Finding Biomarkers from a High-Dimensional Imbalanced Dataset Using the Hybrid Method of Random Undersampling and Lasso

Masithoh Yessi Rochayani[1]*, Umu Sa'adah[2], and Ani Budi Astuti[3]

[1-3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya
Jln. Veteran, Malang 65145, Indonesia
[1]yessirochayani@student.ub.ac.id; [2]u.saadah@ub.ac.id; [3]ani_budi@ub.ac.id

Received: 14th May 2020/ **Revised:** 13th August 2020/ **Accepted:** 18th August 2020

How to Cite: Rochayani, M. Y., Sa'adah, U., & Astuti, A. B. (2020). Finding Biomarkers from a High-Dimensional Imbalanced Dataset Using the Hybrid Method of Random Undersampling and Lasso. *ComTech: Computer, Mathematics and Engineering Applications, 11*(2), 75-81. https://doi.org/10.21512/comtech.v11i2.6452

Abstract **-** The research conducted undersampling and gene selection as a starting point for cancer classification in gene expression datasets with a high-dimensional and imbalanced class. It investigated whether implementing undersampling before gene selection gave better results than without implementing undersampling. The used undersampling method was Random Undersampling (RUS), and for gene selection, it was Lasso. Then, the selected genes based on theory were validated. To explore the effectiveness of applying RUS before gene selection, the researchers used two gene expression datasets. Both of the datasets consisted of two classes, 1.545 observations and 10.935 genes, but had a different imbalance ratio. The results show that the proposed gene selection methods, namely Lasso and RUS + Lasso, can produce several important biomarkers, and the obtained model has high accuracy. However, the model is complicated since it involves too many genes. It also finds that undersampling is not affected when it is implemented in a less imbalanced class. Meanwhile, when the dataset is highly imbalanced, undersampling can remove a lot of information from the majority class. Nevertheless, the effectiveness of undersampling remains unclear. Simulation studies can be carried out in the next research to investigate when undersampling should be implemented.

Keywords: biomarkers, high-dimensional imbalanced dataset, Random Undersampling (RUS), Lasso hybrid method

I. INTRODUCTION

Biomarkers are the indicators that provide essential information about the presence of disease (Taj, Rehman, & Bajwa, 2020). For example, body temperature is a biomarker of fever, and blood pressure is a biomarker of hypertension or hypotension. Doctors can diagnose several diseases using a blood test or urine test because both of them contain biomarkers. Other biomarkers are genes that have been used for cancer diagnosis.

There are so many genes in the human body, and researchers still have not found all of the human genes (Salzberg, 2018). Microarray technology is a tool to study the expression of many genes at once. Microarray gene expression data have features (genes) that very much exceeds the number of observations, which are so-called high-dimensional data. This type of data can contain hundreds of observations and tens of thousands of genes (Hastie, Tibshirani, & Wainwright, 2015). Therefore, finding biomarkers from high-dimensional gene expression data requires a particular method.

Classifying gene expression is important for studying gene characteristics in various diseases such as cancers. However, traditional classification methods cannot work well in high-dimensional data. In general, traditional classification methods require a smaller sample size than the number of variables. One of the traditional methods is logistic regression called Generalized Linear Model (GLM) by utilizing the logit function to model data with the categorical response variable. However, logistic regression will produce biased estimators using high-dimensional data (Sur & Candès, 2019).

Selecting predictor variables at the preprocessing stage is a strategy for modeling high-dimensional data. In general, there are three approaches to set variables, namely the filter, wrapper, and embedded methods. Among these approaches, embedded is more efficient in computing and does not overfit (Guyon & Elisseeff, 2003). One of the methods included in the embedded approach is regularization. Regularization uses a constraint or a penalty in optimizing the objective function of the regression model. The ability of the regularization method to shrink the regression coefficients toward zero makes this method can be used for variables selection. Least Absolute Shrinkage and Selection Operator (Lasso) proposed by Tibshirani (1996) is one of the regularization methods. The Lasso penalty function is defined by Equation (1).

*Corresponding Author

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \tag{1}$$

The $\beta_j, (j = 1, 2, \ldots, p)$ is the regression coefficient, and ρ represents the number of predictor variables. Lasso outperforms other variable selection methods, such as Classification and Regression Tree (CART) and Random Forest, in selecting true variables in psychiatric data (Lu & Petkova, 2014).

Moreover, modeling gene expression data faces high-dimensional challenges and cases of imbalanced class. Strategies for addressing imbalanced class include oversampling to increase the number of instances in the minority classes or undersampling to remove instances from the majority classes. However, for high-dimensional data with thousands of features, the use of the oversampling method, such as Synthetic Minority Over-Sampling Technique (SMOTE), will cause the data size to be much bigger and take more time to execute (Kaur & Gosain, 2018). Therefore, the undersampling method is more appropriate to use for high-dimensional data with imbalanced class. The addition of undersampling before variable selection in high-dimensional data with an imbalanced class can give better results than that without undersampling (Yin & Gai, 2015). The most widely used undersampling method for balancing binary data is Random Undersampling (RUS). It removes some observations in the majority class randomly. This simplicity makes RUS have a low computational cost, compared to Tomek-link or other undersampling methods. Hence, RUS is suitable to be applied to huge data.

Various studies have used regularization methods for gene selection in gene expression data, including Algamal and Lee (2015) with adaptive Lasso, Kang, Huo, Xin, Tian, and Yu (2019) with relaxed Lasso, Wu, Jiang, Shen, and Yang (2018) with L ½ penalty and Zhang, Wang, Sun, Zurada, and Pal (2019) with group Lasso. However, those studies do not pay attention to the imbalanced class. Classification with an imbalanced class causes the model to be more fit for the majority class. Those studies also only evaluate the goodness of the prediction model and do not validate the selected genes obtained based on oncogenomics theory. Validation based on the theory needs to be done to ascertain whether the obtained model can be justified in theory or not.

Based on the mentioned explanation, the researchers investigate whether implementing undersampling before gene selection gives different selected genes from not implementing undersampling. If that is the case, it is to see which method is better at providing the selected genes and model. The used undersampling method is RUS, and for gene selection, it is Lasso. The researchers also investigate whether the genes that appear at the early stage of the selection process are the genes for the most important biomarkers of the disease. Then, the researchers validate the selected genes based on the oncogenomics theory.

II. METHODS

Two gene expression datasets with different imbalance ratios are used to understand the effectiveness of implementing undersampling before gene selection. The first dataset is OVA_Breast, which compares the gene expression in breast tumor tissues and other tumor tissues (colon, endometrium, kidney, lung, omentum, ovary, prostate, and uterus). The second dataset is OVA_Ovary, which compares the gene expression in ovarian tumor tissues and other tumor tissues. Another reason for using these two datasets is that both cancers are the deadliest cancer for women.

These datasets are downloaded from openml.org. These datasets consist of 1.545 observations and 10.935 genes as the predictors. In the OVA_Breast dataset, the classes are labeled by “Breast”, representing the class of breast tumor tissues, and “Other” for other tumor tissues. Meanwhile, in the OVA_Ovary dataset, the classes are labeled “Ovary” and “Other”.

Next, the researchers apply RUS in the majority class to balance the two classes. The researchers conduct gene selection using Lasso in original data and the reduced data, which have been standardized. Since these datasets have a binary response, the used model is binary logistic regression. Binary logistic regression is a logistic regression model that the response variable has two categories denoted by 1 for “success” and 0 for “failure”. The general model of binary logistic regression with ρ predictor variables is defined by Equation (2).

$$\pi(\boldsymbol{x}_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)} \tag{2}$$

Then, $\pi(x_i)$ is the probability of success, β_0 is the intercept, β_j is the logistic regression coefficient, and x_{ij} is the j^{th} predictor variable. Using logit transformation, Equation (2) can be stated as a linear form. It is shown in Equation (3).

$$\begin{aligned} logit\left(\pi(\boldsymbol{x}_i)\right) &= \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \\ &= \beta_0 + \boldsymbol{x}_i{}^T\boldsymbol{\beta} \end{aligned} \tag{3}$$

Logistic regression parameters are estimated using Maximum Likelihood Estimation (MLE). For the binary logistic regression model, the log-likelihood function is in Equation (4).

$$\begin{aligned} l(\beta_0, \boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} y_i\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \\ &+ \ln\left(1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}\right) \end{aligned} \tag{4}$$

In Equation (4), multiplying the log-likelihood function by $\frac{1}{n}$ is intended so that the number of samples does not affect the estimation results. Then, maximizing the log-likelihood function is equivalent to minimizing the negative log-likelihood. The negative log-likelihood is the objective function of logistic regression. The negative log-likelihood of binary logistic regression is defined by Equation (5).

$$\begin{aligned} -l(\beta_0, \boldsymbol{\beta}) &= -\frac{1}{n}\sum_{i=1}^{n} y_i\left(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \\ &+ \ln\left(1 + e^{\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}}\right) \end{aligned} \tag{5}$$

Therefore, the estimated parameter of binary logistic regression $(\hat{\beta})$ can be expressed by Equation (6).

$$\hat{\beta} = \arg\min -l(\beta_0, \beta) \tag{6}$$

The solution to the optimization problem in Equation (6) is obtained by setting the partial derivative of the objective function of $-l(\beta_0, \beta)$, which is equal to zero. However, this function cannot be derived exactly, so a numerical approximation is needed. Before using numerical methods to estimate the coefficients, the negative log-likelihood function is expressed in quadratic approximation (second-order Taylor polynomial), expressed by Equation (7).

$$l_Q(\beta_0, \beta) = -\frac{1}{2n}\sum_{i=1}^{n} w_i\left(z_i - \beta_0 - x_i^T\beta\right)^2 + C(\tilde{\beta}_0, \tilde{\beta}) \tag{7}$$

The $z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)(1-\tilde{p}(x_i))}$ is the estimated response, $w_i = \tilde{p}(x_i)(1-\tilde{p}(x_i))$ is the j^{th} weight, $\tilde{p}(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i\tilde{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i\tilde{\boldsymbol{\beta}})}$ is evaluated at the current parameters, $C(\tilde{\beta}_0, \tilde{\beta})$ is constant, and $\tilde{\beta}$ is the current estimated parameter (Friedman, Hastie, & Tibshirani, 2010).

Coefficients of Lasso $(\hat{\beta}_{Lasso})$ are the solution to the constrained optimization problem. It can be seen in Equation (8).

$$\hat{\beta}_{Lasso} = \arg\min -l_Q(\beta_0, \beta),$$
$$\text{subject to } \|\beta\|_1 \leq t. \tag{8}$$

Using the Lagrange multiplier, the constrained optimization problem in Equation (8) is transformed into an optimization problem without constraint in Equation (9).

$$\hat{\beta} = \arg\min\{-l_Q(\beta_0, \beta) + \lambda\|\beta\|_1\} \tag{9}$$

The λ > 0 is the regularization parameter. A framework called pathwise coordinate descent is proposed to obtain the coefficient of Lasso (Friedman, Hastie, & Holger, 2007; Friedman, Hastie, & Tibshirani, 2010; Mazumder, Friedman, & Hastie, 2011; Tibshirani *et al.,* 2012). This framework consists of three nested loops: outer loop, middle loop, and inner loop. In the outer loop, the regularization parameter λ is updated by decreasing its value. In the middle loop, the quadratic approximation in Equation (7) is updated using the current parameters. Finally, the coordinate descent algorithm is run in the inner loop to solve Equation (9).

The regularization parameter (λ) in the first iteration is the largest λ. The λ makes all regression coefficients equal to zero. As the iteration index increases, the λ decreases, and the number of nonzero coefficients increases. Since the number of nonzero coefficients of Lasso depends on the λ, the optimum λ has to be estimated. The commonly used method to estimate the optimum λ is K-fold Cross-Validation (CV). The estimated optimum λ produces the smallest average binomial deviance (Hastie *et al.,* 2015), as follows:

$$\hat{\lambda} = \underset{\lambda \in \{\lambda_1, \ldots, \lambda_m\}}{\operatorname{argmin}} \mathrm{CV}(\lambda) \tag{10}$$

Where,

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{k=1}^{K} Dev_k(\lambda) \tag{11}$$

Then, binomial deviance (*Dev*) is defined as:

$$Dev = 2\sum_i o_i \log\left(\frac{o_i}{e_i}\right) \tag{12}$$

The 0_i denotes the value of the observation, and e_i is for the estimated value of the model.

The steps of modeling using RUS + Lasso are presented in Figure 1. First, the original dataset is split into the training set and the testing set. After that, random undersampling is conducted on the majority class of the training set, and Lasso is applied to select relevant genes. The best regularization parameter (λ) is estimated using cross-validation. The best λ has the smallest binomial deviance in cross-validation. The model is then generated from the best λ. Finally, validation is conducted using the testing set.

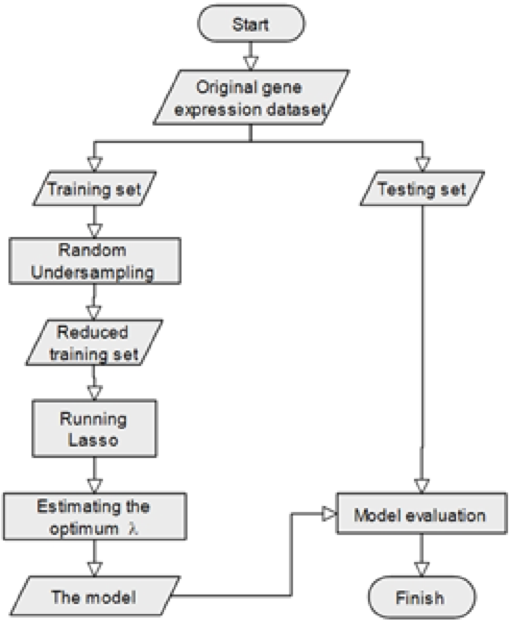


Figure 1 Flowchart of Gene Selection and Modeling Using RUS + Lasso

III. RESULTS AND DISCUSSIONS

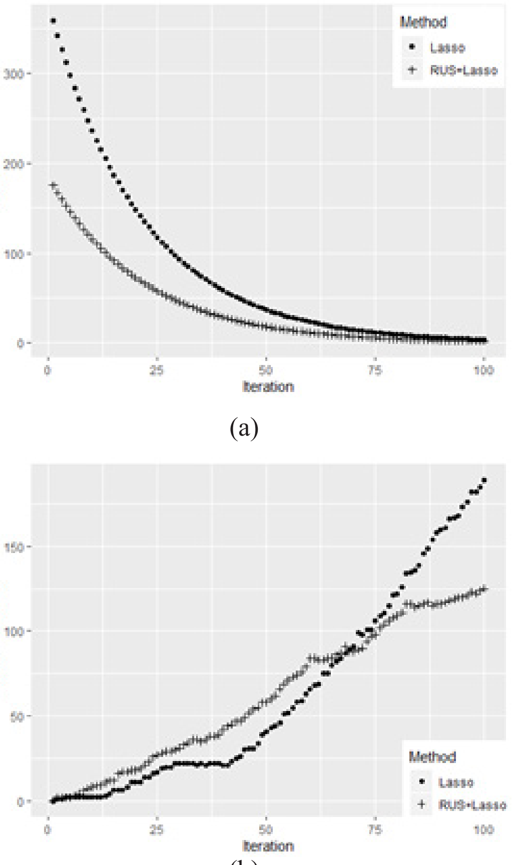
The researchers divide the datasets into a training set and a testing set with a ratio of 80%:20%. Gene selection is conducted only in training data. Therefore, the number of observations for gene selection is 1.236. In the OVA_Breast dataset, there are 274 observations in the “Breast” class and 962 observations in the “Other” class. Meanwhile, for the OVA_Ovary dataset, there are 163 observations of the “Ovary” class and 1.073 observations of the “Other” class. Undersampling is performed using RUS to balance the two classes. The ratio of 274:274 is obtained in the OVA_Breast dataset and 163:163 in the OVA_Ovary dataset.

In the OVA_Breast dataset, gene selection using Lasso is conducted on the original training set and the reduced training set. The glmnet package on R (Friedman *et al.,* 2010) limits the smallest λ or λ in the last iteration to be 0,01 times the initial λ. Then, the iteration index is denoted by *m*, in which it is *m = 1, 2, ..., M*, and $\lambda_M = 0{,}01 \times \lambda_0$. The ratio that decreases λ can be expressed by Equation (13).

$$r = \sqrt[M]{\frac{\lambda_M}{\lambda_0}} \qquad (13)$$

The λ in the m^{th} iteration is $\lambda_m = \lambda_0 r^m$. Since *r* is a positive real number less than 1, the λ decreases as the iteration index (*m*) increases. In both Lasso and RUS + Lasso methods, 100 iterations are used. *Since* $\lambda_{100} = \lambda_0 r^{100}$ and $\lambda_M = 0{,}01 \times \lambda_0$, it is = 0,955.

Figure 2 presents the plot of the regularization parameter and the number of nonzero coefficients for each iteration. In Figure 2(a), the plotted λ value is λ of glmnet output multiplied by *n*, where *n* is the number of observations. It is done because glmnet is computed by multiplying the negative log-likelihood term by $\frac{1}{n}$, but the penalty term is not multiplied. Therefore, to get the actual λ, the λ from glmnet output is multiplied by *n.* Multiplying λ glmnet by *n* causes the actual λ from the Lasso method to be greater than λ from the RUS + Lasso method because the sample is reduced after applying RUS.



(a)

(b)

Figure 2 The Plot of Regularization Parameter and the Number of Nonzero Coefficients for Each Iteration in OVA_Breast Dataset

Figure 2 The Plot of Regularization Parameter and the Number of Nonzero Coefficients for Each Iteration in OVA_Breast Dataset

In Figure 2(a), the λ decreases following the geometric sequence with the exponential decay curve. Meanwhile, Figure 2(b) shows that the number of selected genes increases as the index of iteration increases. From the two pictures, the number of selected genes increases as the index of iteration increases.

Next, the researchers investigate which genes are selected at the beginning of the iteration. In the Lasso method, the first selected gene is 209604_s_at, which is selected in the second iteration. In the fourth iteration, the gene 218502_s_at is selected. The third selected gene, namely 210239_at, starts to appear on the thirteenth iteration. On the other hand, the application of RUS + Lasso produces two selected genes in the second iteration. Those genes are 209604_s_at and 218502_s_at. The third gene is selected in the sixth iteration, which is 210239_at.

The first three genes of RUS + Lasso are the same as Lasso. It may happen because the dataset is not highly imbalanced. The ratio of class “Breast” and class “Other” is 274:962. The number of observations in the majority class is about 3,5 times the number of observations in the minority class, which is not too highly imbalanced.

Based on Figure 1, the number of selected genes depends on the regularization parameter (λ). Therefore, to obtain the best model, the optimum λ has to be estimated. The 10-fold CV results show that the optimum regularization parameter of Lasso is λ = 11,48 that produced 106 genes. Meanwhile, the optimum λ of the RUS + Lasso is λ = 5,35, producing 102 genes. To explore which method gives a better model, the researchers perform a model evaluation using the training set and the testing set. The result of the model evaluation is shown in Table 1.

Table 1 The Accuracy of the Proposed Methods on the OVA_Breast Dataset

Method	**Accuracy**	
	Training	**Testing**
Lasso	98,71%	97,09%
RUS + Lasso	98,91%	96,44%

In Table 1, RUS + Lasso has slightly higher accuracy than Lasso on the training set. However, on the testing set, the accuracy of Lasso is slightly higher. The slight difference in accuracy indicates that the two methods produce a model that can predict new data well.

Before validating the selected based on theory, the genes of probeset ID are converted to gene symbols. Since there are so many genes that are produced at optimum λ, the researchers only validate the first three selected genes. In the OVA_Breast dataset, Lasso and RUS + Lasso have the same first three selected genes: 209604_s_at, 218502_s_at, and 210239_at. The results of the conversion of these probeset ID to gene symbols are in Table 2.

GATA binding protein 3 (GATA3), which is the first gene selected, has been widely studied by scientists, especially in breast cancer. It is very useful as a marker for metastatic breast carcinoma (Cimino-mathews *et al.*, 2013) and a relatively high sensitive marker for breast carcinomas (Shaoxian *et al*., 2017). A high level of GATA3 expression indicates a slow rate of cell proliferation and predicts better survival in breast cancer patients. As the tumor grade increases, the expression of GATA3 decreases (Shaoxian *et al.*, 2017). A low level of GATA3 expression is associated with poor prognosis in breast cancer patients (Liu, Shi, Wilkerson, & Lin, 2012).

Table 2 The First Three Selected Genes in OVA_Breast Dataset

Probeset ID	Gene symbol
209604_s_at	GATA3
218502_s_at	TRPS1
210239_at	IRX5

Trichorhinophalangeal syndrome 1 (TRPS1) is also widely studied by scientists. The decrease in expression of TRPS1 can prevent mitosis of cancer cells, thereby reducing cancer growth (Witwicki *et al.*, 2018). Thus, the strong expression of TRPS1 can be used as a good prognostic marker in breast cancer. Moreover, Iroquois homeobox 5 (IRX5) is the target of therapy in several cancers including breast cancer (Myrthue *et al.*, 2008).

In the OVA_Ovary dataset, the researchers also use 100 iterations, so the ratio of λ is equal to 0,955. Figure 3 shows the plot of the regularization parameter and the number of nonzero coefficients for each iteration in the OVA_Ovary dataset. Then, Figure 3(a) presents the plot of lambda for each iteration. It can be seen that a decrease in the regularization parameter forms an exponential decay curve.

The researchers also investigate the selected genes in the initial iterations. In the OVA_Ovary dataset, Lasso and RUS+Lasso provide different selected genes. In the second iteration of Lasso, two selected genes appear. Those are 209569_x_at and 219873_at. Then, the third gene is selected in the fourth iteration, which is 206067_s_at. Meanwhile, in RUS + Lasso, genes with probeset ID of 1556051_a_at and 204069_at are selected in the second iteration. Then, in the fifth iteration, two genes are selected again. Those genes are 209569_x_at and 209678_s_at.

The differences in the result of the selected genes in the OVA_Ovary dataset may occur due to too many samples in the majority class being removed. In this dataset, the majority class has 1.073 observations, while the minority class has 163 observations. In other words, the number of observations in the majority class is about 6,6 times the number of observations in the minority class. When class balancing is performed, in the majority class, 910 observations are removed. It causes a lot of information to be discarded.

Since the selected genes at the beginning of the iteration are different between Lasso and RUS + Lasso, it is very interesting to see which method provides the best model at optimum λ. The results of 10-fold CV show that the optimum λ of Lasso is λ = 20,91 producing 71 genes, and RUS + Lasso is λ = 6,43 with 97 genes. Table 3 presents the accuracy of the model obtained from Lasso and RUS + Lasso.

Based on Table 3, the accuracy of the model produced by RUS+Lasso is higher than Lasso in the training set. Meanwhile, Lasso in the testing set has higher accuracy. Nevertheless, the accuracy of the two methods is still high.

Both OVA_Breast and OVA_Ovary data show that Lasso is slightly better than RUS + Lasso for the testing set. It suggests that the use of undersampling is less effective for these two data sets. Dal Pozzolo, Caelen, and Bontempi (2015) stated that the factors affecting the effectiveness of undersampling are the degree of imbalance and separation of the two classes. The most effective condition for undersampling occurs when the two classes are not too imbalanced, and the class conditions are not well separated.

Furthermore, the researchers perform theoretical validation on the selected genes at the beginning of iterations of the OVA_Ovary dataset. Table 4 presents the first three selected genes by Lasso in the OVA_Ovary dataset. Meanwhile, Table 5 shows the first four selected genes by RUS + Lasso in the OVA_Ovary dataset since the third and fourth genes appear in the same iteration index. It also displays the results of the conversion of probeset ID to gene symbol.

Table 3 The Accuracy of the Proposed Methods on the OVA_Ovary Dataset

Method	Accuracy	
	Training	Testing
Lasso	95,06%	92,88%
RUS + Lasso	98,77%	89,64%

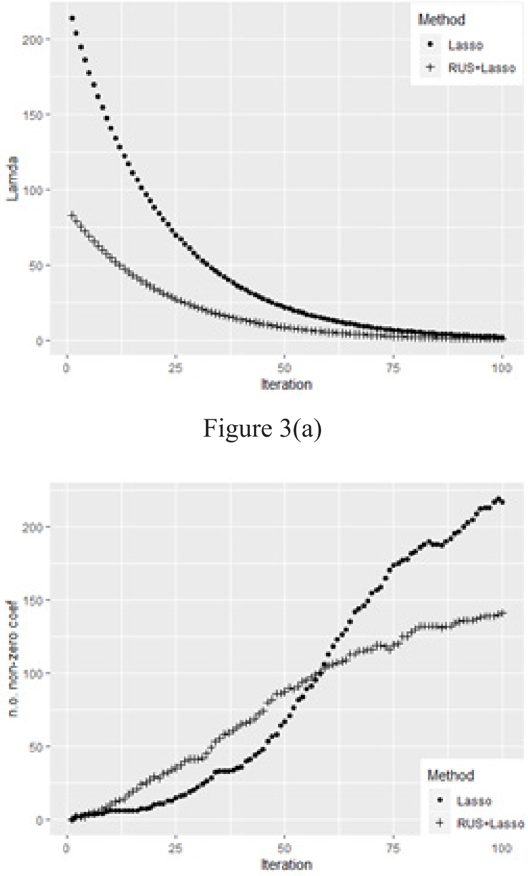


Figure 3(a)

Figure 3(b)

Figure 3 The Plot of Regularization Parameter and the Number of Nonzero Coefficients for Each Iteration in OVA_Ovary Dataset

Table 4 The First Three Genes Selected
by Lasso in the OVA_Ovary Dataset

Probeset ID	Gene symbol
209569_x_at	NSG1
219873_at	COLEC11
206067_s_at	WT1

Table 5 The First Four Genes Selected
by RUS + Lasso in the OVA_Ovary Dataset

Probeset ID	Gene symbol
1556051_a_at	BICD1
204069_at	MEIS1
209569_x_at	NSG1
209678_s_at	PRKCI

The results of gene selection in the OVA_Ovary dataset are surprising. Several selected genes in this dataset are not widely studied by researchers, including neuronal vesicle trafficking associated 1 (NSG1), collectin subfamily member 11 (COLEC11), and BICD cargo adaptor 1 (BICD1). Meanwhile, the role of Wilms' tumor 1 (WT1), Meis homeobox 1 (MEIS1), and (PRKCI) in ovarian cancer has been studied previously.

The expressions of WT1 are analyzed by Liu *et al.* (2014). The research aims to find out the correlation between WT1 expression levels and clinical features in ovarian cancer. The higher the expression of WT1 is, the higher the cancer grade will be. From the research, it is revealed that high levels of WT1 expression in ovarian cancer are associated with aggressive clinical features.

Then, MEIS1 plays a role in ovarian carcinogenesis (Crijns *et al.,* 2007). MEIS1 has a high expression in ovarian tumors compared to normal ovarian surface epithelium and other tumor types. Meanwhile, the expression of PRKCI in some subtypes of ovarian cancer is studied by Tsang, Wei, Itamochi, Tambouret, and Birrer (2017) and Sarkar *et al.* (2017). PRKCI is one of the most overexpressed genes in clear cell ovarian cancer (a subtype of ovarian cancer), and its expression influences cancer cell proliferation (Tsang *et al*., 2017). PRKCI also has a high expression in serous ovarian carcinoma (another subtype of ovarian cancer) (Sarkar *et al.*, 2017).

IV. CONCLUSIONS

In the OVA_Breast data, the proposed methods (Lasso and RUS + Lasso) can produce selected genes that become important breast cancer biomarkers. Those are GATA3, TRPS1, and IRX5. However, in the OVA_Ovary data, several genes have not been widely studied for their role in ovarian cancer. The genes are NSG1, COLEC11, and BICD1. Therefore, researchers in oncogenomics can further explore the role of NSG1, COLEC11, and BICD1 in ovarian cancer.

The model obtained in both Lasso and RUS + Lasso methods has high accuracy. However, the model is complicated because it involves many predictors (genes) and allows predictors to have insignificant effects. In the OVA_Breast data, it obtains 106 genes from Lasso and 102 genes from RUS + Lasso at the optimum regularization parameter. Meanwhile, the OVA_Ovary data get 71 genes from Lasso and 97 genes from RUS + Lasso.

Although the proposed method can be used to find biomarkers, it cannot produce a model that is easy to interpret. The obtained model cannot be used to explore the characteristics of genes in disease, whether the expression of these genes tends to be high or low. Therefore, the decision tree can be used for the next stage of modeling to obtain a model that can be interpreted easily.

From the research, the researchers also find that undersampling does not affect when it is implemented in a less imbalanced class. Meanwhile, when the dataset is highly imbalanced, undersampling can remove a lot of information from the majority class. Nevertheless, the effectiveness of undersampling remains unclear, when to use it and when not to use it. Therefore, simulation studies can be carried out in the next research to learn when undersampling should be implemented on high-dimensional data with imbalanced classes.

ACKNOWLEDGEMENT

We express our deepest gratitude to Prof. Widodo, S.Si., M.Si., Ph.D. Med.Sc., Dean of the Faculty of Mathematics and Natural Sciences Universitas Brawijaya and the former Head of the Central Laboratory of Biological Sciences Universitas Brawijaya (LSIH-UB), who has provided knowledge about genes to us.

REFERENCES

Algamal, Z. Y., & Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, *42*(23), 9326-9332. https://doi.org/10.1016/j.eswa.2015.08.016

Cimino-Mathews, A., Subhawong, A. P., Illei, P. B., Sharma, R., Halushka, M. K., Vang, R., ... & Argani, P. (2013). GATA3 expression in breast carcinoma: Utility in triple-negative, sarcomatoid, and metastatic carcinomas. *Human Pathology, 44*(7), 1341-1349. https://doi.org/10.1016/j.humpath.2012.11.003

Crijns, A. P. G., De Graeff, P., Geerts, D., Ten Hoor, K. A., Hollema, H., Van Der Sluis, T., ... & De Vries, E. G. E. (2007). MEIS and PBX homeobox proteins in ovarian cancer. *European Journal of Cancer*, *43*(17), 2495-2505. https://doi.org/10.1016/j.ejca.2007.08.025

Dal Pozzolo, A., Caelen, O., & Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks? In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 200-215). Springer. https://doi.org/10.1007/978-3-319-23528-8_13

Friedman, J., Hastie, T., & Holger, H. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*(2), 302-332. https://doi.org/10.1214/07-AOAS131

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157-1182.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. New York: Chapman and Hall/CRC.

Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of Theoretical Biology, 463,* 77-91. https://doi.org/10.1016/j.jtbi.2018.12.010

Kaur, P., & Gosain, A. (2018). Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. In *ICT Based Innovations* (pp. 23-30). Springer. https://doi.org/10.1007/978-981-10-6602-3_3

Liu, H., Shi, J., Wilkerson, M. L., & Lin, F. (2012). Immunohistochemical evaluation of GATA3 expression in tumors and normal tissues: A useful immunomarker for breast and urothelial carcinomas. *American Journal of Clinical Pathology, 138*(1), 57-64. https://doi.org/10.1309/AJCP5UAFMSA9ZQBZ

Liu, Z., Yamanouchi, K., Ohtao, T., Matsumura, S., Seino, M., Shridhar, V., ... & Kurachi, H. (2014). High levels of Wilms' tumor 1 (WT1) expression were associated with aggressive clinical features in ovarian cancer. *Anticancer Research*, *34*(5), 2331-2340.

Lu, F., & Petkova, E. (2014). A comparative study of variable selection methods in the context of developing psychiatric screening instruments. *Statistics in Medicine*, *33*(3), 401-421. https://doi.org/10.1002/sim.5937

Mazumder, R., Friedman, J. H., & Hastie, T. (2011). SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, *106*(495), 1125-1138. https://doi.org/10.1198/jasa.2011.tm09738

Myrthue, A., Rademacher, B. L. S., Pittsenbarger, J., Kutyba-Brooks, B., Gantner, M., Qian, D. Z., & Beer, T. M. (2008). The iroquois homeobox gene 5 is regulated by 1,25-dihydroxyvitamin D_3 in human prostate cancer and regulates apoptosis and the cell cycle in LNCaP prostate cancer cells. *Clinical Cancer Research, 14*(11), 3562-3570. https://doi.org/10.1158/1078-0432.CCR-07-4649

Salzberg, S. L. (2018). Open questions How many genes do we have ? *BMC Biology*, *16*(94), 1-3. https://doi.org/https://doi.org/10.1186/s12915-018-0564-x

Sarkar, S., Bristow, C. A., Dey, P., Rai, K., Perets, R., Ramirez-Cardenas, A., ... & McGuire, M. (2017). PRKCI promotes immune suppression in ovarian cancer. *Genes & Development*, *31*(11), 1109-1121. https://doi.org/10.1101/gad.296640.117

Shaoxian, T., Baohua, Y., Xiaoli, X., Yufan, C., Xiaoyu, T., Hongfen, L., ... & Wentao, Y. (2017). Characterisation of GATA3 expression in invasive breast cancer: Differences in histological subtypes and immunohistochemically defined molecular subtypes. *Journal of Clinical Pathology*, *70*(11), 926-934. https://doi.org/10.1136/jclinpath-2016-204137

Sur, P., & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences, 116*(29), 14516-14525. https://doi.org/10.1073/pnas.1810420116

Taj, A., Rehman, A., & Bajwa, S. Z. (2020). Biomarkers and their role in detection of biomolecules. In A. Wu & W. S. Khan (Eds.), *Nanobiosensors: From design to applications* (pp. 73-94). Germany: Wiley-VCH.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., & Tibshirani, R. J. (2012). Strong rules for discarding predictors in Lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(2), 245-266. https://doi.org/10.1111/j.1467-9868.2011.01004.x

Tsang, T. Y., Wei, W., Itamochi, H., Tambouret, R., & Birrer, M. J. (2017). Integrated genomic analysis of clear cell ovarian cancers identified PRKCI as a potential therapeutic target. *Oncotarget*, *8*(57), 96482-96495.

Witwicki, R. M., Ekram, M. B., Qiu, X., Janiszewska, M., Shu, S., Kwon, M., ... & Yu, K. (2018). TRPS1 is a lineage-specific transcriptional dependency in breast cancer. *Cell Reports*, *25*(5), 1255-1267. https://doi.org/10.1016/j.celrep.2018.10.023

Wu, S., Jiang, H., Shen, H., & Yang, Z. (2018). Gene selection in cancer classification using sparse logistic regression with $L_{1/2}$ regularization. *Applied Sciences*, *8*(9), 1-12. https://doi.org/10.3390/app8091569

Yin, H., & Gai, K. (2015). An empirical study on preprocessing high-dimensional class-imbalanced data for classification. In *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems* (pp. 1314-1319). IEEE. https://doi.org/10.1109/HPCC-CSS-ICESS.2015.205

Zhang, H., Wang, J., Sun, Z., Zurada, J. M., & Pal, N. R. (2019). Feature selection for neural networks using group Lasso regularization. *IEEE Transactions on Knowledge and Data Engineering*, *32*(4), 659-673. https://doi.org/10.1109/TKDE.2019.2893266