Multiclass classification of acute lymphoblastic leukemia microarrays data using support vector machine algorithms

Hamidah¹, a¹, Z Rustam¹, b¹, S Utama¹, c, T Siswantining¹, d

¹Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), University of Indonesia, Depok 16424, Indonesia

corresponding author’s e-mail: a¹hamidah61@sci.ui.ac.id, b¹rustam@ui.ac.id, c³suarsih.utama@sci.ui.ac.id, d³titin@sci.ui.ac.id

Abstract. Acute lymphoblastic leukemia (ALL) is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblast. Classification of acute lymphoblastic leukemia subtypes based on fusion genes that have a translocation. The fusion genes are BCR-ABL, E2A-PBX1, Hyperdiploid > 50 chromosomes, MLL, T-ALL, and TEL-AML1. The classification of acute lymphoblastic leukemia subtypes has an important role for the type of treatment that will be received, duration of treatment, medication needed during treatment, and other treatments that may be needed. In this paper, the method used is Multiclass Support Vector Machine Recursive Feature Elimination (MSVM-RFE) as the feature selection and One-Against-One Multiclass Support Vector Machine (OAO-MSVM) with RBF-Kernel with σ = 0.01 and Polynomial-Kernel with d = 4 as the classification methods. For the multiclass classification of acute lymphoblastic leukemia microarrays data, the best method to use is the MSVM Polynomial-Kernel with d = 4 that produces overall accuracy about 94%, precision about 96%, recall about 95%, F1 score about 95%, and the running time is 0.66 seconds.

1. Introduction

According to WHO (World Health Organization), cancer is a disease that causes the second highest death in the world. About 9.6 billion people died in 2018 due to cancer [1]. Cancer has many types depending on the location of the cancer. If cancer occurs in blood cells, it is called leukemia. According to the National Cancer Institute, about 61,780 new cases of leukemia occurred in 2019. About 22,840 people worldwide died from leukemia. For five years in 2009-2015, 62.7% of people with leukemia survived [2]. Leukemia has four subtypes; Acute Lymphoblastic Leukemia (ALL), Chronic Lymphoblastic Leukemia (CLL), Acute Myelogenous Leukemia (AML), and Chronic Myelogenous Leukemia (CML) [3]. Acute leukemia, cancer growth is fast, whereas chronic leukemia, cancer growth is slow. This research will only discuss the acute lymphoblastic leukemia (ALL) and its classification.

Acute lymphoblastic leukemia is the most common type of childhood cancer [4]. In the United States in 2016, there were an estimated 95,764 people living with acute lymphoblastic leukemia, about 54.2% less than 20 years old [5]. Acute lymphoblastic leukemia (ALL) is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblast [6]. Lymphoblast is a progenitor cell that will form white blood cells (leukocytes). Acute lymphoblastic leukemia occurs due to abnormal formation of lymphoblast. Abnormal lymphoblast cell will make mistakes in making copies of chromosomes. This mistakes is called translocation. Translocation is caused by two genes to attach and form an abnormal of fusion gene [7]. The abnormal fusion genes are BCR-ABL, E2A-PBX1, Hyperdiploid > 50
chromosomes, MLL, T-ALL, and TEL-AML1. Those fusion genes are the basis for the classification of acute lymphoblastic leukemia subtypes. The classification of acute lymphoblastic leukemia subtypes has an important role for the type of treatment that will be received, duration of treatment, medication needed during treatment, and other treatments that may be needed by patients with acute lymphoblastic leukemia.

Machine learning is an algorithm-based study that can do classification, prediction, and clustering. Machine learning has several types of learning, one of them is supervised learning that usually used for classification and prediction. Supervised learning has several methods, they are support vector machine (SVM), nearest neighbor, neural network, naïve Bayes, decision tree, etc. Support vector machine (SVM) is one of the most well-known methods. SVM has been implemented to analyze, predict and classify problems in various fields. One of them is the insurance field, such as insolvency prediction in insurance companies [8]. SVM also has been implemented in the medical world in helping to analyze, predict, and classify a disease. They are analysis of gene expression data in chronic kidney disease [9], classification of schizophrenia data [10], classification of infarction for detecting ischemic stroke [11], classification of soft tissue tumors [12], classification of acute sinusitis [13], classification of imbalanced cerebral infarction datasets [14], etc. SVM can also be used as a feature selection method on medical microarrays data, they are brain cancer microarrays data [15], colon cancer microarray data, prostate cancer microarray data, and lymphoma cancer microarray data [16]. However, the SVM algorithm is designed to solve binary class problems, whereas for multiclass problems, SVM cannot solve multiclass problems directly.

In this paper, the data used are microarrays data from acute lymphoblastic leukemia that contain mRNA expression from each sample. Microarray data is a type of medical data that is transformed into a numeric type. Microarray data generally has many features but few samples. The data has 985 features and 248 samples. This condition can create a large computational cost. Therefore, to reduce the large computational costs, this study conducted a feature selection. Several feature selection methods have been implemented in disease classification, such as cancer classification with feature selection using Fisher’s Ratio [17] and breast cancer classification with feature selection using Laplacian Score [18]. The method used in this paper for feature selection is Multiclass Support Vector Machine Recursive Feature Elimination (MSVM-RFE). MSVM-RFE is an extension of the SVM-RFE method from binary problems onto multiclass problems. The SVM-RFE was first introduced by Guyon in 2002 [19], while MSVM-RFE was first introduced by Zhou in 2007 [20]. The classification methods used are One-Against-One Multiclass Support Vector Machine (OAO-MSVM) with RBF-Kernel and Polynomial-Kernel. Moreover, OAO-MSVM is the extension of SVM from binary problems to multiclass problems.

This paper consists of the materials and methods that used to multiclass classification of acute lymphoblastic leukemia data, the experimental results, and the conclusion.

2. Materials and methods

2.1. Dataset

The dataset used in this paper is acute lymphoblastic leukemia microarrays data in St. Jude Children's Research Hospital obtained from https://broadinstitute.org/cgi-bin/cancer/datasets.cgi. Data consists of 985 features and 248 samples. Table 1 shows part of the data taken from St. Jude Children’s Research Hospital.

Table 1. Examples of Acute Lymphoblastic Leukemia Data

| ID         | Class            |
|------------|------------------|
| 34303_at   | -1844.56         |
| 31472_s_at | 530.4            |
| 40789_at   | 1630.23          |
| 32251_at   | 8817.16          |
| 33224_at   | 9132.29          |
| 249.09     | Hyperdiploid > 50 chromosomes |
| 1021.87    | MLL              |
| 3212.40    | E2A-PBX1         |
| 3327.69    | 940.49           |
| 2571.58    | 12988.54         |
| 3284.71    | 2571.58          |
| 3284.71    | 3284.71          |

2
In Table 1, columns 34303_at, 31472_s_at, 40789_at, 32251_at are some of the features used in this study. These features are mRNA expression codes of acute lymphoblastic leukemia patients. The class column is the classification class of the acute lymphoblastic leukemia. There are six class, they are BCR-ABL, E2A-PBX1, Hyperdiploid > 50 chromosomes, MLL, T-ALL, and TEL-AML1. These classes are the fusion genes where translocation occurs. The following is the number of each sample from each class:

- Class A (BCR-ABL): 15 samples
- Class B (E2A-PBX1): 27 samples
- Class C (Hyperdiploid > 50 chromosomes): 64 samples
- Class D (MLL): 20 samples
- Class E (T-ALL): 43 samples
- Class F (TEL-AML1): 79 samples

2.2. Support Vector Machine (SVM)
SVM is one of the most frequently used methods in classification. SVM and its modification is currently one of the most well-known classification techniques with computational advantages over their competitors [21]. SVM aims to find hyperplane where the margin value is maximum. The Hyperplane is [22]:

\[ f(x) = \text{sign}(w^T x + b) \]  

where

- \( w \) = weight parameter
- \( x \) = data vector
- \( b \) = bias parameter

Figure 1 shows an illustration of SVM [23].
\[ \begin{align*}
\text{min} & \quad \frac{1}{2} ||w||^2 \\
\text{subject to} & \quad y_i [w^T x_i + b] \geq 1
\end{align*} \] (2)

where \( y_i \) is class of \( i \)-th data and \( x_i \) is vector feature of \( i \)-th data.

If there is an error tolerance in the classification, then the primal optimization of SVM is [24]:

\[ \begin{align*}
\text{min} & \quad \frac{1}{2} ||w||^2 + C \sum_{i=1}^{D} \xi_i \\
\text{subject to} & \quad y_i [w^T x_i + b] \geq 1 - \xi_i \\
\xi_i & \geq 0
\end{align*} \] (4)

where \( \xi_i \) are slack variables and constant \( C > 0 \) sets the relative importance of maximizing margins and minimizing the amount of slack [24]. The equations stated in (4), (5), and (6) are called soft margin SVM that was introduced by Cortes and Vapnik in 1995 [24].

When a problem is not linearly separable in input space, soft margin SVM cannot find a robust separating hyperplane that minimizes the number of misclassified data points and that generalizes well [23]. To overcome this, the kernel strategy was applied in SVM. Let \( X^n \) is an input space and \( F \) is a feature space. \( \forall x_i, x_j \in X^n \), the kernel is defined as \( K(x_i, x_j) = < \varphi(x_i), \varphi(x_j) > \), where \( \varphi: X^n \to F \) [22].

The primal formulation of the kernel SVM is [22]:

\[ \begin{align*}
\text{min} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^{D} \xi_i \\
\text{subject to} & \quad y_i (w^T \varphi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, \forall i
\end{align*} \] (7)

Using the method of Lagrange multipliers, we can obtain the dual optimization of the kernel SVM which is expressed in terms of variables \( \lambda_i \) [22]:

\[ \begin{align*}
\max & \quad \sum_{i=1}^{D} \lambda_i - \frac{1}{2} \sum_{i=1}^{D} \sum_{j=1}^{D} \lambda_i \lambda_j y_i y_j x_i x_j \\
\text{subject to} & \quad \sum_i \lambda_i y_i
\end{align*} \] (9)

The dual formulation leads to an expansion of the weight vector in terms of the input examples [22]:

\[ w = \sum_{i=1}^{D} \lambda_i y_i \varphi(x_i) \] (11)

The dual formulation is a maximum problem that more efficient to solve than the primal formulation [23].

2.3. Feature Selection Method

In this part, the method used for feature selection is the Multiclass Support Vector Machine-Recursive Feature Elimination (MSVM-RFE). MSVM-RFE is an extension of the SVM-RFE algorithm to solve multiclass problems. After using the SVM method to find the weight parameters of each feature, the features will be sorted by MSVM-RFE to get the best rank of features.
2.3.1. Support Vector Machine Recursive Feature Selection (SVM-RFE). SVM-RFE was first introduced by Guyon in 2002 [19]. This method will make a feature ranking list based on the square of the vector weights obtained from the SVM method. Figure 2 shows the algorithm of SVM-RFE.

![Figure 2. SVM-RFE Algorithm](image)

It can be seen at Figure 2, SVM is used to get the weight vector \( w_k \) of each feature. The features are ranked based on the square of the weight vector, \( r_k = w_k^2 \). The ranking of the feature starts from features with the smallest value of \( r_k \). From this algorithm, a ranking of features will be obtained.

2.3.2. Multiclass Support Vector Machine Recursive Feature Selection (MSVM-RFE). MSVM-RFE is an extension of the SVM-RFE method from binary problems to multiclass problems. MSVM-RFE was first introduced by Zhou in 2007 [20]. The MSVM-RFE algorithm is almost the same as the SVM-RFE algorithm. The difference is, in MSVM-RFE, the algorithm run by dividing the sample into several subsamples. Based on the algorithm of MSVM-RFE in figure 3, feature selection is done in 3 steps. First, divide the data based on the One-Against-One approach, which divides the data into \( \frac{m(m-1)}{2} \) binary class pairs where \( m \) is number of class in multiclass problem. In this paper, there are six classes (\( m = 6 \)). Based on this principle, binary classes are built into 15 class pairs. Second, find the vector weights using the SVM algorithm. Third, rank the best features based on the weight vector squared.

![Figure 3. MSVM-RFE Algorithm](image)
2.4. Classification Methods
After selecting features and getting the best ranking, the next step is to do classification. The methods used in this paper are One-Against-One Multiclass Support Vector Machine (OAO-MSVM) with RBF-Kernel and Polynomial-Kernel.

2.4.1. One-Against-One Multiclass Support Vector Machine (OAO-MSVM). OAO-MSVM was proposed by Knerr, Personnaz, and Dreyfus in 1990 [25]. OAO-MSVM forms \( \frac{m(m-1)}{2} \) binary class pairs where \( m \) is the number of classes in the multiclass problem. In this paper, there are six classes (\( m = 6 \)). Based on this principle, binary classes are built into 15 class pairs. The objective function used in the OAO-MSVM method is the same as the SVM method in general, that is the primal formulation in equation (6). However, for training data from \( \{(k, j)|k = 1, 2, ..., m; j = 1, 2, ..., m; k \neq j\} \) classes, the constraints for \((x_i, y_i)\) are [23]:

\[
(w_{kj}^T \varphi(x_i) + b_{kj}) \geq 1 - \xi_{kj}, \text{ for } y_i = k, \tag{12}
\]

\[
(w_{kj}^T \varphi(x_i) + b_{kj}) \leq -1 + \xi_{kj}, \text{ for } y_i = j, \tag{13}
\]

\[
\xi_{kj} \geq 0 \tag{14}
\]

where \( w_{kj} \) is the weight vector of the pairs of class \( k \) and class \( j \), \( b_{kj} \) is the bias of the pairs of class \( k \) and class \( j \), and \( \xi_{kj} \) is the slack variable of the pairs of class \( k \) and class \( j \).

3. Experimental results
The data used in this paper is microarray data with 985 features and 248 samples. Data consists of six classes, they are:

A. BCR-ABL
B. E2A-PBX1
C. Hyperdiploid > 50 chromosomes
D. MLL
E. T-ALL
F. TEL-AML

The data is divided into 80% training data and 20% testing data. This paper uses a one-against-one approach for binary SVMs. Because the data used is a multiclass problem, the training data is divided into 15 subset binary classes: \{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{A, F\}, \{B, C\}, \{B, D\}, \{B, E\}, \{B, F\}, \{C, D\}, \{C, E\}, \{C, F\}, \{D, E\}, \{D, F\}, and \{E, F\}.

Feature selection is used to reduce computational costs and improve accuracy. With the MSVM-RFE procedure, the selected features are the 335 best features. The classification method used is Multiclass SVM with the RBF-Kernel with \( \sigma = 0.01 \) and the Polynomial-Kernel with \( d = 4 \). Figure 6 is an example of a confusion matrix that is used to evaluate the classification results for class A. Evaluation results of class B, C, D, E, and F used the form of confusion matrix similar to Figure 4.
To evaluate the results of the classification of each method, the values of accuracy, precision, recall, and F1 score are used as parameters. Table 2 shows the formulas used to calculate these parameters. Figure 5 shows the evaluation of classification results of each class without feature selection and Figure 6 shows the evaluation of classification results of each class with feature selection.

**Table 2. The formula used to evaluate**

| Parameters   | Formula                                                   |
|--------------|-----------------------------------------------------------|
| Accuracy     | \( \frac{(TN + TP)}{(FN + TP + FP + TN)} \times 100\% \)  |
| Recall       | \( \frac{TP}{(FN + TP)} \times 100\% \)                  |
| Precision    | \( \frac{TP}{(FP + TP)} \times 100\% \)                  |
| F1 score     | \( 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \) |

**Figure 4.** Confusion Matrix for Class A

**Figure 5.** Comparison between MSVM RBF-Kernel and MSVM Polynomial-Kernel of each class without feature selection
**Table 3.** Comparison of overall accuracy of each classifiers methods without feature selection

| Classifier                     | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) | Running Time (seconds) |
|-------------------------------|--------------|---------------|------------|--------------|------------------------|
| MSVM RBF-Kernel with $\sigma = 0.01$ | 89%          | 89%           | 89%        | 89%          | 2.16                   |
| MSVM Polynomial-Kernel with $d = 4$ | 93%          | 94%           | 93%        | 93%          | 1.12                   |

**Table 4.** Comparison of overall accuracy of each classifiers methods with feature selection using MSVM-RFE

| Classifier                     | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) | Running Time (seconds) |
|-------------------------------|--------------|---------------|------------|--------------|------------------------|
| MSVM RBF-Kernel with $\sigma = 0.01$ | 86%          | 86%           | 86%        | 86%          | 1.23                   |
| MSVM Polynomial-Kernel with $d = 4$ | 94%          | 96%           | 95%        | 95%          | 0.66                   |

4. Conclusion

In this paper, a one-against-one strategy is used to broaden the problem of binary SVM. With this strategy, 15 subsets contain binary pairs were formed. MSVM-RFE has been used as a feature selection method. Feature selection method increase accuracy and increase running time on MSVM Polynomial-Kernel with $d = 4$. But on the MSVM RBF-Kernel with $\sigma = 0.01$, MSVM-RFE only increase the running time. For the multiclass classification of acute lymphoblastic leukemia microarrays data, the
best method to use is the MSVM Polynomial-Kernel with \( d = 4 \) that produces overall accuracy about 94\%, precision about 96\%, recall about 95\%, F1 score about 95\%, and the running time is 0.66 seconds.

**Acknowledgement**

This research supported financially by University of Indonesia, with a DRPM UI - PITTA 2018 research grant scheme.

**References**

[1] World Health Organization Cancer (World Health Organization, Switzerland) accessed on 14 August 2019 see https://www.who.int/cancer/en/

[2] National Cancer Institute Cancer Stats Facts: Leukemia (U.S. Department of Health and Human Services, Washington, DC) accessed on 14 August 2019 see https://seer.cancer.gov/statfacts/html/leuks.html.

[3] Coebergh J W W, Reedijk A M J, Vries E, Marcos C, Jakab Z, Stelianova-Foucher E, and Kamps W A 2006 European Journal of Cancer 42 2019-2036

[4] St. Jude Children’s Research Hospitals Acute Lymphoblastic Leukemia (ALL) (St. Jude Children’s Research Hospitals, Danny Thomas Place, Memphis) accessed on 14 August 2019 see https://www.stjude.org/disease/acute-lymphoblastic-leukemia-all.html

[5] National Cancer Institute Cancer Stats Facts: Leukemia-Acute Lymphoblastic Leukemia (ALL) (U.S. Department of Health and Human Services, Washington, DC) accessed on 14 August 2019 see https://seer.cancer.gov/statfacts/html/lymph.html

[6] Tewary D, Mondal J, and Sardar S 2014 Acute Lymphoblastic Leukemia - An Overview 3 110

[7] National Comprehensive Cancer Network 2019 Acute Lymphoblastic Leukemia (National Comprehensive Cancer Network, US, 2019) accessed on 15 August 2019 see https://www.nccn.org/patients/guidelines/all/files/common/downloads/files/all.pdf

[8] Rustam Z, Yaurita F 2018 Insolvency Prediction in Insurance Companies using Support Vector Machines and Fuzzy Kernel C-Means Journal of Physics: Conference Series 1028

[9] Rustam Z, Sudarsono E, and Sarwinda D 2019 Random-Forest (RF) and Support Vector Machine (SVM) Implementation for Analysis of Gene Expression Data in Chronic Kidney Disease (CKD) IOP Conference Series: Materials Science and Engineering 546 p 052066

[10] Rustam Z and Rampisela T V 2018 International Journal of Engineering Technology (IJET) 7 6873-6877

[11] Bagasta A R, Rustam Z, Pandelaki J, and Nugroho W A 2019 Comparison of Cubic SVM with Gaussian SVM: Classification of Infarction for Detecting Ischemic Stroke IOP Conference Series: Materials Science and Engineering 546 p 052016

[12] Zahras D, Rustam Z, Sarwinda D 2019 Soft Tissue Tumor Classification using Stochastic Support Vector Machine IOP Conference Series: Materials Science and Engineering 546 p 052089

[13] Rustam Z, Pandelaki J, and Siahaan A 2019 Kernel Spherical K-Means and Support Vector Machine for Acute Sinusitis Classification IOP Conference Series: Materials Science and Engineering 546 p 052011

[14] Rustam Z, Utami D A, Hidayat R, Pandelaki J, and Nugroho W A 2018 International Journal on Advanced Science, Engineering and Information Technology 9 685-691

[15] Panca V and Rustam Z 2016 Application of Machine Learning on Brain Cancer Multiclass Classification AIP Conference Proceedings 1862 p 030133

[16] Rustam Z and Maghfirah N 2018 Correlated Based SVM-RFE as Feature Selection for Cancer Classification using Microarray Databases AIP Conference Proceedings 2023 p 02025

[17] Rachman A A, Rustam Z 2016 Cancer Classification using Fuzzy C-Means with Feature Selection Conference: 2016 12th International Conference on Mathematics, Statistics, and Their Application (ICMSA)

[18] Lestari A W, Rustam Z 2017 Normed Kernel Function-Based Fuzzy Possibilistic C-Means (NKPCM) Algorithm for High Dimensional Breast Cancer Database Classification with
Feature Selection is Based on Laplacian Score AIP Conference Proceedings 1862

[19] Guyon I, Weston, J, and Barnhill S 2002 Machine Learning 46 396
[20] Zhou X and Tuck D 2007 Bioinformatics 23 1109
[21] Cristianini N and Taylor J S 1999 An Introduction to Support Vector Machines (Cambridge University Press)
[22] Cristianini N and Taylor J S 2000 An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods (Cambridge University Press)
[23] Awad M and Khanna R 2015 Efficient Learning Machine Springer Science + Business Media New York p 39-55
[24] Cortes C and Vapnik V 1995 Machine Learning 20 273–297
[25] Knerr S, Personnaz L, and Dreyfus G 1990 Single-Layer Learning Revisited: A Stepwise Procedure for Building and Training a Neural Network. Neurocomputing: Algorithms, Architectures and Applications Springer p 71