Pancreatic Cancer Classification Using the Kernel-based Support Vector Machine (KSVM)

R Khairi¹, S G Fitri¹ and Z Rustam¹*

¹Department of Mathematics, University of Indonesia, 16424 Depok, Indonesia
*e-mail: rustam@ui.ac.id

Abstract. Pancreatic cancer is a disease in which malignant (cancerous) tumor cells develop in pancreatic tissue; organ behind the lower abdomen and in front of the spine, which helps the body use and store energy from food by producing hormones to control blood sugar levels and digestive enzymes to break down food. Usually, pancreatic cancer is rarely detected at an early stage. One sign of a person with pancreatic cancer is diabetes, especially if it coincides with rapid weight loss, jaundice, or pain in the upper abdomen that spreads to the back. Among various types of cancer, pancreatic cancer has the lowest survival rate of only about 3%-6% of those diagnosed who can survive for five years. If patients are diagnosed on time for treatment, their chances of survival will increase. There is a tumor marker commonly used to follow the course of pancreatic cancer, namely CA 19-9 which can be measured in the blood. Healthy people can have small amounts of CA 19-9 in their blood. High levels of CA 19-9 are often a sign of pancreatic cancer. But sometimes, high levels can indicate other types of cancer or certain noncancerous disorders, including cirrhosis and gallstones. Because a high level of CA 19-9 is not specific for pancreatic cancer, CA 19-9 cannot be used by itself for screening or diagnosis. It can help monitor the progress of your cancer and the effectiveness of cancer treatment. In this study, the Kernel-based Support Vector Machine method is used to classify CA19-9 blood test results into two sections including data on patients diagnosed with pancreatic cancer or normal patients. This method will get an accuracy of around 95%.

Keywords: Kernel-based, Vector Machine.

1. Introduction
Cancer can begin in any organ in the human body, which consists of trillions of cells. In all types of cancer, a person's cells begin to divide endlessly and spread to the surrounding tissue so that they can form tumors. Cancer is a genetic disease that causes cancer can be inherited from our parents. Some specific environmental exposures such as substances and chemicals in tobacco smoke, ultraviolet radiation can also cause cancer. There are so many types of cancer. Usually, this type of cancer is named based on the tissue or organ where cancer grows [1]. For example, pancreatic cancer is a type of cancer that starts in the pancreas [2].

The pancreas is an organ located behind the upper left abdomen. The pancreas is surrounded by other organs such as the small intestine, liver, and spleen. The pancreas changes the food that enters the body into fuel for body cells. The pancreas has two main functions: the exocrine function which aids digestion and the endocrine function which regulates blood sugar. There are several diseases of the pancreas...
including pancreatitis, precancerous conditions, and pancreatic cancer. Each of these diseases can show different symptoms and require different treatments [3].

According to the National Cancer Institute (NCI), pancreatic cancer ranks 11th as the most common cancer in the United States. About 97.5% of all new pancreatic cancer diagnoses are people aged 45-55 years, 89% of all new diagnoses are at people aged 55-66 years, 66.5% of all new cases are in people older than 65. Men have a slightly higher risk of developing pancreatic cancer than women [4]. Not yet known what the exact cause of pancreatic cancer, but some factors can increase the risk of developing the disease, namely smoking, family history of pancreatic cancer or hereditary cancer syndrome, and chronic pancreatitis [3]. Usually, the symptoms of early-stage pancreatic cancer may not be seen. However, common symptoms of pancreatic cancer include jaundice, digestive problems, rapid weight loss, and stomach ache. In addition to the common symptoms of pancreatic cancer, there is some symptoms of advanced pancreatic cancer include: worsening of the upper abdomen or back pain, especially after eating or lying down, extreme fatigue, swelling, bedsores, newly diagnosed diabetes, and depression [4]. Some treatments for pancreatic cancer can be done depending on the type of cancer, the stage of cancer and the needs and personal preferences of the patient. General treatments include chemotherapy, gastroenterology, interventional radiology, radiation therapy, and surgery [3, 4].

One common evaluation for diagnosing pancreatic cancer is a laboratory test. We use laboratory tests to examine antigens associated with tumors, such as high levels of carbohydrate antigens (CA) 19-9 in the blood [4]. Although carbohydrate antigens (CA) 19-9 are known as biomarkers of pancreatic cancer, they are not commonly used for general screening, because of their low sensitivity and specificity [5, 6, 7, 8]. CA 19-9 level in most healthy adults should be less than 37 U/mL [9, 10]. High CA 19-9 level (more than 37 U/mL) were found in pancreatic carcinoma patients in clinical observation [6, 9]. Test results can vary depending on age, gender, medical history, the method used for the test, and other things. Apart from being part of the diagnosis of pancreatic cancer along with other tests, the CA 19-9 blood test is also needed when patients undergo cancer treatment to see how well the treatment is working. Then a CA19-9 blood test is also needed after being treated for cancer to find out whether the patient's cancer has recurred [10].

In this paper, we use a data set of carbohydrate antigens (CA) 19-9 blood test results obtained from the laboratory of Al Islam Bandung Hospital in Bandung, West Java. This data set will be classified into two parts: patient data diagnosed with pancreatic cancer and normal patient data. This classification uses the Kernel-based Support Vector Machine which is one of the most powerful and accurate methods among well-known data mining algorithms. Initially, SVM was developed by Vapnik in the 1990s which is rooted in statistical learning theory. As time has progressed, SVM has been developed very quickly both in theory and practice [11]. SVM is a binary classification method that creates a model that can be generalized well [12, 13] with an optimal global solution [12, 14]. In SVM, the two target values are separated by hyper plane so that the closest distance between data and margin is maximized. In this study, SVM is used because it has the potential for better generalization and accuracy [12, 13].

2. Method

2.1. Kernel Function

In the classification problem, some training data sets may not be linearly separable. This problem can be solved with the kernel, which is a method for projecting each vector to a higher dimensional feature space, where all data can be presented linearly using a function called the kernel function [15, 16]. The kernel function for every \( x \in \mathbb{R}^n \) is given by [17, 18, 19]:

\[
K(x, y) = (\phi(x), \phi(y))
\]  

(1)

where \( \phi(x) \) is the feature space mapped by input data \( x \) [18].
The distance between $\emptyset(x)$ and $\emptyset(y)$ in the feature space can be defined as [18, 19]:

$$(d(x,y))^2 = K(x,x) - 2K(x,y) + K(y,y)$$  \hfill (2)

One of the commonly used kernel functions is a linear kernel function which is defined as follows [15, 18, 20]:

$$K(x,y) = (x \cdot y)$$  \hfill (3)

2.2. Support Vector Machine (SVM)

The SVM algorithm was created by Vapnik in the 1960s, where SVM is a category of supervised learning methods. Until now the SVM algorithm has been widely used and is known as one of the effective machine learning because it has a high classification efficiency. The basic concept of SVM is an attempt to find the optimal hyperplane that can be used to classify data into different data classes. The main purpose of the SVM algorithm is to find the optimal hyperplane that maximizes margins; the distance between the hyperplane itself and the nearest data for each class. Margin is directly proportional to the possibility of smaller errors in generalization [21]. The optimal hyperplane formed is illustrated in Figure 1.

![Figure 1](image)

**Figure 1.** Optimal Hyper Plane Formed from Maximum Margin.

Suppose there is some data as $N$. Then suppose $(x_i, y_i)$ as a data set with $i = 1, ..., N$, where $x_i \in R^n$ is input data and $y_i \in \{-1, 1\}$ is the class label (output) of the pancreatic dataset, namely the pancreatic cancer class, and normal class. Hyperplane to be formed can be written with the following equation [22]:

$$y(x) = w^T x + b$$  \hfill (4)

where $w$ is a vector of the weight parameter values, and $b$ is a bias that has a scalar value.

Hyperplane that is formed will separate the data into two classes in the dataset, namely the cancer class or normal, or the SVM method; class that is positive or negative. The process of separating this dataset is carried out with the following conditions [22]:

$$w^T x + b \geq 1, y_i = +1$$  \hfill (5)

$$w^T x + b \leq 1, y_i = -1$$  \hfill (6)

The above equations, in general, can be stated in the following statement [22]:

$$y_i (w^T x + b) \geq 1, i = 1, 2, ..., n$$  \hfill (7)

The distance between the two hyper planes can be defined with the equation below [22]:

...
\[ \frac{|w^T x_i + b|}{\|w\|} = \frac{1}{\|w\|} \]  

(8)

The resulting total distance between the two hyperplanes is \( \frac{2}{\|w\|} \). To maximize margins, \( \|w\| \) is minimized by [22]:

\[ \min \frac{1}{2} \|w\|^2 \]  

with the provision of:

\[ y_i(w^T x + b) \geq 1, \forall i = 1, 2, ..., N \]  

(10)

In the formation of the hyperplane, if there is training data that cannot be separated linearly, then the problem can be overcome by adding the slack variable \( (\varepsilon_i \geq 0) \). This case is called the misclassification error. Adding the slack variable changes the formula which can be written as follows [22]:

\[ \min \frac{1}{2} \|w\|^2 + C \sum \varepsilon_i \]  

(11)

with the provision of:

\[ y_i(w^T x_i + b) \geq 1 - \varepsilon_i \]  

(12)

If \( \varepsilon > 1 \), there will be a misclassification at that time. There is a parameter to adjust the balance in maximizing margins and minimizing misclassification, which is parameter C. Parameter C is used to avoid overfitting. [22].

2.3. Confusion Matrix

The confusion matrix, also known as the error matrix, is a visual evaluation tool for the performance of classification algorithms used in machine learning. The confusion matrix is a four-cell contingency table, where each row of the matrix represents the prediction class and each column of the matrix represents the actual class, or vice versa [23].

Table 1. Confusion Matrix.

| Predicted Condition (Predicted Class) | True Condition (Actual Class) |
|--------------------------------------|-------------------------------|
|                                      | Positive Condition | Negative Condition |
| Positive Condition                   | True Positive (TP) | False Positive (FP) |
| Negative Condition                   | False Negative (FN) | True Negative (TN) |

In this case, in table 1. There are two classes that show a patient diagnosed with pancreatic cancer and not diagnosed (normal). The number of data whose prediction results are classified into the same class as the actual state is labeled true (T). However, if the prediction results are classified into different classes from the actual state labeled false (F). Positive (P) and negative (N) labels indicate the model predictions that conflict with the actual values of the observations [24].

To find out the performance of the machine learning method used, there are several parameter calculations that we can get from the confusion matrix, including the following equation [24]:
\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\] (13)

\[
\text{Precision} = \frac{TP}{TP + FP}
\] (14)

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

\[
\text{F1-Score} = \frac{2TP}{2TP + FP + FN}
\] (16)

The accuracy value can be obtained from the confusion matrix quite easily because the number of True Positive (TP) and True Negative (TN) is divided by the total number of observations tested. In simpler terms, the proportion of observations is classified correctly. The precision value is the ratio of True Positive (TP) to the number of predicted positive cases. Precision is used to find out what percentage of patients predicted pancreatic cancer correctly from the overall prediction of patient data for pancreatic cancer. Recall reflects the ratio or ratio of True Positive (TP) to the actual number of positive cases. Recall is used to find out what percentage of patients who are predicted to have pancreatic cancer correctly from the overall data of patients who have pancreatic cancer in the actual condition. F1-Score is a harmonic mean value of precision and recall values. Used if false negative and false positive have far different values. The higher the F1-Score, the better the model [23].

3. Experiment

In this paper, we use a dataset of CA 19-9 blood tests in the laboratory of Al Islam Bandung Hospital, Bandung, West Java in Table 2.

| CA 19-9 (U/mL) | Hemoglobin (g/dL) | Leukocyte (cell/uL) | Hematocrit (%) | Platelets (cell/uL) | Diagnosis |
|---------------|-------------------|---------------------|----------------|---------------------|-----------|
| 114.8         | 13.6              | 11900               | 39.9           | 315000              | 1         |
| 1142          | 7.5               | 11800               | 23.6           | 547000              | 1         |
| 16.69         | 12.1              | 5100                | 36.4           | 247000              | 0         |
| 79.84         | 8.8               | 8300                | 28.6           | 300000              | 1         |
| 44.33         | 8.4               | 24600               | 25.2           | 662000              | 1         |
| 18.4          | 14.5              | 8100                | 43             | 212000              | 0         |
| 39.75         | 8.8               | 11900               | 27.9           | 645000              | 1         |
| ⋮             | ⋮                 | ⋮                   | ⋮              | ⋮                   | ⋮         |
| 51.37         | 11.5              | 7000                | 36.4           | 530000              | 1         |
| 6.27          | 9.3               | 12900               | 31.8           | 728000              | 0         |
| 86.21         | 10.1              | 12800               | 34.1           | 346000              | 1         |

The data set consists of 203 data consisting of 6 attributes. In the diagnosis section, 0 is defined for normal data, and 1 is defined for diagnosed pancreatic cancer data. Normal CA 19-9 is less than 37U/mL,
normal hemoglobin is 13-18g/dL, normal leukocytes are 4000-10000 cells/uL, normal hematocrit is 40-54%, and normal platelet count is 150000-450000 cells/uL.

The results of the classification of this cancer data set will later be limited to determine whether CA 19-9 blood test results are diagnosed with pancreatic cancer or not, this study did not arrive at the determination of the stage of pancreatic cancer.

4. Result and Discussion

The following chart illustrates the accuracy of the results of the classification of Pancreatic Cancer using a Kernel-based Support Vector Machine in Figure 2.

![Figure 2. Accuracy (%) Using Kernel-based Support Vector Machine.](image)

From the graph above it can be seen that the highest accuracy level of 95.23% is obtained when training data is used as much as 90% of the overall data using the parameters C=50 and C=100.

Then from the SVM method using a linear kernel and with the parameter C=100, a confusion matrix can be obtained for pancreatic cancer data in Table 3.

| Data Training (%) | C=10  | C=50  | C=100 |
|-------------------|-------|-------|-------|
| 10                | 62.29 | 60.10 | 68.30 |
| 20                | 65.64 | 68.09 | 70.55 |
| 30                | 68.53 | 76.22 | 81.11 |
| 40                | 77.86 | 79.50 | 88.11 |
| 50                | 78.43 | 84.31 | 89.34 |
| 60                | 81.70 | 82.92 | 87.25 |
| 70                | 75.40 | 91.80 | 87.80 |
| 80                | 75.60 | 90.24 | 91.80 |
| 90                | 85.71 | 95.23 | 95.12 |

**Table 3.** Confusion Matrix in SVM with Linear Kernel and 90% Training Data.

| Predicted Condition (Predicted Class) | True Condition (Actual Class) |   |
|--------------------------------------|--------------------------------|---|
|                                      | Positive                       | 38.10%  |
| Positive                             | Negative                       | 4.76%   |
| Negative                             |                                | 0%      |
|                                      | Negative                       | 57.14%  |
The confusion matrix in Table 3 shows that the Kernel-based SVM method estimates that there are as many as 38.10% of patients affected by pancreatic cancer, and there are as many as 57.14% of patients not affected by pancreatic cancer.

Then the following table represents other parameters that can be calculated; F1Score for SVM method using a linear kernel and with parameters C=10, C=50, and C=100 in table 4.

| Data Training (%) | F1-Score C=10 | F1-Score C=50 | F1-Score C=100 |
|-------------------|---------------|---------------|----------------|
| 10                | 0.60          | 0.46          | 0.66           |
| 20                | 0.65          | 0.67          | 0.70           |
| 30                | 0.67          | 0.75          | 0.88           |
| 40                | 0.78          | 0.80          | 0.90           |
| 50                | 0.79          | 0.85          | 0.87           |
| 60                | 0.82          | 0.83          | 0.88           |
| 70                | 0.76          | 0.92          | 0.92           |
| 80                | 0.76          | 0.90          | 0.95           |
| 90                | 0.86          | 0.95          | 0.95           |

5. Conclusion and Future Work
From the experiments we have done, we can conclude that the pancreatic cancer classification using Kernel-based Support Vector Machine has high accuracy. The highest accuracy we got was 95.23%, using 90% training data and parameters C=50 and C=100. Going forward, we can use larger data sets and some modifications to get better accuracy and show that the Kernel-based Support Vector Machine (K SVM) can be used as a reference to help doctors in their final decision.

Acknowledgment
This research is supported financially by the Ministry of Research, Technology and Higher Education of the Republic of Indonesia (KEMENRISTEKDIKTI), with the PTUPT 2020 research grant scheme. And we thank the leaders of hospitals, doctors, and laboratory assistants at Al Islam Hospital Bandung, Bandung, West Java, who have shared knowledge and experience to help this research.

References
[1] National Cancer Institute. What is Cancer?. Accessed on September 20, 2019. https://www.cancer.gov/about-cancer/understanding/what-is-cancer
[2] American Cancer Society. Pancreatic Cancer: About Pancreatic Cancer. Accessed on September 20, 2019. https://www.cancer.org/cancer/pancreatic-cancer/about/what-is-pancreatic-cancer.html
[3] Columbia University Irving Medical Center. The Pancreas Center. Accessed on September 20, 2019. https://columbiasurgery.org/pancreas/pancreas-and-its-functions
[4] Cancer Treatment Centers of America. Pancreatic Cancer. Accessed on September 20, 2019. https://www.cancercenter.com/cancer-types/pancreatic-cancer
[5] Lee, H.S. et al. 2018. Combine use of CEMIP and CA 19-9 enhances diagnostic accuracy for pancreatic cancer. Scientific Reports 8, 3383.
[6] Tempero, M.A. et al. 1987. Relationship of Carbohydrate Antigen 19-9 and Lewis Antigens in Pancreatic Cancer. Cancer Research 47, 5501-5503.
[7] Van den Bosch, R.P. et al. 1996. Serum CA 19-9 determination in the management of pancreatic cancer. Hepatogastroenterology 43, 710–713.
[8] Pleskow, D.K. et al. 1989. Evaluation of a serologic marker, CA19-9, in the diagnosis of pancreatic cancer. Annals of Internal Medicine 110, 704–709.

[9] Chang, Ching-Wen. et al. 2018. Fast detection of tumor marker CA 19-9 using AlGaN/GaN high electron mobility transistors. Sensors and Actuators B: Chemical 267, 191-197.

[10] University of Rochester Medical Center. CA 19-9. Accessed on September 20, 2019. urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=ca_19_9

[11] Wu, X. et al. 2009. The Top Ten Algorithms in Data Mining. Chapman and Hall/CRC, 37-59.

[12] Rustam, Z. and Rampisela, T.V. 2019. Support vector machines and twin support vector machines for classification of schizophrenia data. International Journal of Engineering and Technology (IJET) 7, 6873-6877.

[13] Cortes, C. and Vapnik, V. N. 1995. Support-vector networks. Machine Learning 20, 273-297.

[14] Burges, C.J.C. and Crisp, D. J. 1999. Uniqueness of the SVM solution. Neural Information Processing Systems 12, 223-9.

[15] Cheng, K. et al. 2017. Mixed kernel function support vector regression for global sensitivity analysis. Mechanical System and Signal Processing 96, 201-214.

[16] Al-Anazi, A.F.K. 2011. Support Vector Machines for Petrophysical Modelling and Lithoclassification. ProQuest Dissertations and Theses; Thesis (Ph.D.)—University of Calgary (Canada) 72-10, 395.

[17] Putri, R.A. et al. 2019. Kernel Based Fuzzy C-Means Clustering for Chronic Sinusitis Classification. IOP Conference Series: Materials Science and Engineering 546.

[18] Hartini, S. and Rustam, Z. 2019. Classification of Breast Cancer using Fast Fuzzy Clustering based on Kernel. IOP Conference Series: Materials Science and Engineering 546.

[19] Bishop, C.M. 2006. Pattern Recognition and Machine Learning. New York: Springer.

[20] Liu, L. et al. 2014. Research on Kernel Function of Support Vector Machine. In: Huang YM., Chao HC., Deng DJ., Park J. (eds) Advanced Technologies, Embedded and Multimedia for Human-centric Computing. Lecture Notes in Electrical Engineering 260, 827-834.

[21] Arieritari, N.P.A.A. and Rustam, Z. 2018. Support Vector Machines for Classifying Policyholders Satisfactorily in Automobile Insurance. IOP Conference Series: Journal of Physics 1028.

[22] Rustam, Z. et al. Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infraction Datasets. International Journal on Advanced Science Engineering Information Technology 9.

[23] Powers, D.M.W. 2011. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies.

[24] Fawcett, T. 2006. An Introduction to ROC Analysis. Pattern Recognition Letters.