimensional space and to build an optimal hyperplane for splitting data using the kernel function [16,23]. In order to obtain this, SVM works by measuring the maximum margin or the distance between the nearest vectors and the hyperplane [27].

Given \( \{x_i, y_i\}_{i=1}^{M} \) as the dataset, \( x_i \in R^n \) are the input vectors and \( y_i \{ -1, +1 \} \) are the class labels. Therefore, the idea of SVM is to create an optimal hyperplane that separates the two classes defined as:

\[ w^T \cdot x + b = 0 \]  \hspace{1cm} (2)

where \( w \) is the weight or orthogonal vector, \( b \) is the bias or distance between origin and the hyperplane, and \( x \) is the feature vector or sample [28]. Figure 2 shows the illustration of SVM with two classes, positive and negative.

![Figure 2. The illustration of SVM [29].](image)

Each vector \( x_i \) in the positive and negative class of the dataset should satisfy the following conditions:

\[ w^T \cdot x_i + b \geq 1 \text{ for } y_i = +1 \]  \hspace{1cm} (3)

\[ w^T \cdot x_i + b \leq -1 \text{ for } y_i = -1 \]  \hspace{1cm} (4)

which are equivalent to the inequality written as follows:
\[ y_i(w^T \cdot x_i + b) \geq 1 \text{ for } i = 1, 2, \ldots, M \] (5)

The optimization problem of SVM is written as follows:

\[
\min \frac{1}{2} \|w\|^2 \\
\text{s.t } y_i(w^T \cdot x_i + b) \geq 1 \text{ for } i = 1, 2, \ldots, M
\] (6)

To solve the optimization problem, Lagrange function is applied [28]. Then, the optimization problem becomes:

\[
\min Lp = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{M} a_i y_i (w^T \cdot x_i + b) + \sum_{i=1}^{M} a_i \\
\text{s.t } y_i(w^T \cdot x_i + b) \geq 1 \text{ for } i = 1, 2, \ldots, M
\] (8)

for \( a_i \geq 0 \) where \( a_i \) is Lagrange multiplier. Then, the decision function of SVM is:

\[
f(x_i) = \sum_{i=1}^{M} a_i y_i < x_i, x_i > + b
\] (9)

where \( x_i \) is a testing data.

For several error or non-separable cases, the slack variable \( \xi_i \) and parameter \( C \) (\( C > 0 \)) are added to the optimization problem [28]. Therefore, the optimization problem is:

\[
\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{M} \xi_i \\
\text{s.t } y_i(w^T \cdot x_i + b) \geq 1 - \xi_i \text{ for } i = 1, 2, \ldots, M
\] (10)

In some cases where the classes in the dataset are not linearly separable, the kernel function can be applied [28]. The formula of this function is written as follows:

\[
K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)
\] (12)

\( \varphi(x) \) is a function that is used to map input vectors \( x \) into a higher dimension where the classes are linearly separated [28]. The kernel function used in this study was Radial Basis Function (RBF) with \( \gamma \) as its parameter. The RBF kernel formula is shown as follows:

\[
K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0
\] (13)

2.5. Confusion matrix

In classification, the confusion matrix is used to measure the performance of machine learning methods. Also, it shows the comparison between the outcomes of the instance that were classified by the model and the actual results [30]. Table 2 shows the confusion matrix.

| Predicted | Actual |
|-----------|--------|
| Positive  | TP     | FP |
| Negative  | FN     | TN |
Based on the Table 2 above, the model performance can be evaluated by calculating the accuracy, recall, precision, and F1-score of the model. The formulas of the calculations are shown in the equations below:

\[
\text{accuracy} = \frac{TP + TN}{TP + FN + TN + FP}
\]  
(14)

\[
\text{recall} = \frac{TP}{FN + TP}
\]  
(15)

\[
\text{precision} = \frac{TP}{FP + TP}
\]  
(16)

\[
\text{F1-score} = \frac{\text{recall} \times \text{precision} \times 2}{\text{recall} + \text{precision}}
\]  
(17)

where TP (True Positive) represents the number of instances having ovarian cancer that were correctly predicted. Meanwhile, FN (False Negative) represents the number of instances having ovarian cancer that were incorrectly predicted. TN (True Negative) represents the number of non-ovarian cancer instances that were correctly predicted. Also, FP (False Positive) represents the non-ovarian cancer instances that were incorrectly predicted.

3. Results and discussion
This study implemented the KNN and SVM methods in the classification of ovarian cancer with different percentages of training data which range from 10% to 90%.

The optimal parameter used for KNN in this study was \( k = 7 \) which was optimized using the Grid Search method. The results of classification using KNN are listed in the table below.

| Training Data (%) | Accuracy (%) | Recall (%) | Precision (%) | F1-score (%) |
|-------------------|--------------|------------|---------------|--------------|
| 10                | 73.22        | 86.44      | 75.55         | 80.62        |
| 20                | 75.46        | 86.79      | 77.96         | 82.13        |
| 30                | 75.52        | 85.26      | 79.41         | 82.23        |
| 40                | 76.22        | 84.11      | 81.17         | 82.61        |
| 50                | 78.43        | 86.15      | 81.15         | 83.57        |
| 60                | 81.70        | 87.50      | 85.96         | 86.73        |
| 70                | 83.60        | 90.00      | 85.71         | 87.80        |
| 80                | 85.36        | 89.65      | 89.65         | 89.65        |
| 90                | 90.47        | 94.11      | 94.11         | 94.11        |

Based on Table 3, KNN method performed best using 90% of the training data with an accuracy value of 90.47%. The highest recall, precision, and F1-score value was 94.11% which was also obtained from 90% of the training data. However, the lowest results on accuracy was 73.22% with recall, precision, and F1-score values were 86.44%, 75.55%, and 80.62% respectively. The recall has higher values than precision values for most training data. This condition means that the number of negative instances that incorrectly predicted was higher compared to the number of positive instances that incorrectly predicted.
In the classification using SVM, the RBF kernel was applied. Also, the parameters were $C = 1000$ and $\gamma = 0.1$, which were optimized using the Grid Search method. The following table is the results of classification using SVM with RBF kernel.

| Training Data (%) | Accuracy (%) | Recall (%) | Precision (%) | F1-score (%) |
|-------------------|--------------|------------|---------------|--------------|
| 10                | 63.38        | 51.66      | 87.32         | 65.05        |
| 20                | 68.71        | 65.09      | 83.13         | 73.01        |
| 30                | 75.52        | 75.53      | 85.54         | 80.22        |
| 40                | 86.88        | 85.18      | 94.52         | 89.60        |
| 50                | 88.23        | 86.15      | 94.91         | 90.31        |
| 60                | 90.24        | 90.38      | 94.00         | 92.15        |
| 70                | 90.16        | 87.80      | 97.29         | 92.30        |
| 80                | 90.24        | 87.50      | 95.45         | 91.30        |
| 90                | 90.47        | 92.30      | 92.30         | 92.30        |

From Table 4 above, training data was used which differ from 10% to 90% for SVM with RBF kernel. Meanwhile, experiments using 90% of the training data produced the highest accuracy value of 90.47%. The value for recall, precision, and F1-score was 92.30%. Furthermore, the lowest accuracy value was 63.38% using 10% of the training data with recall, precision, and F1-score values of 51.66%, 87.32%, and 65.05% respectively. The precision values for most training data were higher than the recall values. This condition means that the number of positive instances that incorrectly predicted was higher compared to the number of negative instances that incorrectly predicted.

It observed that both KNN and SVM methods produced the best accuracy value of 90.47% using 90% of training data. However, we need to evaluate the results of other performance measures such as recall, precision, and F1-score. The high recall value and the low precision value indicate that the number of negative instances that incorrectly predicted was high. This condition can cause a disadvantage because healthy women have to pay unnecessary medical expenses. Meanwhile, the high precision value and low recall value indicate that the number of positive instances that incorrectly predicted was high. This condition is fatal as it can cause an early death for the women that have ovarian cancer because it is not treated immediately. Therefore, F1-score is needed to evaluate the performance of the methods. Also, a high value of F1-score shows a high performance of the classification [31]. The F1-score value for KNN and SVM was 94.11% and 92.30% respectively. Hence, KNN is a better method than SVM to classify ovarian cancer.

4. Conclusion
This study applied KNN and SVM with RBF kernel as supervised machine learning methods and compared them for classifying the data of ovarian cancer. The dataset from Al Islam Bandung Hospital used in this study consisted of 5 attributes and 203 instances. There were 130 instances labeled as ovarian cancer and 73 non-ovarian.

The KNN idea was to determine a class label of new sample based on the majority class. This was achieved by measuring the closest distance. Meanwhile, the idea of SVM method was to obtain an optimal hyperplane that separates the data into classes. This was achieved by measuring the maximum margin between the closest vectors and the hyperplane.

Based on Table 3 and 4, the experiments using 90% of the training data produced the best accuracy value of 90.47% for both KNN and SVM. Meanwhile, the F1-score value was 94.11% for KNN and
92.30% for SVM with RBF kernel. Therefore, the results showed that KNN is a better method than SVM with RBF kernel to classify ovarian cancer. This study is hoped to be a reference for doctors when making a diagnosis. For future research, a larger dataset is recommended to get better results.

Acknowledgement
This research is fully supported financially by The Indonesian Ministry of Research, Technology, with a KEMENRISTEK/BRIM PDUPT 2021 research grant scheme.

References
[1] Torre L A, Islami F, Siegel R L, Ward E M and Jemal A 2017 Global Cancer in Women: Burden and Trends Cancer Biomar 26 444–57
[2] Momenimovahed Z, Tiznobaik A, Taheri S and Salehiniya H 2019 Ovarian cancer in the world: epidemiology and risk factors Int J Womens Health 11 287–99
[3] Ferlay J, Shin H R, Bray F, Forman D, Mathers C and Parkin D M 2010 Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008 Int J Cancer 127 2893–917
[4] Rauh-Hain J A, Krijn T C, Del Carmen M G and Olawaiye A B 2011 Ovarian cancer screening and early detection in the general population Rev Obstet Gynecol 4 15–21
[5] Octaviani T L, Rustam Z and Siswantining T 2019 Ovarian Cancer Classification using Bayesian Logistic Regression IOP Conf. Ser.: Mater. Sci. Eng. 546 052049
[6] Salehi F, Dunfield L, Phillips K P, Krewski D and Vanderhyden B C 2008 Risk factors for ovarian cancer: an overview with emphasis on hormonal factors J Toxicol Environ Health B Crit Rev 11 301–21
[7] Burges A and Schmalfeldt B 2011 Ovarian cancer: diagnosis and treatment Dtsch Arztebl Int 108 635–41
[8] Gajjar K, Ogden G, Mujahid M I and Razvi K 2012 Symptoms and risk factors of ovarian cancer: a survey in primary care ISRN Obstet Gynecol 2012 754197
[9] Gupta D and Lis C G 2009 Role of CA125 in predicting ovarian cancer survival - a review of the epidemiological literature J Ovarian Res 2 13
[10] Huang Z, Yan H, Chavan D, Yuan Z, Yang X, Zhang Y, Song K and Kong B 2018 Effective treatment of a patient with stage IV ovarian cancer: A case report Oncol Lett 15 588–91
[11] Das P M and Bast R C Jr 2008 Early detection of ovarian cancer Biomark Med 2 291–303
[12] Raval D, Bhatt D N, Kumhar M K, Parikh V and Vyas D A 2016 Medical Diagnosis System Using Machine Learning International Journal of Computer Science & Communication (IJCSC) 7 177–82
[13] Timmerman D, Testa A C, Bourne T, Ferrazzi E, Ameye L, Konstantinovic M L, Van Calster B, Collins W P, Vergote I, Van Huffel S and Valentin L 2005 Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group J Clin Oncol 23 8794–801
[14] Nunes N, Yazbek J, Ambler G, Hoo W, Naftalin J and Jurkovic D 2012 Prospective evaluation of the IOTA logistic regression model LR2 for the diagnosis of ovarian cancer Ultrasound Obstet Gynecol 40 355–9
[15] Rampisela T V and Rustam Z 2018 Classification of Schizophrenia Data Using Support Vector Machine (SVM) J. Phys.: Conf. Ser. 1108 012044
[16] Arfiani, Rustam Z, Pandelaki J and Siahaan A 2019 Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification IOP Conf. Ser.: Mater. Sci. Eng. 546 052011
[17] Nadira T and Rustam Z 2018 Classification of cancer data using support vector machines with features selection method based on global artificial bee colony AIP Conf. Proc. 2023 020205
[18] Rustam Z and Faradina R 2018 Face recognition to identify look-alike faces using support vector machine J. Phys.: Conf. Ser. 1108 012071
[19] Siswantining T, Kamalia A, Rustam Z, Subroto F and Semendawai A S 2019 Classification of thalassemia data using K-nearest neighbor and Naïve Bayes *Int. J. Adv. Sci. Technol.* 28 15–9

[20] Huljanah M, Rustam Z, Utama S and Siswantining T 2019 Feature Selection using Random Forest Classifier for Predicting Prostate Cancer *IOP Conf. Ser.: Mater. Sci. Eng.* 546 052031

[21] Fabris F, Magalhães J P and Freitas A A 2017 A review of supervised machine learning applied to ageing research *Biogerontology* 18 171–88

[22] Peterson L E 2009 K-nearest neighbor *Scholarpedia* 4 1883

[23] Zhang Z, Li Y, Jin S, Zhang Z, Wang H, Qi L and Zhou R 2018 Modulation Signal Recognition Based on Information Entropy and Ensemble Learning *Entropy* 20 198

[24] Kataria A and Singh M D 2013 A Review of Data Classification Using K-Nearest Neighbour Algorithm *Int. J. Emerging Technol. Adv. Eng.* 3 354–60

[25] Zhang W 2017 Machine Learning Approaches to Predicting Company Bankruptcy *J. Risk Financial Manag.* 6 364–74

[26] Abu Alfeilat H A, Hassanat A B A, Lasassmeh O, Tarawneh A S, Alhasanat M B, Eyal Salman H S and Prasath V 2019 Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review *Big Data* 7 221–48

[27] Nithya B and Ilango V 2019 Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction *SN Appl. Sci.* 1 641

[28] Abukhait J, Mansour A M and Obeidat M 2018 Classification based on Gaussian-kernel Support Vector Machine with Adaptive Fuzzy Inference System *Prz Elektrotechnicz* 94

[29] Tharwat A 2019 Parameter investigation of support vector machine classifier with kernel functions *Knowl Inf Syst* 61 1269–302

[30] Aszhari F R, Rustam Z, Subroto F and Semendawai A S 2020 Classification of thalassemia data using random forest algorithm *J. Phys.: Conf. Ser.* 1490 012050

[31] Tharwat A 2018 Classification assessment methods *Applied Computing and Informatics*