Feature Selection using Random Forest Classifier for Predicting Prostate Cancer

Mia Huljanah\textsuperscript{1}, Zuherman Rustam\textsuperscript{1,*}, Suarsih Utama\textsuperscript{1} and Titin Siswantining\textsuperscript{1}

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

*Correspondence author : rustam@ui.ac.id

Abstract. Prostate cancer is cancer that attacks the prostate gland, usually affecting men over 50 years. Prostate cancer is a disease that develops slowly. Based on this, rapid and precise detection is needed so that the disease can be treated immediately. This study focuses on the application Feature Selection using the Random Forest Classifier to detect prostate cancer. The Random Forest Classifier is a method of classifying data by determining the decision tree. The use of more trees will affect the accuracy to be obtained for the better. The Random Forest Classifier can classify data that has incomplete attributes and can be used to handle large sample data. Selection of features is an important process because it can affect the accuracy of classification. This method increases accuracy by about 87%. Thus, the selection of features can improve accuracy in the detection of prostate cancer.

Keywords: Random Forest Classifier, Prostate Cancer, Machine Learning, Feature Selection

1. Introduction

Cancer is one of the leading causes of death worldwide, around 9.6 million deaths were caused by cancer in 2018. Cancer is a disease that arises due to the abnormal growth of body tissue cells that turn into cancer cells, while tumors are conditions in which cell growth is abnormal so that it forms lesions or in many cases, lumps in the body. The most common cancers are Lung Cancer and Breast Cancer with each 2.09 million cases, Colorectal Cancer with 1.80 million cases, and Prostate Cancer with 1.28 million cases in 2018 [1].

One of the most common cancers in men is Prostate Cancer, cancer that attacks prostate gland. The prostate gland is a walnut-sized gland that is contained in the male reproductive system, which is located between the neck of the bladder and the urinary tract (urethra). The prostate secretes a white liquid that nourishes and transports sperm, which is called semen. The male hormone secreted by the testes directly affects the growth and function of the prostate. In Indonesia, prostate cancer is also known as third cancer that most often affects men in Indonesia, where one in 10 men, especially the elderly, suffer from this cancer. In Australia, the prevalence of prostate cancer is higher than breast cancer. Recorded, there are more than 3000 Australian men who die each year from this cancer. This figure makes Australia a country with the highest mortality due to prostate cancer in the world [2].

Prostate cancer is defined as a deadly cancer because the disease develops slowly, is able to stay in the body for years, and often appears undetected - no symptoms are felt at the onset of cancer cells but only after an advanced stage. Symptoms that can be felt in patients with prostate cancer, among others, feel the urge to urinate suddenly, feels difficult and uncomfortable when urinating, there is blood in urine or semen, pain behind the waist or groin.
This disease is usually detected and diagnosed with a blood test using Prostate-Specific Antigen test. Prostate-Specific Antigen is a protein produced by cells of the prostate gland. The value of the PSA level at normal levels ranges from 1.0 - 4.0 ng/ml. High PSA levels do not necessarily indicate prostate cancer because some other prostate diseases can also increase PSA levels so that further specialist checks are needed. However, the risk of prostate cancer in people who experience an increase in PSA levels is at least 2.0 ng/ml, which means the risk of prostate cancer is 7.1%. If the PSA level increases to 2.0-3.9 ng/ml, the risk increases to 18.7%. PSA of 4.0-5.9 ng/ml has a risk value of 21.3%. While at the level of 6.0 - 7.9% the value of the risk is in the range of 28.6%. In the range of 8.0 - 9.9 ng/ml, the risk is 31.7% and for levels above 10.0 ng/ml, the risk is 56.5% [3,4].

Fofanov et al (2019) used a targeted post-ligation amplification sequencing approach, called next-generation copy number alteration assay, to predict tumor metastatic potential based on prone areas to copy changes score in metastatic prostate cancer [5]. In Sharan et al (2018), Support Vector Machine-Recursive Elimination Features (SVM-RFE) and Absolute Cosine filters method are used to categorize prostate cancer based on the texture of individual tissue components [6].

Preston et al (2019) tested the initial PSA level during midlife predicting the risk of aggressive prostate cancer in black men using statistical analysis methods. The results show that PSA levels in middle age strongly predicted total and aggressive prostate cancer among black men and white men [7]. This paper focuses on selecting the most influential features in predicting prostate cancer and increasing accuracy of selecting features in datasets using classification.

Classification is one case that uses labeled data in Supervised learning. Classification divides the dataset into training data and test data to predict new classes by studying the categories and labels of the class [8]. One method that quite famous in classification is random forest.

Random forest is one way of applying the stochastic discrimination approach to classification. The Classification process will run if all the trees have been formed. When the classification process is complete, initialization is done with as much data based on its accuracy value. The advantage of using the random forest is that it is able to classify data that has incomplete attributes, can be used for classification and regression but are not very good for regression, more suitable for classifying data and can be used to handle large sample data.

2. Method

2.1. Features Selection

Feature Selection is a basic concept in machine learning that has a considerable impact on the performance of the model. Feature selection is important for classification because this process removes irrelevant features so that it can improve model performance, make the model easier to understand, and reduce running time. Features Selection is divided into three (3) types in general, Filter methods, Wrapper methods, and Embedded methods [9].

Filter method is a method commonly used in preprocessing data. This method combines ranking techniques with the main criteria and uses sorting techniques to select variables. The ranking method filters out irrelevant features before starting the classification process. Advantages of this method are simplicity, amazing results and relevant features, and independent of any machine learning algorithm.

The wrapped method looks for features that are suitable for the machine learning algorithm used. If the filter method is used before the machine learning algorithm, the wrapped method is used before and when machine learning works until a suitable feature is found. features are evaluated using predictive accuracy in classification case and goodness cluster in clustering case.

Embedded methods are methods that maintain each iteration of the model training process and extract features that contribute most to training for certain iterations carefully. The most commonly used embedded method is the regularization method, a method that punishes features by giving a coefficient threshold. Some examples of regularization algorithms are LASSO, Elastic Net, and Select From Model.
Select From Model is a feature selection that removes features if corresponding coefficients or important features are below the parameter threshold provided. This method is used with estimators that have important features or coefficients.

2.2. Random Forest Classifier
Random Forest is one of the supervised learning algorithms that are flexible, easy to use, and without creating hyper-parameters. This algorithm is quite effective in classifying. In Random Forest, there is a limit to the minimum number of trees that must be built, so that in this amount all data has been classified. Where the amount is very dependent on each data [10].

The number of breaker attributes affects the minimum number of trees for each data. The number of trees has a large influence on the level of accuracy. Starting at the minimum number of trees, increasing the number of trees, increasing the accuracy produced. There is the best limit of accuracy, where after accuracy is achieved, even though the number of trees plus accuracy will remain stable. The accuracy produced by this algorithm is influenced by the number of breaker attributes. The Random Forest that uses the number of breaker attributes equal to the number of attributes available will provide low accuracy.

2.3. Algorithm
Feature Selection using Select from Model starts with determining the threshold value to give a boundary between the features to be selected and the features that will be eliminated, then all features will be sorted by Gini importance score from the smallest to the largest. Furthermore, features with Gini importance score that are below the threshold value will be eliminated. Selected features will be used in the Random Forest algorithm (See Figure.1).

---

**Figure 1.** Select from Model Feature Selection Algorithm

| 1 Determine the threshold |
| 2. Sort the value of the feature |
| 3. Elimination the feature below the threshold |
| 4. Enter selected features in Machine Learning Algorithm |
| 5. Test the Performance of the Model |

---

**Figure 2.** Random Forest Classifier Algorithm

**Input:** A training set $S := (x_1, y_1), ..., (x_n, y_n)$, $F$ features, and number of trees in forest $B$

1. Select $M$ trees from the dataset
2. Construct a decision tree from the $M$ trees.
3. Repeat step 1 and step 2, $B$ times.
4. At each node:
   5. Construct $f$ as a tiny subset of $F$
   6. Split on best feature in $f$
7. New records are given to the category that wins the most votes

**Results:** $D$ selected features that have highest accuracy

---

Figure 2 shows an algorithm from the Random Forest Classifier. This algorithm runs by inputting the $S$ training set, $F$ feature where $F$ is constant, and the $M$ trees where $M$ is constants. Randomly select the $M$ tree from the dataset, then the selected $M$ tree is used to build decision trees. Decision trees are made $B$ times. At each node, make the smallest subset of feature $F$ and separate the best features for each $f$. The result is the selected $D$ feature which has the highest score.
2.4. Confusion Matrix
In classification, method performance can be measured using confusion matrix. Confusion matrix contains a comparison of the results of the classification carried out by the system with the results of the classification that should be [11].

There are 4 (four) conditions in confusion matrix for measuring performance, namely, True Positive (TP) which indicates the positive data entered into the system is detected correctly by the system, False Positive (FP) indicates the negative data entered into the system is detected correctly by the system, True Negative (TN) indicates negative data entered into the system is detected incorrectly by the system, and False Negative (FN) indicates positive data entered into the system is detected incorrectly by the system. See Table 1 to see the Confusion Matrix.

| Class       | Actual        |
|-------------|---------------|
|             | Positive      | Negative      |
| Predict     |               |               |
| Positive    | True Positive | False Negative|
| Negative    | False Positive| True Negative  |

Table 1. Confusion Matrix.

Based on the value of True Negative (TN), False Positive (FP), False Negative (FN), and True Positive (TP) can be obtained the value of accuracy, precision, and recall. Accuracy values describe how accurately the system can classify data correctly. In other words, the value of accuracy is a comparison between data that is correctly classified and the overall data. Accuracy values can be obtained by Equation 1. Precision values explain the level of accuracy between the information requested with and the answer given by the system (see Equation 2). Whereas recall is the success rate of the system in rediscovering information (see Equation 3) [12].

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \tag{1}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} * 100\% \tag{2}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} * 100\% \tag{3}
\]

3. Experiment
In this study, we use data from Al-Islam Bandung Hospital. In this data consists of 185 patients and 7 features. Each features states age, PSA (Prostate-Specific Antigen), Haemoglobin, Leukocytes, Haematocrit, Thrombocytes, and Diagnose. Normal PSA level range from 1.0-4.0 ng/mL. Normal levels Haemoglobin range from 13-18 g/dL, Leukocytes are on 4000-10000 cell/µL, and Thrombocytes at intervals 150000-450000 cell/µL. Normal Haematocrit levels are at 40-54%. Diagnose is the condition of the patient, with 1 for a positive diagnosis of prostate cancer and 0 for negative (See Table 2).
Table 2. Prostate Cancer Datasets.

| Age | PSA   | Hemoglobin | Leukocytes | Hematocrit | Thrombocytes | Diagnose |
|-----|-------|------------|------------|------------|--------------|----------|
| 0   | 73    | 13.17      | 12.8       | 17200      | 35.4         | 166000   | 1        |
| 1   | 64    | 25.15      | 15.6       | 12900      | 46.7         | 226000   | 1        |
| 2   | 71    | 8.95       | 14.5       | 11100      | 39.4         | 224000   | 1        |
| 3   | 71    | 9.51       | 14.5       | 11100      | 39.4         | 224000   | 1        |
| 4   | 72    | 8.57       | 14.4       | 8900       | 40.5         | 286000   | 1        |
| 5   | 72    | 9.14       | 13.2       | 10500      | 37.2         | 215000   | 1        |
| 6   | 70    | 2.16       | 14.0       | 7800       | 41.4         | 325000   | 0        |
| 7   | 73    | 22.2       | 14.0       | 18300      | 42.3         | 406000   | 1        |
| 8   | 71    | 1.14       | 12.0       | 7900       | 38.2         | 416000   | 0        |
| 9   | 59    | 6.11       | 10.9       | 7500       | 33.1         | 275000   | 1        |
| 10  | 58    | 2.17       | 13.2       | 6700       | 39.3         | 241000   | 0        |

4. Result
This study using Select from Model Feature Selection and Random Forest Classifier in Python 3.0. Results of Random Forest Classifier algorithm of the 7 initial features with nTrees = 1000 decision trees, performed on the training sample, are given in Table 3.

Table 3. Accuracy of Random Forest Classifier with all Features.

| Data Training (%) | Accuracy (%) | Precision (%) |
|-------------------|--------------|---------------|
| 50                | 98.92473118  | 97.917        |
| 55                | 99.01960784  | 98            |
| 60                | 99.0990991   | 98.113        |
| 65                | 99.17355372  | 98.246        |
| 70                | 99.23076923  | 98.413        |
| 75                | 99.28057554  | 98.485        |
| 80                | 99.32432432  | 98.63         |
| 85                | 99.36708861  | 98.734        |
| 90                | 97.63313609  | 95.9          |
| 95                | 94.88636364  | 91.297        |

Experiments using 85% of training data showed the best accuracy of 99.376% with a precision of 98.734%. As for system performance can be seen using the confusion matrix in Table 4. The table shows that from 158 data (equivalent to 85% training data), 79 positive data are correctly classified by the system, 78 negative data are incorrectly classified, 0 negative data are correctly classified, and 1 Positive data is classified incorrectly by the system.

Using data training 85%, we found that PSA had a 59.438% effect in diagnosing prostate cancer. As for the influence of other features in diagnosing this disease, Leukocytes are 11.77%, Hematocrit is 8.238%, Hemoglobin is 7.512%, Thrombocytes is 7.225%, and Age is 5.761%. (See Table 5).

Table 4. Confusion Matrix for 85% Data Training.

| Class | Actual |      |      |
|-------|--------|------|------|
|       |        | Positive | Negative |
| Predict | Positive | 79   | 1    |
|        | Negative | 0    | 78   |
Table 5. Features Importance.

| Data Training (%) | Age (%) | PSA (%) | Haemoglobin (%) | Leukocytes (%) | Hematocrit (%) | Thrombocytes (%) |
|-------------------|---------|---------|-----------------|----------------|----------------|-----------------|
| 85                | 5.761   | 59.438  | 7.512           | 11.777         | 8.283          | 7.225           |

Furthermore, the training data is tested again by selecting existing features. This experiment shows very good results (See Table 6).

Table 6. Accuracy of Random Forest Classifier with Features Selection.

| Data Training (%) | Accuracy (%) | Precision (%) | Recall (%) | f_1 score (%) |
|-------------------|--------------|---------------|------------|---------------|
| 85                | 100          | 100           | 100        | 100           |

Table 7 shows that from 158 data (equivalent to 85% training data), 80 positive data are correctly classified by the system, 78 negative data are incorrectly classified, 0 negative data are correctly classified, and 0 positive data is classified incorrectly by the system.

Table 7. Confusion Matrix for 85% Data Training with Features Selection.

| Class | Actual | Predict | Positive | Negative |
|-------|--------|---------|----------|----------|
|       |        |         | 80       | 0        |
| Positive | |         |
| Negative | |         | 0        | 78       |

Running time needed by the system in predicting prostate cancer can be seen in Table 8.

Table 8. Running Time.

| Data Training (%) | Time (seconds) |
|-------------------|---------------|
| 85                | 11.913014888763428 |

5. Discussion

In this study, we found that the Random Forest Classifier provides good results for each training data, which is around 90-99.5%. As expected, the feature selection can improve the accuracy of the models we make. This is shown in the results of using 85% training data, accuracy increases from 99.367% to 100%.

The study also found that PSA affects the prediction of prostate cancer by more than 50%. The thing to note is that high PSA does not necessarily indicate the presence of prostate cancer. High PSA values may indicate urinary tract infections or inflammation of the prostate. However, if the PSA value is higher, the risk of prostate cancer is higher.

In Table 5, it can be seen that age affects the prediction of 5.761%. This is in line with Bechis et al (2011) which states that 26% of men ≥ 75 years old are presented with high-risk diseases [13]. In the
study of Fujita et al (2016), found that the number of leukocytes associated with the degree of prostate enlargement and lower urinary tract symptoms [14]. This certainly affects the prediction of prostate cancer. This result also shown by experiments, leukocytes became the second most influential feature in predicting prostate cancer by 11.777%.

The unexpected thing in this study was the influence of hematocrit in predicting prostate cancer. As far as the author's observation, there have no studies linking the influence of hematocrit on prostate cancer. In this study, hematocrit was the third most influential feature with a score of 7.225%.

Judging from the confusion matrix, it can be concluded that the Random Forest Classifier has good performance in predicting prostate cancer. Table 8 shows that Random Forest Classifier can predict prostate cancer quickly, around 12 seconds.

6. Conclusion
In this research, three important things were produced. First, the Random Forest Classifier can predict prostate cancer well and quickly. Second, the most influential feature in predicting prostate cancer is PSA, leukocyte, and haematocrit levels. Third, Feature Selection increases the accuracy of the model to 100%.

Acknowledgement
This research was financially supported by University of Indonesia, with PIT.9 2019 research grant scheme (ID number NKB-0039/UN2.R3.1/HKP.05.00/2019).

References
[1] World Health Organization. 2018. Cancer. Available at https://www.who.int/news-room/fact-sheets/detail/cancer [accessed on 2019].
[2] YayasanKanker Indonesia. 2018. MelawanKeganasanKankerProstat. Available at http://yayasanankerindonesia.org/article/melawan-keganasan-kanker-prostat [accessed on 2019].
[3] Prostate Cancer Foundation of Australia. What You Need to Know about Prostate Cancer. Available at http://www.prostate.org.au/awareness/general-information/what-you-need-to-know-about-prostate-cancer/ [accessed on 2019].
[4] American Urological Association Education and Research, Inc. 2013. PSA testing for the pretreatment staging and posttreatment management of prostate cancer: 2013 Revision of 2009 Best Practice Statement. Linthicum, MD. [accessed on 2019]. Available at http://www.auanet.org/common/pdf/education/clinical-guidance/Prostate-Specific-Antigen.pdf. [accessed on 2019].
[5] Fofanov V Y, Upadhyay K, Pearlman A, Loke J, O V, Shao Y, Freedland S, Ostrer H. 2019. Rapid Next-Generation Sequencing Method for Predicting of Prostate Cancer Risks. The Journal of Molecular Diagnostics, Vol. 21, No. 1, pp. 49-57
[6] Sahran, Shahnorbunun., et all. 2018. Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading. Artificial Intelligence in Medicine 87 pp. 78-90
[7] Preston, Mark A., et all. 2019. Baseline Prostate-spesific Antigen Level in Midlife and Aggressive Prostate Cancer in Black Men. European Urology 75 pp.399-407
[8] Nadira, T. Rustam, Z. Classification of Cancer Data Using Support Vector Machines with Features Selection Method Based on Global Artificial Bee Colony
[9] Brownlee, Jason. 2014. An Introduction to Feature Selection. [Available at https://machinelearningmastery.com/an-introduction-to-feature-selection/ [accessed on 2019].
[10] Malik, Usman. 2018. Random Forest Algorithm with Python and Scikit-Learn. . Available at https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/ [accessed on 2019].
[11] Prasetyo, E., 2012. Data Mining konsepdanAplikasimenggunakan MATLAB. Yogyakarta: Andi.
[12] Saxena, Shruti. 2018. Precision vs Recall. Available at https://towardsdatascience.com/precision-vs-recall-386cf9f89488 [accessed on 2019].

[13] Bechis, S.K. Carroll P R, Cooperberg M R. 2011. Impact of Age at Diagnosis on Prostate Cancer Treatment and Survival. *Journal of Clinical Oncology* 29 pp.235-241.

[14] Fujita K, Hosomi M, Nakagawa M, Tanigawa G, Imamura R, Uemura M, Nakai Y, Takayama H, Yamaguchi S, and Nonomura N. 2013. White blood cell count is positively associated with benign prostatic hyperplasia. International *Journal of Urology*. 21 (3) pp.308-12.