Prostate Cancer Classification Using Random Forest and Support Vector Machines

Z Rustam¹, and N Angie{*}

¹Department of Mathematics, University of Indonesia, Depok 16424, Indonesia

*e-mail: nurliaangied@sci.ui.ac.id

Abstract. Nowadays, it gets more types of diseases in the medical sector. For this reason, the role of technology is very important in assisting medical staff to overcome the problem. This research discusses about Prostate Cancer. Prostate Cancer is suffered commonly by males. There are no exact causes how Prostate Cancer occurs in males, but there are several risk factors of a Prostate Cancer, such as age, ethnic group, family history, diet, smoking, and world area. In this research, the classification to diagnose Prostate Cancer is using two methods, those are Random Forest (RF) and Support Vector Machines (SVM). By comparing accuracy of those two methods, we will know which method is better with a dataset that we have from Al-Islam Bandung Hospital, Indonesia. The result is given that Random Forest has a better accuracy than Support Vector Machines. The accuracy shows 97.30% with 80% of data training.

Keywords: Prostate cancer, classification, random forest, support vector machines

1. Introduction

In this research, we discuss about Prostate Cancer. Prostate Cancer is suffered commonly by males. Prostate cancer develops when the cells in the prostate out of control. The prostate itself is a gland [1]. The main function of prostate is to help body makes semen-the fluid that carries sperm [1]. There are two types of prostate cancer, those that grow slowly and quickly [1]. Someone who has a Prostate Cancer that grows slowly, does not need any treatment [1]. However, for quick-growing prostate cancer, treatment is needed to stop its spread because this quick-growing cancer tends to be aggressive [1].

There are several symptoms of Prostate Cancer, there are pain at lower pelvic, lower back, bone, hips, or upper thighs, painful ejaculation, frequently or trouble urinating, hematuria, and loss weight and appetite [2]. Prostate Cancer’s clinical incidence has been increasing. The risk factors of a Prostate Cancer are age, family history, vasectomy, race, and dietary fat [3]. The most prominent risk factor is a dietary fat [3].

The medical staffs will use methods to diagnose the Prostate Cancer from the symptoms. The methods include checking the infection, taking a PSA Test, and doing an examination of digital rectal [4]. For advanced cancer, there are further testing such as Magnetic Resonance Imaging (MRI) Scan, Computerized Tomography (CT) Scan, or Positron Emission Tomography (PET) Scan [4]. In this research, we use RF and SVM for diagnosing patients of Prostate Cancer.
Previous researches using other methods of classification $X^2$ Test and Binary Logistic Regression [5], Automatic Localization [6], Imaging Features [7, 8, 9]. In addition, the Random Forest was already being used in Land Cover Classification [10], Feature Selection [11], and Bioinformatics [12]. Meanwhile, the Support Vector Machines is currently applied in Detection Systems for Intrusions [13], Cancer Classification [14, 15], Detection of Traffic Incident [16], Sorting of Hyper Spectral Imagery [17], Schizophrenia Classification [18], Face Recognition [19, 20], and Analysis of Gene Expression Data [21].

2. Methods

2.1 Data

This research uses a dataset of patients suffering from Prostate Cancer at the Al-Islam Bandung Hospital, Indonesia. The dataset has six features, which are Age, Prostate Specific Antigen (PSA), Haemoglobin, Leukocytes, Haematocrit, and Thrombocytes. There are 185 observations with the values that consists a diagnosis for supporting the program we use. Based on Diagnosis, 0 was stated for the patients without prostate cancer and 1 was stated for patients with it, as shown in the table below:

| Age | PSA  | Haemoglobin | Leukocytes | Haematocrit | Thrombocytes | Diagnosis |
|-----|------|-------------|------------|-------------|--------------|-----------|
| 73  | 13.17| 12.8        | 17200      | 35.4        | 166000       | 1         |
| 38  | 1.31 | 15.4        | 11100      | 43.7        | 283000       | 0         |
| 44  | 0.93 | 11.3        | 7800       | 37.8        | 269000       | 0         |
| 44  | 0.48 | 14.6        | 9800       | 42.2        | 263000       | 1         |
| 45  | 3.75 | 16.3        | 10700      | 47          | 272000       | 0         |
| 47  | 1.65 | 13.1        | 18400      | 38.5        | 255000       | 0         |
| 49  | 1.21 | 15.3        | 16400      | 45.4        | 472000       | 0         |
| 50  | 33.74| 14.5        | 7300       | 42.1        | 202000       | 1         |
| 51  | 2.68 | 11.4        | 16800      | 34.3        | 275000       | 0         |
| 51  | 1.21 | 15.2        | 12900      | 45.2        | 338000       | 0         |

2.2 Support Vector Machine

Support Vector Machines (SVM) uses hyperplane as a data separator in classifying a problem. In two dimensions, this hyperplane is a line that divides a plane into two classes with a measuring margin where in each class lies on the either side. In this method, given training data labeled (supervised learning), then the algorithm produces an optimal hyperplane to categorize new examples. This research mostly refers to the book which is written by Christianini [22].

A sample of training dataset is given as $x_i \in \mathbb{R}^N$ and a set of associated label is notated $y_i \in \{-1,1\}$ for $i = 1,2,\ldots,N$, and $N$ is the quantity of data. The function of Support Vector Machines is finding optimum hyperplane shown in Equation (1):

$$f(x) = w \cdot x + b = 0$$ (1)
We summarize Support Vector Machines problems as a formulation of Quadratic Programming (QP) shows in Equation (2) with constraints which is shown in Equation (3):

$$\min \left( \frac{1}{2} ||w||^2 \right)$$  \hspace{1cm} (2)

$$y_i(w \cdot x_i + b) \geq 1, \forall i = 1, \ldots, N$$  \hspace{1cm} (3)

We use Lagrange Multiplier for resolving this challenge

$$L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{N} \alpha_i \left( y_i ((w \cdot x_i + b) - 1) \right), i = 1,2, \ldots, N$$  \hspace{1cm} (4)

By derivating $L(w, b, \alpha)$ respected to $w$ and $b$ both equal to zero, these two conditions is obtained.

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0$$  \hspace{1cm} (5)

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0$$  \hspace{1cm} (6)

In order to eliminate $w$ and $b$, we have to substitute (5) and (6) to (4), then derive a dual form:

$$L(\alpha) = \max \left( -\frac{1}{2} \sum_{i=1}^{N} \sum_{r=1}^{N} \alpha_i \alpha_r y_i y_r (x_i \cdot x_r) + \sum_{i=1}^{N} \alpha_i \right)$$  \hspace{1cm} (7)

subject to:

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, \ldots, N$$  \hspace{1cm} (8)

A kernel function, defined as $K(x_i, x_r) = \varphi(x_i) \cdot \varphi(x_r)$, was used to support the method in this research. Equation (7) can be written as:

$$L(\alpha) = \max \left( -\frac{1}{2} \sum_{i=1}^{N} \sum_{r=1}^{N} \alpha_i \alpha_r y_i y_r (x_i \cdot x_r) + \sum_{i=1}^{N} \alpha_i \right)$$  \hspace{1cm} (9)

By solving Equation (9) with constraints in Equation (8), it determines the Lagrange Multipliers and finally the function $f(x)$ obtained [24]:

$$w = \sum_{i=1}^{N} \alpha_i y_i x_i$$  \hspace{1cm} (10)

then,

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i x_i + b$$  \hspace{1cm} (11)
\[ b = \frac{1}{N_s} \sum_{i \in S} (y_i - \sum_{i \in S} a_i y_i) \]  \hspace{1cm} (12)

2.3 Random Forest [21]
Random Forest was developed by Leo Breiman [25]. It is a group of uncut classification or regression trees made from the selection of random samples from training data. Predictions are made by combining predictions of ensemble [26]. At each decision split, the features are chosen randomly [26]. Correlation between trees is reduced by randomly selecting features that increase predictive power and result in higher efficiency [26]. Each tree grows as [27]:
- By random sampling \( N \), if the number of cases in the training set is \( N \) but with replacement, from the original data. This sample will be used as training to grow trees.
- For \( M \) the number of input variables, the variable \( m \) is chosen such that \( m << M \) is determined at each node, the variable \( m \) is chosen randomly from \( M \).
- Every tree in Random Forest is planted as much as possible. No pruning is used.
Thus the advantages of Random Forest are [27] overcoming the problem of excessive compatibility, less sensitive to outlier data, parameters can be easily adjusted and therefore eliminate the need for tree pruning, the importance of variables and accuracy are generated automatically.

2.3.1 Algorithm of Random Forest for Classifying The Dataset.
Random Forest Classifier has the algorithm for itself that can be seen in Figure 2:

In the training set \( S := (x_1, y_1), \ldots, (x_n, y_n) \), input \( A, F \) features, sum of trees in the forest \( B \)
1. Select \( M \) trees from the dataset
2. Construct a Decision Tree from the \( M \) trees
3. Repeat step 1 and 2, \( B \) times
4. At each node:
5. For a tiny subset of \( F \), construct \( f \)
6. In \( f \), separate the best features
7. New records are given to the category that win the most votes
Result: \( D \) selected features that have highest accuracy

![Figure 2. Random Forest Classifier Algorithm [11].](image)

2.4 Confusion Matrix
For the classification problems, we need a method which is called Confusion matrix. This method is usually used to calculate accuracy in the concepts of data mining or the decision support systems.
Accuracy explains how a data is being classified.
\( T_P \) : Number of samples having prostate cancer diagnosed correctly
\( F_P \) : Sum of healthy people that were incorrectly identified to have prostate cancer
\( T_N \) : Number of healthy individuals correctly spotted
\( F_N \) : The amount of samples with prostate cancer that were incorrectly classified as healthy

| Actual Value | Recognize Value |
|--------------|----------------|
| Positive     | \( T_P \)       | \( F_P \)       |
| Negative     | \( F_N \)       | \( T_N \)       |

The formula for accuracy is seen below:
\[ \text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \] 

(13)

3. Experimental Results

This research used R Studio as a software for Random Forest and Support Vector Machines.

3.1 Prostate Cancer Using Random Forest

The results of Accuracy which are given by Random Forest Method can be seen at Table 3:

| Data Training | Accuracy   | Running Time |
|---------------|------------|--------------|
| 0.1           | 89.16%     | 0.08s        |
| 0.2           | 90.54%     | 0.08s        |
| 0.3           | 89.15%     | 0.06s        |
| 0.4           | 90.99%     | 0.07s        |
| 0.5           | 92.39%     | 0.06s        |
| 0.6           | 91.89%     | 0.05s        |
| 0.7           | 94.55%     | 0.07s        |
| **0.8**       | **97.30%** | **0.06s**    |
| 0.9           | 94.44%     | 0.06s        |

Based on Table 3, the highest accuracy is recorded when data training 80% with 97.30% and has 0.06 seconds for the running time, while the lowest accuracy is recorded when data training 30% with 89.15% and has 0.06 seconds for the running time.

| Data Training | Accuracy   | Running Time |
|---------------|------------|--------------|
| 0.1           | 81.92%     | 0.06s        |
| 0.2           | 66.89%     | 0.06s        |
| 0.3           | 66.67%     | 0.04s        |
| 0.4           | 66.67%     | 0.05s        |
| 0.5           | 63.04%     | 0.04s        |
| 0.6           | 66.22%     | 0.06s        |
| 0.7           | 67.27%     | 0.05s        |
| 0.8           | 72.97%     | 0.06s        |
| **0.9**       | **83.33%** | **0.05s**    |
Based on Table 4, the highest accuracy is recorded when data training 90% with 83.33% and has 0.05 seconds for the running time, while the lowest accuracy is recorded when data training 50% with 63.04% and has 0.04 seconds for the running time.

Figures 3 and 4 below showing us comparison the accuracy and the running time of prostate cancer classification between Random Forest and Support Vector Machines:

![Figure 3. Graph of Accuracy Prostate Cancer Classification.](image)

Figure 3 shows the result of accuracy by Random Forest and Support Vector Machines. It seems like accuracy of Random Forest is more consistent than Support Vector Machines. This is indicated by the graph of Random Forest tends to go up, meanwhile in Support Vector Machines, the graph is less consistent. Figure 4 shows the result of running time by Random Forest and Support Vector Machines. We can see that Random Forest seems to need more time to run than Support Vector Machines.

![Figure 4. Graph of Running Time Prostate Cancer Classification.](image)

4. Conclusion

We can conclude that Random Forest is better for classifying prostate cancer data than Support Vector Machines, especially on the accuracy. The accuracy of Random Forest reaches a highest rate at 97.30% with 80% of data training and running time was 0.06 seconds, while Support Vector Machines only reaches rate at 83.33% with 90% of data training and running time was 0.05 seconds. In addition, there is a hope that this method allows medical staff to easily classify diseases or other medical conditions.
problems. Furthermore, to create more optimal models, it is better to use larger dataset for solving another classification problems.

**Acknowledgement**

This research supported financially by the Ministry of Research, Technology, and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with PTUPT 2020 research grant scheme.

**References**

[1] Prostate Cancer UK accessed on October 2, 2019 see https://prostatecanceruk.org/prostateinformation/about-prostate-cancer

[2] The Official Foundation of The American Urological Association accessed on October 2, 2019 see https://www.urologyhealth.org/urologic-conditions/prostate-cancer

[3] Kenneth J. and Peggy S. 1993 *Risk Factors for Prostate Cancer* SN 0003-4819

[4] NHS Prostate Cancer accessed on October 2, 2019 see https://www.nhs.uk/conditions/prostatecancer/diagnosis/

[5] de Lima C O, Devita K L, Vasconcelos L R B, do Prado M, and Campos C N 2017 *Correlation between Endodontic Infection and Periodontal Disease and Their Association with Chronic Sinusitis: A Clinical-tomographic Study*, American Association of Endodontists

[6] Campbell A P, Metson R, and Bergmark R W 2017 *Orbital Complications of Acute Sinusitis*, Operative Techniques in Otolaryngology - Head and Neck Surgery.

[7] Reede D, Shinder R, Velayudhan V, Smoker W R K, and Chaudhry Z A 2017 *Imaging of Intracranial and Orbital Complications of Sinusitis and Atypical Sinus Infection: What the Radiologist Needs to Know*, Current Problems in Diagnostic Radiology.

[8] Lakanh K J T 2018 *Sinus Headaches Sinusitis Versus Migraine*, Physician Assist Clin. 3, 181-192.

[9] Wyler B and Mallon W K 2019 *Sinusitis Update*, Emerg Med Clin N Am, 37, 41–54.

[10] Pall Oskar Gislason, Jon Atli Benediktsson, Johannes R. Sveinsson, Random Forests for land cover classification, Pattern Recognition Letters, Volume 27, Issue 4, 2006, Pages 294-300, ISSN 0167-8655.

[11] Rustam Z, Huljanah M, Suarsih U, and Titin S 2019 *Feature Selection using Random Forest Classifier for Predicting Prostate Cancer*, IOP Conf 546:052031

[12] Qi Y. (2012) Random Forest for Bioinformatics. In: Zhang C., Ma Y. (eds) Ensemble Machine Learning. Springer, Boston, MA.

[13] Rustam Z and Zahras D 2018 Comparison Between Support Vector Machine and Fuzzy C-Means as Classifier for Intrusion Detection System, Journal of Physics: Conference Series 1028 (1), 012227.

[14] Rustam Z, Primasari I, and Widya D 2018 Classification of Cancer Data Based on Support Vectors Machines with Feature Selection using Genetic Algorithm and Laplacian Score, AIP Conference Proceedings, 2023 (1), 020234.

[15] Rustam Z 2018 Classification of Cancer Data using Support Vector Machines with Feature Selection Method Based on Global Artificial Bee Colony, AIP Conference Proceedings, 2023 (1), 020205.

[16] Xiao J 2019 SVM and KNN Ensemble Learning for Traffic Incident Detection, Physica A 517,29–35.

[17] Bing G, Xiaoping W, Chunhui Z, and Lejun Z 2018 *Classification of Hyperspectral Imagery Based on Spectral Gradient, SVM and Spatial Random Forest*, Infrared Physics and Technology 95, 61–69.

[18] Rustam Z and Rampisela T V 2018 Classification of Schizophrenia data using Support Vector Machine (SVM), Journal of Physics: Conference Series 1108 (1), 012044

[19] Rustam Z and Faradina R 2018 Face Recognition to Identify Look-Alike Faces using Support Vector Machine, Journal of Physics: Conference Series 1108 (1), 012071.
[20] Rustam Z and Ruvita A A 2018 Application Support Vector Machine on Face Recognition for Gender Classification, Journal of Physics: Conference Series 1108 (1), 012067.

[21] Rustam Z, Sarwinda D, and Sudarsono E 2019 Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease.

[22] N. Christianini and J.S. Taylor. (1999). An Introduction to Support Vector Machines and Other Kernel based Learning Methods, Cambridge.

[23] Elish, Karim O. and Mahmoud O. Elish. “Predicting defect-prone software modules using support vector machines.” Journal of Systems and Software 81 (2008): 649-660.

[24] Z Rustam and F. Yaurita. Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means. Journal of Physics: Conference Series. 2018.

[25] Breiman, L., Random Forests, Machine Learning 45(1), 5-32, 2001.

[26] Jehad Ali, R. Khan, N. Ahmad, I. Maqsood, “Random Forests and Decision Trees.”, International Journal of Computer Science Issues, Vol 9, Issue 5, No. 3, 1697-0814, 2012.

[27] Introduction to Decision Trees and Random Forests, Ned Horning; American Museum of Natural History’s.