ICIC Express Letters
Volume **13**, Number **8**, August **2019**

ICIC International ©2019 ISSN 1881-803X
pp. 697–703

ENSEMBLE BASED MACHINE LEARNING FOR OPTIMIZING TOXICITY IN CANCER DRUG DISCOVERY

Heri Kuswanto[1,*], Erlin Sukmaputri[1] and Hayato Ohwada[2]

[1]Department of Statistics
Faculty of Mathematics, Computing and Data Science
Institut Teknologi Sepuluh Nopember (ITS)
Surabaya, East Java 60111, Indonesia
*Corresponding author: kuswanto.its@gmail.com

[2]Department of Industrial Administration
Faculty of Science and Technology
Tokyo University of Science
2641 Yamazaki, Noda-shi, Chiba-ken 278-8510, Japan
ohwada@rs.tus.ac.jp

Received January 2019; accepted April 2019

Abstract. *Cancer is a disease caused by abnormal growth due to the cells of the body's tissues that turn into cancer cells. Radiotherapy is one of the cancer treatments that has a side effect of killing normal cells around cancer cells. Radioprotector is made to reduce normal cell death and increase cancer cell death. This research identifies the compounds corresponding to the toxicity with normal cell death rate below and above 20%. The data used in this study is the level of toxicity to classify compounds for radioprotector consisting of 84 compounds with 217 predictors (features). Two ensemble based machine learning approaches are applied to overcoming the problem of high dimensionality of the data, namely Logistic Regression Ensembles (LORENS) and Ensemble Support Vector Machine (AdaBoost-SVM). The AdaBoost-SVM is applied to the important features selected by Mean Decreasing Gini (MDG) index. The results showed that the AdaBoost-SVM outperforms LORENS significantly. The accuracy is 0.7889 obtained by examining 5% of most important features.*

Keywords: Cancer, High dimensionality, Compound, Toxicity, LORENS, AdaBoost-SVM

1. Introduction.

Cancer is characterized by uncontrolled cell growth and the ability of these cells to attack other biological tissues, either by direct growth in adjacent tissues (invasion) or by migration of cells to distant (metastatic) sites in the body [1]. One common cancer treatment is radiotherapy, i.e., a radiation therapy treatment for cancer using radiation such as gamma rays, x-rays or high-energy electrons. The way radiotherapy works is to give radiation doses that turn off the tumor in a predetermined area (target volume) while the surrounding normal tissue gets the minimum dose. Treatment using radiotherapy has significant side effects, namely killing normal cells around cancer cells. This is because radiotherapy damages DNA in cancer cells and makes DNA stimulate p53 for apoptosis (cell death). p53 is a tumor suppressor gene that acts to stop tumor development. When cells are exposed to radiotherapy, DNA in cancer cells is damaged and results in apoptosis occurring in normal cells and killing these cells. In order to overcome the effects of radiotherapy, Ariyasu et al. [2] designed radioprotector or radiation protection by looking for components of compounds related to p53 protein. The design involves forming 84 compounds that were thought to be good for the radioprotector. Furthermore, two experiments have been conducted. In the first experiment, the compound

DOI: 10.24507/icicel.13.08.697

was given to normal cells to measure toxicity. The second experiment was carried out by giving compounds to cells that had been exposed to gamma radiation (10 Gy) to measure the radiation protection function. The indicator used in both trials is the cell death rate. The synthetic compounds needed for radioprotector are those that have low cell death rates on toxicity and high cell death rates in the radiation protection function.

This study carries out an analysis on the level of toxicity where toxicity is the ability of a molecule or chemical compound that can cause damage to certain parts of living things [3]. Using the dataset of [2], a study by [4] predicted radiation protection and toxicity using Random Forest (RF) and Support Vector Machine (SVM). The results showed that the Random Forest is better used to predict toxicity while SVM is used to predict radiation protection. Another study conducted by [5] compared the results of the classification accuracy of compounds for optimization of radiation protection and toxicity using Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGB), and k-Nearest Neighbor (kNN). All of these approaches use selected features as the basis of classification. The study found that using 10% of the most important features led to the optimal accuracy except the XGB that was optimal under 15% most important features.

The data produced by [2] is a high dimensional data where the number of features is larger than the number of observations (compounds). Both studies carried out feature selection in handling high dimensional data using feature importance. According to [6] the challenge in the case of high dimensional data is poor accuracy due to the phenomenon of curse of dimensional and overfitting models on training data. There are two approaches to overcoming the challenges of high dimensional data, namely reducing the dimensions of the dataset or by applying methods that are independent of dimensional data. The common way is to do a feature selection on variables or use ensemble-based classifications.

This study investigates the performance of ensemble based machine learning approaches to overcome the high dimensional problem on the compounds of radioprotector dataset. This study is different with the previous studies on the way to treat the feature during the classification process, and is expected to improve the prediction accuracy. One of the methods applied is LORENS. Lim et al. [7] argued that LORENS is also able to handle unbalanced response variables and improve accuracy, sensitivity, and specificity compared to other classification methods. LORENS has been widely applied in several previous studies, e.g., by [8], where LORENS was used in classifying consumer defection cases with very large sample sizes. Furthermore, [9] conducted a classification on gene expression in Alzheimer's disease by comparing LORENS with the Naive Bayes method. It is found that LORENS outperforms Naive Bayes. Both [8] and [9] found that LORENS performed well and outperformed the competing methods.

This research also applies ensemble of Support Vector Machine (SVM). The SVM method finds the best separator function or hyperplane that separates two classes in the input space. Over the past few years, SVM has been applied to the problem of high dimensional data, e.g., remote sensing classification, web documents and microarray analysis [6]. The ensemble method on SVM uses Adaptive Boosting (AdaBoost). The AdaBoost uses SVM as a base classifier for classification, and it has been applied in many previous studies such as by [10] which discussed the classification of remote sensing with high-dimensional data using SVM, ANN, and Maximum Likelihood. The study showed that SVM outperforms the others. Research by [11] discussed the prediction of gene expression of several types of cancer and ensemble SVM has been proven to have optimal accuracy compared to single SVM and kNN. In addition, the study of [12] showed that ensemble SVM is a good approach to predict breast cancer on small and large scale data. Other studies that applied Ensemble SVM (AdaBoost-SVM) on the case of high dimensional data are [13-15] among others. Based on the above description, this study uses LORENS and Ensemble of SVM methods in predicting the level of toxicity for cancer drug design. The classified compounds with low toxicity can be recommended for radioprotector.

2. **Research Methodology.** The data analyzed in this study is a secondary data produced by [2], namely data on normal cell protective compounds on radioprotector. The dataset comprises of 84 compounds and 217 compound components (hereafter we called as features). There are two classes of toxicity level, i.e., low and high toxicity calculated based on death rate cells. The classification is based on cell death rates where class 0 indicates toxicity with cell death rates less than 20% and class 1 for toxicity with cell death rates of above 20%. Table 1 below listed the variables as well as the features involved in the analysis.

Table 1. Variables and features

Variable	Name of feature
Response (Y)	Class target $Y(0)$ = Toxicity with normal cell death rate 20-100% $Y(1)$ = Toxicity with normal cell death rate < 20%
Predictor (x_i)	x_1 = pKa x_2 = Br_Count x_3 = C_Count x_4 = Cl_Count $\vdots$ x_{214} = Molecular_3D_PolarSASA x_{215} = Molecular_3D_SASA x_{216} = Molecular_3D_SAVoL x_{217} = Molecular_Volume

The analysis using LORENS involves the following steps: partitioning the predictors (features) into some subspaces, determination of the threshold as well as number of ensemble. In this case, two different thresholds will be examined, i.e., optimal threshold and fixed threshold 0.5. Furthermore, the probability of each compound with its corresponding predictors is calculated from the logistic regression model in order to assign the compound into the response class. Finally, majority voting is applied to determining the class. Different with LORENS which uses all dataset in the classification, AdaBoost SVM requires feature selection prior to the analysis. The important features are selected with Mean Decreasing Gini (MDG) Index. The AdaBoost-SVM will be run with four different proportion of most important features, i.e., top 5%, 10%, 25% and 35% important features. The AdaBoost-SVM requires searching Cost (C) and Gamma parameters that led to optimal accuracy. Grid search procedure will be applied to finding the optimal range of those two parameters, and then the accuracy can be calculated.

3. **Results and Discussion.** This study discusses the results of compound classification used to determine the compounds that are good for protecting normal cells on the radioprotector based on the level of toxicity. The classification is done by using LORENS and Ensemble SVM.

3.1. **Classification using LORENS.** LORENS is a computational approach to solve classification problems. In order to get the best classification results, LORENS is run several times with a number of different partitions. In this case, we set 5, 7, 10, 12, 15, 17, 20, 22, 25, 30, 40, and 50 partitions. LORENS uses an optimal threshold as the basis of classification. This research compares the performance of LORENS using optimal threshold and pre-determined threshold 0.5. The number of commonly used ensemble is 10; however, this study uses 11 ensembles in order to avoid confusion due to the vote equality during the majority voting. In the analysis using LORENS, 10-folds cross validation is used, which means that the data will be divided into 10 parts (folds) with the same number. Each part will be treated as testing data and training data. It is started with using the data in the 1st fold as testing data, while the 2nd to 10th folds are used as training data. Furthermore, data in the second fold is used as testing data and the rest of the fold is used as training data, and so forth until the 10th fold. There are 27830 logistic regression models formed under the setting of 11 ensembles, 10 folds and 12 partitions.

The first step in carrying out LORENS analysis is to divide the data into several subspaces or partitions. Each subspace contains predictors with a fair amount. This study has 217 predictors that will be randomly assigned into subspace depending on the number of partitions. As the result, each subspace will have different logistic regression model with its corresponding probability. If the probability is greater than 0.5, then compound is classified into class 1. If the probability is less than 0.5, then the compound is assigned to class 0. Table 2 below presents the accuracy of LORENS obtained with optimal and 0.5 thresholds for different number of partitions. In fact, the values of optimal threshold for all folds are very close to 0.5. Therefore, we can expect that the performance between both threshold settings will not be significantly different.

Table 2. Accuracy of LORENS

Number of partitions	Accuracy (%) Optimal threshold	Accuracy (%) Fixed threshold 0.5
5	55.9524	59.5238
7	58.3333	60.7143
10	58.3333	61.9048
12	55.9524	64.2857
15	54.7619	58.3333
17	57.1429	59.5238
20	60.7143	65.4762
22	63.0952	65.4762
25	64.2857	61.9048
30	**69.0476***	**69.0476***
40	65.4762	67.8571
50	67.8571	66.6667

The table shows that using 30 partitions leads to optimum classification results both for 0.5 and optimum threshold. The accuracy is about 69.0476%. We see also that the accuracy is not linearly correlated with the number of partitions. The performance of classification using both thresholds is very similar, as expected.

3.2. **AdaBoost-SVM.** This subsection discusses the results of applying AdaBoost-SVM to classifying the compound for toxicity. The method is not designed for the case of high dimensionality. Therefore, the features need to be selected to make it feasible to apply the AdaBoost-SVM. The feature will be selected based on its importance using Mean Decreasing Gini (MDG) method. The result of the feature selection is the ranking of each feature where the higher the MDG value, the more important the feature is. The parameters used to calculate MDG from random forest are mtry (number of selected features) and ntree (number of trees formed). Both parameters are combined to search for the maximum accuracy. The analysis shows that the optimal accuracy is 0.6749 obtained from the combination of 5 mtry and 800 ntree. By using these parameters, the

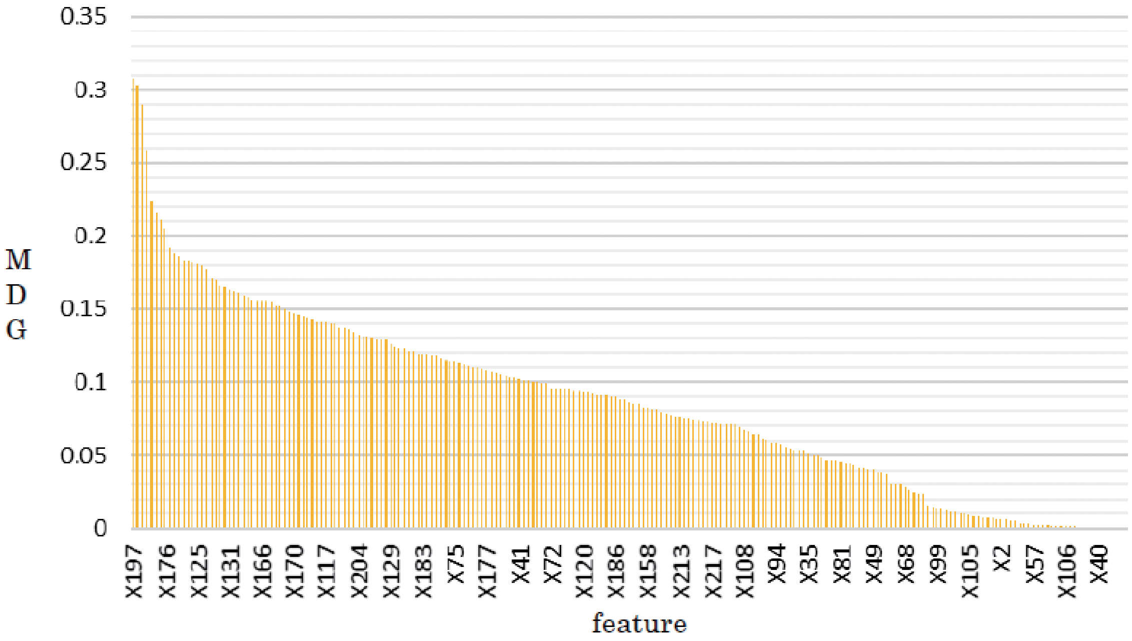

Figure 1. MDG values for each feature

MDG calculation is then performed to determine the importance of each variable. The results of the MDG calculations are shown in Figure 1.

We see that the highest MDG is 0.3075 obtained for X197. This shows that based on the MDG method, the most influential variable in determining classification is X197 (Energy), followed by X176 (Jurs-PNSA), X125 (CHI-3-C), etc. The Adaboost-SVM will be applied to the 5%, 10%, 25%, and 35% most importance features, resulting on 11, 22, 55 and 76 number of features respectively. The AdaBoost-SVM classification is carried out by specifying the parameters on a single SVM classification model, i.e., using RBF kernel, C, and Gamma.

The optimum C and Gamma values are searched using a search grid within the range 10^{-4} to 10^4. Using 10-fold cross validation, the analysis is carried out using four different iterations, i.e., 5, 10, 15, and 20 iterations. Table 3 summarizes the grid search result from the classification using the AdaBoost-SVM method on different numbers of important features.

Table 3. Grid search of SVM

C	Gamma	Average of total accuracy			
		5%	10%	25%	35%
$10^{-4} - 10^{-1}$	$10^{-4} - 10^{0}$	0.5508	0.5453	0.5442	0.5364
	$10^{1} - 10^{4}$	0.5122	0.5028	0.5028	0.509
$10^{0} - 10^{4}$	$10^{-4} - 10^{0}$	0.6783	0.6969	0.6652	0.6378
	$10^{1} - 10^{4}$	0.5217	0.5156	0.484	0.4922

Based on the values in Table 3, the optimal C and Gamma parameters are obtained within range C of $10^0 - 10^4$ and range Gamma of $10^{-4} - 10^0$, with the accuracy above 0.6. Furthermore, the AdaBoost-SVM classification is performed with parameters within those ranges as can be seen in Table 4 showing the result of analysis using 5% important features.

The highest total accuracy is 0.7889 obtained with C = 1 and Gamma = 0.1 with 10 iterations. The highest total accuracy for the other cases are 0.7861, 0.7847, and 0.75 for

Table 4. Total accuracy for the corresponding C and Gamma using 5% important features

SVM parameters		Number of iterations	Total accuracy
C	Gamma		
1	0.1	5	0.7514
1	0.1	10	0.7889
10	0.01	15	0.7639
1000	0.001	20	0.7389

10%, 25% and 35% important features respectively. However, these values are obtained under different combinations of Cost, Gamma and iteration.

3.3. **Discussion.** This subsection discusses the comparison of the classification accuracy obtained from LORENS and AdaBoost-SVM. LORENS was developed based on the idea of using all features to generate the class of the response during the classification, although the algorithm involves feature partition. Meanwhile, AdaBoost-SVM uses selected features for the classification. From the empirical reason point of view, both classification approaches have their potential benefit. LORENS uses all information in the feature with the following consideration. The drug discovery process is a complicated work and needs a very high cost. Choosing only several important features means neglecting the information within the unselected features which might be useful, as the common idea of ensemble approaches. Meanwhile, AdaBoost-SVM uses only selected (most important) features considering the fact that it saves the computational time especially if the data dimension is very high. Table 5 summarizes the comparison of classification using LORENS and AdaBoost-SVM for the toxicity case. Again, the AdaBoost-SVM was run with four different percentages of selected features, i.e., 5%, 10%, 25%, and 35%.

Table 5. Comparion of classification accuracy obtained with LORENS and AdaBoost-SVM

Method	Percentage features (%)	Accuracy (%)	Parameters of AdaBoost-SVM		
			C	Gamma	Iteration
LORENS	100%	0.6904	–	–	–
AdaBoost-SVM	5%	0.7889	1	0.1	10
	10%	0.7861	10	0.001	15
	25%	0.7847	10	0.001	10
	35%	0.7500	100	0.0001	15

Based on the table, we see that the AdaBoost-SVM outperforms LORENS by improving the accuracy about 9%, achieved by using 5% of the most important features with parameters Cost = 1, Gamma = 0.1 and 10 iterations. It is worth noting that this result does not necessarily mean that AdaBoost-SVM will always outperform LORENS as it is case dependent.

4. **Conclusions.** The analysis showed that the high dimensionality in toxicity data is better classified with AdaBoost-SVM. The accuracy reached 78% which is significantly higher than the classification using LORENS. This result is obtained by examining only 5% important features. The AdaBoost-SVM improves the accuracy obtained by other machine learning methods such as Random Forest and single SVM which reached about 71.6% accuracy, as shown in [4]. It suggests that the ensemble based machine learning methods can be used as a promising alternative to classify compounds dealing with toxicity. The selected features that are important for the toxicity case obtained from MDG are

similar with those obtained with Gini Index, where X197 (Energy) is the most important feature. The present study found that the ensemble based machine learning approach works well to optimize the toxicity in the cancer drug discovery case. Nevertheless, the experiment was intended to optimize not only the toxicity, but also the radiotheraphy. Therefore, research on improving the ensemble based machine learning for bivariate case can be a promising future research agenda.

Acknowledgment. This work is partially supported by the Ministry of Research and Higher Education – Indonesia. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] E. Meiyanto, D. M. Supardjan and D. Agustina, Efek antiproliferatif pentagamavunon, *Jurnal Kedokteran Yarsi*, vol.14, pp.11-15, 2006.

[2] S. Ariyasu, A. Sawa, A. Morita, K. Hanaya, M. Hoshi, I. Takahashi and S. Aoki, Design and synthesis of 8-hydroxyquinoline-based radioprotective agents, *Bioorganic and Medicinal Chemistry*, vol.22, no.15, pp.3891-3905, 2014.

[3] W. F. Durham, Dangerous properties of industrial materials, in *Toxicity*, N. I. Saz (edt.), Van Nostrand Reinhold Co., New York, 1975.

[4] A. Matsumoto, S. Aoki and H. Ohwada, Comparison of random forest and SVM for raw data in drug discovery: Prediction of radiation protection and toxicity case study, *International Journal of Machine Learning and Computing*, vol.6, no.2, 2016.

[5] M. Kimura, S. Aoki and H. Ohwada, Predicting radiation protection and toxicity of p53 targeting radioprotectors using machine learning, *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp.1-4, 2017.

[6] V. Pappu and P. M. Pardalos, High dimensional data classification, in *Clusters, Orders, and Trees: Methods and Applications: In Honor of Boris Mirkin's 70th Birthday*, 2014.

[7] N. Lim, H. Ahn, H. Moon and J. Chen, Classification of high-dimensional data with ensemble of logistic regression models, *Journal of Biopharmaceutical Statistics*, vol.20, pp.160-171, 2010.

[8] H. Kuswanto, A. Asfihani, Y. Sarumaha and H. Ohwada, Logistic regression ensembles for predicting customer defection with very large sample size, *Procedia Computer Science*, vol.72, pp.86-93, 2015.

[9] H. Kuswanto and R. W. Werdhana, Classification of Alzheimer related genes using LORENS with important and significant features, *Internetworking Indonesia Journal*, vol.10, no.1, pp.29-34, 2018.

[10] M. Pal and P. M. Mather, Support vector machines for classification in remote sensing, *International Journal of Remote Sensing*, pp.1007-1011, 2005.

[11] A. Dragomir and A. Bezerianos, Improving gene expression sample classification using support vector machine ensembles aggregated by boosting, *Cancer Genomics & Proteomics*, pp.63-70, 2006.

[12] M.-W. Huang, C.-W. Chen, W.-C. Lin, S.-W. Ke and C.-F. Tsai, SVM and SVM ensembles in breast cancer, *Plus ONE*, vol.12, no.1, 2017.

[13] X. Li, L. Wang and E. Sung, Adaboost with SVM-based component classifier, *Engineering Applications of Artificial Intelligence*, pp.785-795, 2006.

[14] S. M. Valoillahzadeh, A. Sayadiyan and M. Nazari, *Face Detection Using Adaboosted SVM-Based Component Classifier*, https://arxiv.org/ftp/arxiv/papers/0812/0812.2575.pdf, 2008.

[15] H. J. Morra, Z. Tu, L. G. Apostolova, A. E. Green, A. W. Toga and P. Thompson, Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation, *IEEE Trans. Medical Imaging*, vol.29, no.1, pp.30-34, 2010.