Comparation of SVM Algorithm and Neural Network With Feature Optimization Based on Genetic Algorithm in Determining Immunotherapy Success in Cancer Disease

Yudi Ramdhani, Rizki Tri Prasetio, Ryan Hidayat
Computer Science Program
Universitas Adhirajasa Reswara Sanjaya
Bandung, Indonesia 40291
yudi@ars.ac.id rizki@ars.ac.id, kazekageryan22@gmail.com

Doni Purnama Alamsyah
Entrepreneurship Department, BINUS Business School Undergraduate Program
Bina Nusantara University
Jakarta, Indonesia 11480
doni.syah@binus.ac.id

Abstract

Cancer patients are usually advised to follow a number of treatments. Such as chemotherapy, radiotherapy, and surgery. Now increasingly popular is therapy with the method of immunotherapy. Immunotherapy is the latest breakthrough in cancer treatment. This therapy uses the body's own immune system to fight cancer cells. In this study, a comparison was made between the Support Vector Machine (SVM) algorithm and the Neural Network by using the Rapid Minner application and the Immunotherapy dataset, showing the results that the Neural Network has a higher accuracy value with an average accuracy value of 79.08%, 95.52%, 95.63% in three stages while for Support Vector Machine the average accuracy value is 78.22%, 83.82%, 86.48% in three successive stages from the results which concluded that the Neural Network has a higher level of accuracy than Support Vector Machine for Immunotherapy dataset classification, using the Genetic Algorithm Optimization Feature Proven effective in increasing the accuracy of both classification algorithms.

Keywords
SVM, Neural network, classification, Immunotherapy.

1. Introduction
In the era of big data, the tremendous growth of data dimensions, many redundant and irrelevant features in data collection, resulting in not only an increase in computational costs but also a decrease in classification accuracy. Therefore, Feature Selection (FS), which aims to select a subset of relevant features to reduce overfitting and improve classification performance, has attracted great attention in data pre-processing (Zhou, 2021). Wart is a skin disease caused by the mediated Human Papillomavirus (HPV). Experts are trying to find customized and effective treatment methods for patients. Cryotherapy and Immunotherapy are the best-known and most common methods of treating common warts. Doctors need to determine which method will give better results for each patient (Asbar & Saptari, 2017). This diagnosis requires a classifier with high predictive accuracy and must also recommend some useful insights that can be interpreted to the doctor. This in turn will help medical practitioners to decide on an effective time and treatment plan for each patient (Roopal Jain, 2018).

Cancer immunotherapy provides long-lasting clinical benefits in only a minority of patients, mainly due to the lack of reliable biomarkers for accurate prediction of treatment outcome and evaluation of response (Santosh Kumar Paidi, 2020). The cost of cancer treatment is quite expensive, so it is necessary to have the right treatment technology with a high success rate for curing cancer.

2. Literature Review

A. Data Mining

The performance of data mining techniques used in predicting cancer is greatly reduced without a good combination of key features and also the use of inappropriate machine learning algorithms (AyonDey, 2016). Therefore, it is very important to identify the best combination of significant features that work very well with the best performing algorithm. This research focuses on finding data mining techniques with significant features that will perform well in predicting cancer. However, it is not easy to identify the right technique and select the significant features.

The techniques used in the prediction of cardiovascular disease are insufficient, and proper examination is needed to identify significant features and data mining techniques that will improve performance. (Jesmin Nahara, 2013) proper evaluation and comparison to test the combination of different features along with data mining techniques, remains to be focused. Thus, the need for thorough experimentation arises to provide precise identification of data mining techniques and significant features to ensure acceptable and accurate cancer predictions (AyonDey, 2016).

B. Support Vector Machine

In order to explain the structural characteristics of the model and the problem of overfitting, several new model structures have been developed over the last few decades. The support vector machine, a group of margin classifier models proposed by Vapnik and his group at AT&T Bell Laboratories in the 1990s, is an effective model type with high generalizability in practice (Cortes, 1995). Different from empirical risk minimization-based statistical learning methods, SVM aims to minimize structural risk, which shows a strong ability to avoid overfitting (Paragraph, 2005). In the SVM model, a decision hyperplane is constructed to form a split gap to divide the two class instances with the maximum margin, as shown in Figure 1.

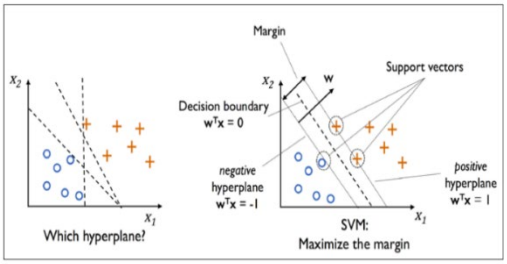

Figure 1. Hyperplane

Due to the widespread generalizability of SVM compared to conventional learning approaches, SVM has been applied to many fields. In particular, as a data-driven predictive technique, the SVM model has attracted the most attention to disease diagnosis in recent years, such as gait diagnosis of cerebral palsy, detection of gastric lymph node cancer, and diagnosis of prostate cancer (Wang H. Z., 2018).

C. Neural Network

The development of Neural Network science dates back to 1943 when Warren McCulloch and Walter Pitts introduced the first computational neural network model. They combine several simple processing units together which can provide an overall increase in computing power (Derwin Suhartono, 2012).

Detailed data analysis with the interrelationships in modern analysis is very important. Based on Bayesian statistics, a network package has been presented which is used for the classification and prediction of each occurrence of a complete process continuously or continuously (Feindt & Kerzel, 2006). Basically, simple Neural Network is categorized as Soft Computing science which represents the function of the human brain which is able to provide stimulation, process, and provide output obtained from the stimulation function and process of the human brain. Or you could say it also consists of an input layer neurons, 1 or 2 hidden layer neurons, and the final layer of output neurons (Feindt & Kerzel, 2006). Each relationship is associated or interrelated with numerical data called 'weight'. The ability of the human brain is very high and can process a lot of information, for example a child who is just learning to speak even though the data is obtained from the environment that is received by the ear to the brain through

the nerves even though he still doesn't know what kind of algorithm he uses to learn the pronunciation. This computational ability is what makes it an edge in science (Wang, 2003).

D. Adaboost

Adaptive boosting is an algorithm introduced by Freund and Schapire in 1995. This algorithm acts to strengthen the performance of weak learner algorithms so that they can create relatively strong classifiers. The normalized distribution vector is shown in Equation (M. Azka Putra, 2018).

$$\mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{i=1}^{N} w_i^t}$$

E. Rapid Miner

Rapidminer is a user interactive environment for machine learning and Data mining processes. It is an open-source, free project implemented in Java. It represents a modular approach to designing even very complex problems - the concept of modular operators which allows the design of complex nested operator chains for a large number of learning problems (Amrita Naika, 2016). RM uses XML to describe the knowledge discovery processes of operator tree modeling. RM has flexible operators for input and output of data in different file formats. It contains more than 100 learning schemes for classification, regression and clustering tasks (Amrita Naika, 2016).

F. Feature Selection on Clasification

Feature Selection techniques are categorized based on computational differences. They are classified as filter-based, wrapper-based and hybrid filter-wrapper methods. The filter-based method selects data features using statistical information from the data set (Jain, 2018). The algorithm uses a specific search strategy to search for subsets in the feature space that are judged by independent measures. Wrapper methods on the other hand rely on classifiers instead of relying on data sets and are mostly adopted in classification problems (Jain, 2018). They use a supervised learning model and aim to maximize classification accuracy. Overall performance is improved by leveraging feedback from the classifier. The hybrid method combines the advantages of filter-based and wrapper-based techniques. They use independent parameters to select the best subset for a particular cardinality and the last best subset among the best subsets across different cardinality is selected using a mining algorithm (Jain, 2018).

3. Methods

3.1 Research Methods

The research method we use is a comparison between two data mining classification methods between the Support Vector Machine (SVM) algorithm and the Neural Network algorithm, one of the algorithms that is often used in data mining. As well as using the Optimization of the sharing feature of the Genetic Algorithm to achieve the classification results. The data is taken from the UCI Machine Learning Repository and the application used by us is Rapidminner.

A. Research Design

In this study, the appropriate technique will be selected and applied to the Immunotherapy Dataset on the SVM classification and Neural network. The first stage in this research is to divide the immunotherapy dataset by dividing it into two, namely training data and test data. The next step is to select the best features in the Immunotherapy dataset using Genetic Algorithms (GA). Then the features that have been selected are classified using the Support Vector Machine and Neural Network algorithms, the best model from the classification results is compared to see which model is the best classification results and classification results will be measured with accuracy values. Further research design stages are presented in Figure 2.

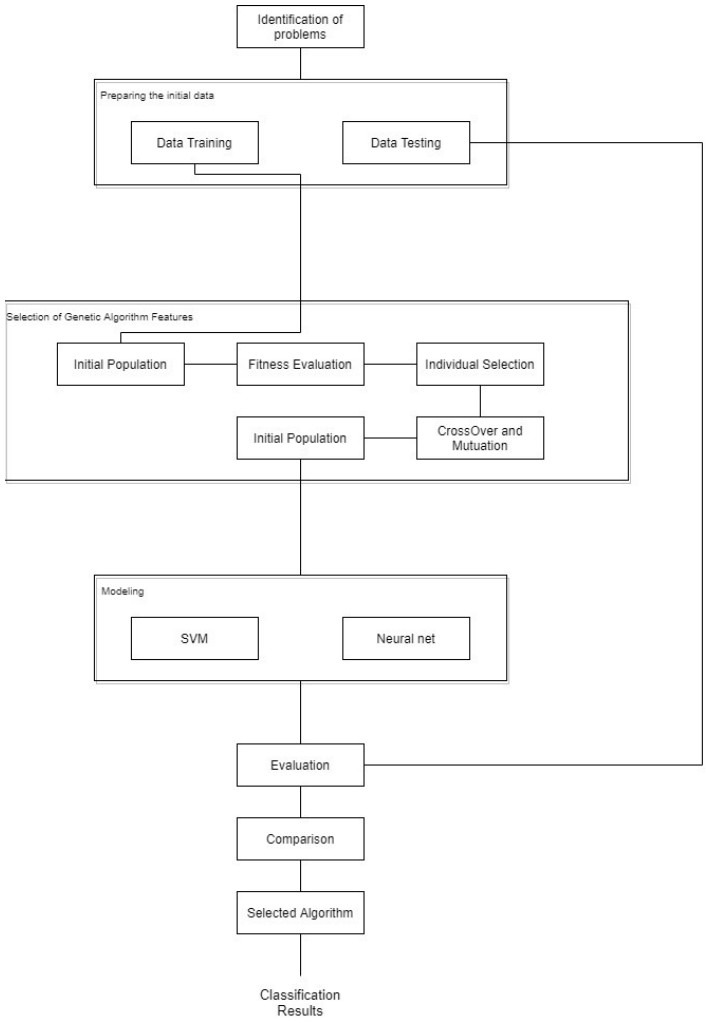

Figure 2. Research Design

B. Data Collecting

The data used for training and testing is public data which has been carefully classified by four experts from the following universities:

- Name: Fahime Khozeimeh, MD
  Institution: Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
- Name: Pouran Layegh, Professor of Dermatology
  Institution: Mashhad University of Medical Sciences, Mashhad, Iran
- Name: Roohallah Alizadehsani, PhD student
  Institution: Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Victoria 3217, Australia.
- Name: Mohamad Roshanzamir, PhD candidate
  Institution: Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran.

Immunotherapy data which has eight attributes and 90 data can be accessed at the following link: https://archive.ics.uci.edu/ml/datasets/Immunotherapy+Dataset.

4. Results and Discussion

The initial stages in the classification of the Immunotherapy Dataset will be carried out first and the classification using the Support Vector Machine and Neural network without using Adaboost and optimization of the Genetic Algorithm can be seen in Table 1.

Table 1. Result of SVM and Neural Network

No	Dataset Ratio	Algoritma	
		SVM	Neural Net
		accuracy	accuracy
1	50%	80,00%	77,78%
2	60%	77,78%	80,56%
3	70%	77,78%	81,48%
4	80%	77,78%	77,78%
5	90%	77,78%	77,78%
average		78,22%	79,08%

From the experimental results using split validation, it can be seen that the average results for the neural network are 0.86% in accuracy for the results, but in this dataset the results are not evenly distributed or not balanced, as follows for the highest results for neural networks and Support Vector Machines.

a) Support Vector Machine

Table 2. Result of Support Vector Machine

	true 1	true 0	class precision
pred. 1	35	8	81.40%
pred. 0	1	1	50.00%
class recall	97.22 %	11.11%	

The highest accuracy result of the SVM algorithm is 80% without using optimization.

b) Neural Network

Table 3. Result of Neural Network

	true 1	true 0	class precision
pred. 1	21	5	80.77%
pred. 0	0	1	100.00%
class recall	100.00%	16.67%	

The highest accuracy result from the Neural Network algorithm is 81.48% without using optimization.

For the dataset trained in the Neural Network algorithm, the data is evenly distributed but the accuracy is not categorized as very good and for testing the Support Vector Machine algorithm for accuracy it cannot be said to be

good, therefore to achieve good results, the next experiment will use adaboost so that the data is balanced. then the comparison for the following two algorithms can be seen in Table 4.

Table 4. AdaBoost Result for Dataset of Immunotherapy

No	Dataset Ratio	Algoritma	
		SVM	Neural Net
		accuracy	accuracy
1	50%	86,67%	93,33%
2	60%	86,11%	91,67%
3	70%	85,19%	92,59%
4	80%	83,33%	100,00%
5	90%	77,78%	100,00%
Average		83,82%	95,52%

It can be seen from the results of the classification using Adaboost to balance the data, it can be concluded that the Immunotherapy dataset is more suitable for use in the Neural Network Classification Algorithm from the average split validation ratio of 0.5 - 0.9 the dataset shows results above the average good category. Especially at the split ratio Validation 0.8 – 0.9 managed to achieve 100% accuracy but for the balance of the data it is still not achieved. In this experiment, we will use the Genetic Algorithm and Adaboost Feature Optimization to improve the results of the neural Network Algorithm classification and Support Vector Machine can be seen for the results in Table 5.

Table 5. Clasification Result of Optimation Genetic Algorithm Fiture and Adaboost

No	Dataset Ratio	Algoritma	
		SVM + GA	Neural + GA
		accuracy	accuracy
1	50%	88,89%	91,11%
2	60%	86,11%	94,44%
3	70%	85,19%	92,59%
4	80%	83,33%	100,00%
5	90%	88,89%	100,00%
average		86,48%	95,63%

The following results of the final stage of the experiment can be studied from the results above that Using the Optimization of the Genetic Algorithm feature Adds a little average results. In the Support Vector Machine algorithm, it increases by 2.66% and for the Neural Network algorithm it increases by 0.11%, relatively very little to increase the average accuracy, but for the increase in the ratio in each algorithm, it increases by more than 1%, especially significant changes to the algorithm. Support Vector Machines. To find out that the optimization of genetic algorithm features has increased significantly, different tests can be carried out, a t-Test Paired Two Sample for Means. The results of the Paired Two Sample for Means t-Test test resulted in the P value of the t-Test of 0.017. The results of this test show that the application of genetic algorithms to optimize the features of Genetic Algorithm and Adaboost can significantly improve the accuracy performance of the Dataset Immunotherapy Algorithm Classification of Support Vector Machines and Neural Networks, Indicators of the Proposed Method Objectively on Measurement Methods taken on Neural Networks and Support Vector Machines significantly. Indicated by P Value of t-Test < 0.05. The results of the t-Test test can be seen in Table 6.

Table 6. T-Test Paired Two Sample for Means Algoritma SVM and Neural Net

	Svm	Neural net
Mean	0,86482	0,95628
Variance	0,000583	0,001732
Observations	5	5
Pearson Correlation	-0,24284	
Hypothesized Mean Difference	0	
Df	4	
t Stat	-3,86228	
P(T<=t) one-tail	0,009056	
t Critical one-tail	2,131847	
P(T<=t) two-tail	0,018112	
t Critical two-tail	2,776445	

Based on the results of the t-Test significance test, it can be concluded that the genetic algorithm has a good performance to improve the performance of the classification algorithm through the feature selection stage.

5. Conclusions

The results of the classification of the Immunotherapy dataset using the Support Vector Machine and Neural Network algorithms can be concluded that the Neural network algorithm is superior to the Support Vector Machine seen from the average value of the Neural Network algorithm, getting the average results from 3 research stages, namely 79.08% of the classification without using optimization, for the average value of 95.52% using Adaboost optimization to improve the performance of the Neural Network Algortima, and the highest average of the Neural Network algorithm is 95.63% using genetic algorithm feature optimization and algorithm optimization using Adaboost. The Support Vector Machine algorithm gets an average result with a value of 78.22% for classification without using optimization, while with a value of 83.82% using Adaboost optimization, and the last with a value of 86.48% using genetic algorithm feature optimization and optimization. the algorithm uses the Adaboost algorithm, it looks quite far from the results obtained in the 3 stages of testing, Genetic Algorithm that can handle various optimizations depending on the objective function (fitness) whether balanced or unbalanced, linear or non-linear, continuous or discontinuous, or with random noise (Gorunescu, 2011). Genetic Algorithm that is applied using optimization features and Adaboost is proven to be effective in increasing the accuracy value significantly. This is based on 3 experiments carried out using split validation from various ratios, this method is effective because the T-test results show p value <0.05 which indicates this result is significant.

Acknowledgment

The work supported by Research and Technology Transfer Office, Bina Nusantara University, as a part of Bina Nusantara University's International Research Grant entitled "The Neuro research of Green Advertising: Study of Green Consumer Behavior" with contract number: No. 017/VR.RTT/III/2021 and contract date: 22 Maret 2021

References

Amrita Naika, L. S. (2016). Correlation review of classification algorithm using data mining tool: WEKA , Rapidminer , Tanagra ,Orange and Knime. international Conference on Computational Modeling and Security.

Asbar, Y., & Saptari, M. A. (2017). Analisa Dalam Mengukur Kualitas Pelayanan Terhadap Kepuasan Konsumen Menggunakan Metode PIECES. JURNAL VISIONER & STRATEGIS, 39-47.

Ayat, N. E. (2005). Automatic model selection for the optimization of SVM kernels. Pattern Recognition 38.

AyonDey, A. (2016). Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. nternational Journal of Computer Applications .

CORINNA CORTES, V. V. (1995). Support-Vector Networks. Machine Leaming, 20, 273-297 .

Derwin Suhartono, S. M. (2012). BINUS UNIVERSITY shcool of computing science. Retrieved from Bina Nusantara: https://socs.binus.ac.id/2012/07/26/konsep-neural-network/

Feindt, M., & Kerzel, U. (2006). The NeuroBayes neural network package. Nuclear Instruments and Methods in Physics Research A.

Gorunescu, F. (2011). Intelligent systems . Gorunescu, Ed.

Jain, R. S. (2018). Feature Selection for Cryotherapy and Immunotherapy Treatment Methods Based on Gravitational Search Algorithm. International Conference on Current Trends Towards Converging Technologies.

Jesmin Nahara, T. I.-P. (2013). Computational intelligence for heart disease diagnosis: A medicalknowledge driven approach. Expert Systems with Application.

M. Azka Putra, N. A. (2018). Wart treatment method selection using AdaBoost with random forests as a weak learner. Communications in Science and Technology .

Roopal Jain, P. M. (2018). Feature Selection for Cryotherapy and Immunotherapy Treatment Methods Based on Gravitational Search Algorithm . Jain, R., Sawhney, R., & Mathur, P. (2018). Feature Selection for Cryotherapy and Immunotherapy Treatment Methods Based on Gravitational Search Algorithm. 2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT).

Santosh Kumar Paidi, J. R. (2020). Label-free Raman spectroscopy and machine learning enables sensitive evaluation of differential response to immunotherapy.

Wang, H. Z. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. . European Journal of Operational Research,.

Wang, S.-C. (2003). Artificial Neural Network. Interdisciplinary Computing in Java Programming.

Zhou, Y. (2021). A problem-specific non-dominated sorting genetic algorithm for supervised feature selection. Information Sciences.