A Computational Method to Assess Post-operative Risk of Lung Cancer Patients

Kittipat Sriwong, Kittisak Kerdprasop, Paradee Chuaybamroong, and Nittaya Kerdprasop

***Abstract*—Lung cancer surgery is risky such that sometime patients died after surgery. To reduce loss, we try to create a computational model to anticipate in advance the post-operative survival among the lung cancer patients using statistical and machine learning algorithms. The dataset used in our model building process is data of patients who underwent lung cancer surgery comprising of 470 records with 17 attributes. These data were collected at Wroclaw Thoracic Surgery Centre, Poland during the years 2007 to 2011. For the purpose of validating the built model, we partitioned this dataset into training set and test set with the ratio 70% : 30% and random it 10 times to obtain 10 pairs of training-test set. The training dataset is used as input to build prediction models for the post-operative survival in the lung cancer patients by applying logistic regression and support vector machine (SVM) algorithms. The obtained two models are then compared to choose the best one with the highest predictive performance based on the mean accuracy of the ten iterations. As a result of comparison using test dataset, prediction model built from the logistic regression reaches 82.38% on its average accuracy, while the SVM approach yields 75.67% of its average accuracy.**

***Index Terms*—Post-operative survival assessment, lung cancer, machine learning, logistic regression, support vector machine.**

I. Introduction

Cancer, or malignant tumor, is a major health problem in most countries. It is reported by the World Health Organization [1] as the second leading cause of death worldwide, next to the heart disease and stroke which are the world's top killers. The death from cancer is accounted to almost 20% and the number is higher in the low and middle income countries. The top-five deadly cancers in descending order are lung cancer, liver cancer, colorectal cancer, stomach cancer, and breast cancer [1]-[3]. The four common risk factors for cancers are tobacco smoke, alcohol use, unhealthy food, and the lack of physical activity.

To improve survival chances of people having cancers, early diagnosis is important for providing effective treatment plan. A cancer treatment normally requires one or more curing modalities such as surgery, radiotherapy, and chemotherapy. Monitoring patient's condition after getting surgery is critical to the achieving of high cure rate and the prolog of patient's life. We thus interested in applying the machine learning methods to help modeling patient's survival chance after curing cancer by means of surgery.

Machine learning is a computational method recently applied to assist cancer diagnosis and risk factor modeling [4], [5], [6]. In this work, we perform a comparative study of two modeling techniques: logistic regression and support vector machine. Logistic regression is a special kind of multiple linear regression that has been designed to deal with categorical target attribute [7], whereas support vector machine [8] is a learning algorithm that can capture both linear and non-linear relationships between the categorical target attribute and the categorical/numeric explanatory attributes. Our research methodology is explained in Section II. The results are shown in Section III with discussions provided in Section IV. We conclude this paper in Section V.

Manuscript received April 7, 2020; revised July 12, 2020. This work was supported by grants from the National Research Council of Thailand and Suranaree University of Technology through the funding of Data and Knowledge Engineering Research Unit.

K. Sriwong, K. Kerdprasop, N. Kerdprasop are with the School of Computer Engineering, Suranaree University of Technology, Thailand (e-mail: nittaya@sut.ac.th).

P. Chuaybamroong is with the Department of Environmental Science, Thammasat University, Rangsit Campus, Thailand.

II. Research Methodology

A. Dataset Characteristics

To perform comparative modeling methods, we used thoracic surgery dataset from the UCI Machine Learning Repository [9]. The dataset was collected during the years 2007-2011 at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland.

This dataset has 470 data instances and 17 attributes with 2 different classes of T (true) and F (false). The class T means the risk of not survive one year critical period after surgery, and the class F means the risk is false, that is, patient can survive after the one year critical period. From then total 470 patients' records, the class T (risk of not survive) contains 70 data instances; the other 400 instances are in class F (can survive the critical one-year period). The details of 17 data attributes are summarized and explained in Table I.

In this dataset, there are three numeric attributes: PRE4, PRE5, and Age. Their statistical summaries are presented in Table II. The other fifteen attributes are categorical and their countable values are summarized in Table III.

TABLE I: Details of Data Attributes

Attribute	Meaning	Value
DNG	Diagnosis codes for primary tumor, secondary tumor, or multiple tumors	{ DGN1, DGN2, DGN3, DGN4, DGN5, DGN6, DGN8 }

DOI: 10.7763/IJMO.2020.V10.762

PRE4	Forced vital capacity	Numeric
PRE5	Exhalation volume at the end of the first second of forced expiration	Numeric
PRE6	Patients' performance based on Zubrod scale	{ PRZ0, PRZ1, PRZ2 }
PRE7	Does the patient feel pain before surgery?	{True, False}
PRE8	Is there haemoptysis before the surgery?	{True, False}
PRE9	Is there dyspnoea before the surgery?	{True, False}
PRE10	Does patient cough before the surgery?	{True, False}
PRE11	Does patient show weakness before surgery?	{True, False}
PRE14	Size of the original tumor	{ OC11, OC12, OC13, OC14 }
PRE17	Does patient have type 2 of diabetes mellitus?	{True, False}
PRE19	Does patient have myocardial infarction (heart attack) within 6 months?	{True, False}
PRE25	Does patient have peripheral arterial diseases?	{True, False}
PRE30	Does patient smoke?	{T, F}
PRE32	Does patient have asthma symptom?	{T, F}
AGE	Age of the patient	Numeric
Risk1Yr	Chance of one year survival period after surgery	{True, False}

TABLE II: STATISTICS OF NUMERIC ATTRIBUTES

Attribute	Minimum	Maximum	Mean	S.D.
PRE4	1.44	6.30	3.28	0.87
PRE5	0.96	86.30	4.57	11.76
AGE	21	87	62.53	8.71

TABLE III: DISTRIBUTIONS OF CATEGORICAL ATTRIBUTES

Attribute	Value	Count
DNG	DGN1	1
	DGN2	52
	DGN3	349
	DGN4	47
	DGN5	15
	DGN6	4
	DGN8	2
PRE6	PRZ0	130
	PRZ1	313
	PRZ2	27
PRE7	True	31
	False	439
PRE8	True	68
	False	402
PRE9	True	31
	False	439
PRE10	True	323
	False	147
PRE11	True	78
	False	392
PRE14	OC11	177
	OC12	257
	OC13	19
	OC14	17
PRE17	True	35
	False	435
PRE19	True	2
	False	468
PRE25	True	8
	False	462
PRE30	True	386
	False	84
PRE32	True	2
	False	468
Risk1Yr	True	70
	False	400

B. *Analytical Framework*

In our comparative study of logistic regression and SVM modeling methods, we identify the attribute *Risk1Yr* as our predicting target. The other 16 attributes play the predictor role. The steps in modeling and evaluating are graphically illustrated in Fig. 1.

C. *Model Assessment Criteria*

In this work, we compare the performance of the logistic regression and the SVM models based on the three measurement metrics: overall accuracy, true positive rate (TPR), and true negative rate (TNR). The accuracy, normally computed in percentage, is the ability of the model to predict correctly both patients having risk not surviving the critical one year period after surgery (positive class) and those who can survive the critical period (negative class).

The TPR, also called *sensitivity*, is the metric that pays more attention to the correct prediction of patients not surviving the critical period as appose to the actual cases of death. The TNR, or *specificity*, can be interpreted in the same way as TPR but the attention is on the patients in negative class. The model's performance evaluation is based on the prediction results on test data and such results are traditionally represented as a matrix, called confusion matrix, as shown in Table IV. The computations of accuracy, TPR, and TNR are shown in equations 1-3, respectively.

TABLE IV: CONFUSION MATRIX FOR MODEL ASSESSMENT

	Model predicts as positive	Model predicts as negative
Actual positive cases	TP (true positive)	FN (false negative)
Actual negative cases	FP (false positive)	TN (true negative)

$$\text{Overall Accuracy} = (TP + TN) / \text{All test data} \quad (1)$$

$$\text{True positive rate} = TP / (TP + FN) \quad (2)$$

True negative rate = $TN / (TN + FP)$ (3)

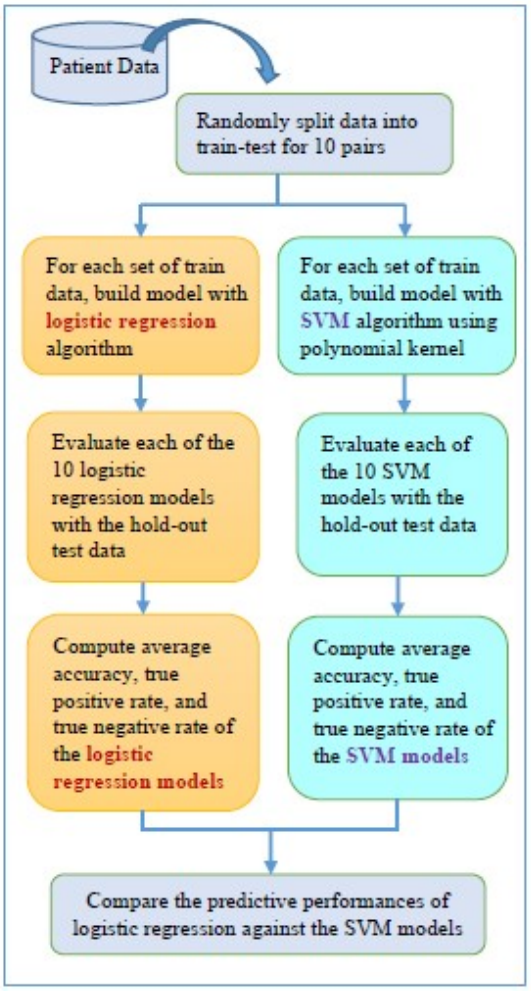

Fig. 1. Framework of our surgery survival modeling.

III. Comparative Analysis Results

A. Data Exploration

To evaluate the overall performance of logistic regression model versus the SVM model, we use separate train-test data (train data 70% to test data 30% proportion) and iterate the model prediction 10 times. The results are shown in Table V.

TABLE V: Comparison of Overall Accuracy (%) of Logistic Regression Model and SVM Model

Iteration	Logistic regression model	SVM model
1	84.67	73.33
2	82.64	73.69
3	81.75	75.18
4	77.62	72.73
5	84.30	79.34
6	85.93	74.81
7	83.97	73.28
8	79.77	75.14
9	78.47	80.56
10	84.67	76.64
average	**82.38**	**75.67**

B. TPR and TNR Comparisons

The comparative results of logistic regression and SVM models assesses on the TPR and TNR metrics are shown in Tables VI and VII, respectively. The TPR, TNR, and overall accuracy of the two models are graphically compared and shown in Fig. 2.

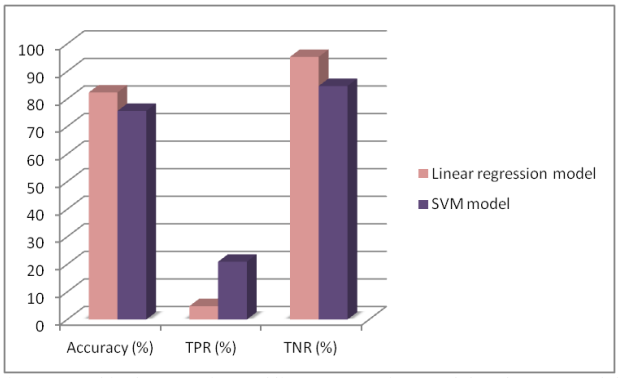

Fig. 2. Graphical comparison of linear regression model and SVM model.

TABLE VI: True Positive Rate (TPR) Comparison of Logistic Regression Model Versus SVM Model

Iteration	Logistic regression model	SVM model
1	0.053	0.105
2	0.100	0.250
3	0.042	0.167
4	0.000	0.263
5	0.214	0.214
6	0.000	0.222
7	0.000	0.111
8	0.036	0.276
9	0.040	0.240
10	0.000	0.250
average	**0.0485**	**0.2098**

TABLE VII: True Negative Rate (TNR) Comparison of Logistic Regression Model Versus SVM Model

Iteration	Logistic regression model	SVM model
1	0.962	0.824
2	0.944	0.839
3	1.000	0.876
4	0.902	0.798
5	0.934	0.869
6	0.991	0.829
7	0.973	0.832
8	0.978	0.847
9	0.941	0.924
10	0.967	0.835
average	**0.9592**	**0.8473**

IV. Results and Discussion

It can be seen from the result in Table III that in overall the logistic regression model can predict more accurate than the SVM model. On average, the accuracy of the logistic regression model is 82.38% accurate, whereas the SVM model with polynomial kernel yields lower predictive performance at the 75.67% accuracy.

When consider the issue of sensitivity, the SVM model shows better performance at the true positive rate 0.2098 on average, while the logistic regression model shows lower performance of sensitivity at 0.0485. On comparing specificity, the logistic regression model is better than the SVM model with the TPR rate at 0.9592 on average.

It is noticeable that the two models have trouble predicting positive cases, that are, the cases of patients not surviving a critical period of one year after getting lung surgery to cure cancer. These low performances among the two models are due to the imbalance problem existing in the dataset. This dataset contains only 77 cases of patients not survive the critical period, whereas the remaining 400 cases are those who can survive the surgery treatment. The high imbalance ratio of 77:400, or approximately 1:5, is the major cause of model's inefficiency. To consider the overall predictive performance and the specificity of the model, we can see that logistic regression performs better than the SVM algorithm. The logistic regression model is shown in Fig. 3.

```
Risk1Yr = (-0.2272 * PRE4)
        + (-0.0303 * PRE5)
        + (-0.009506 * AGE)
        + (-17.47 * [DNG=DGN1])
        + (-3.297 * [DNG=DGN2])
        + (-3.852 * [DNG=DGN3])
        + (-3.425 * [DNG=DGN4])
        + (-1.652 * [DNG=DGN5])
        + (-17.03 * [DNG=DGN6])
        + (0.2937 * [PRE6=PRZ0])
        + (-0.149 * [PRE6=PRZ1])
        + (-0.7153 * [PRE7=False])
        + (-0.1743 * [PRE8=False])
        + (-1.368 * [PRE9=False])
        + (-0.577 * [PRE10=False])
        + (-0.5162 * [PRE11=False])
        + (-1.653 * [PRE14=OC11])
        + (-1.214 * [PRE14=OC12])
        + (-0.4738 * [PRE14=OC13])
        + (-0.9266 * [PRE17=False])
        + (14.06 * [PRE19=False])
        + (0.09789 * [PRE25=False])
        + (-1.084 * [PRE30=False])
        + (13.39 * [PRE32=False])
        + (-19.35)
```

Fig. 3. A predictive model to estimate one-year survival chance of patients.

V. Conclusions

We present in this work the comparative results of applying two computational modeling techniques, logistic regression and support vector machine (SVM), to predict survival chance of patients underwent the surgery to cure cancer. The focus of prediction is the one year survival after surgery, which is the critical period of patients who are treated with surgery plan. Modeling survival chance is based on the machine learning techniques that are recently gained popularity in the medical domain. We study logistic regression technique because it has been extensively applied to estimate risk factors in medicine and life science. We compare logistic regression with the SVM because the later technique has been proven by the machine learning community that it can yield a promising result comparing to several existing machine learning techniques.

From the experimental results, we can conclude that SVM is more sensitive to logistic regression on predicting death among lung cancer patients who take the surgery treatment plan. This conclusion is due to the better result of SVM than the logistic regression regarding the true positive rate metric.

For the specificity measurement of estimating survival chance of patients, logistic regression model outperforms the SVM model. This conclusion is based on the measurement of true negative rate. Logistic regression is also better than SVM when consider the overall predictive accuracy of the model.

On observing characteristics of both machine learning techniques that there is no single method performs the best in every aspect of prediction, we thus plan to further our study by applying the ensemble method. That is, we are in the planning stage of combining the two models to yield better results on both true positive rate and true negative rate measurements.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The first author is responsible for collecting data and doing the experimentations. The second author sketches the research plan, confirms the experimentation results, and discuss the research outcomes. The third author helps discussing and confirming the research results. The last author is the corresponding author who is responsible for preparing the manuscript as well as validating research work in every step.

References

[1] WHO Media Centre. (2017). Cancer, World Health Organization. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[2] P. Anand, A. B. Kunnumakara, C. Sundaram, K. B. Harikumar, S. T. Tharakan, O. S. Lai, B. Sung, and B. B. Aggarwal, "Cancer is a preventable disease that requires major lifestyle changes," *Pharmaceutical Research*, vol. 25, no. 9, pp. 2097-211, 2008.

[3] B. W. Stewart and C. P. Wild, World Cancer Report 2014, International Agency for Research on Cancer, Lyon, 2014.

[4] T. Ayer, J. Chhatwal, O. Alagoz, C. E. Kahn, R. W. Woods, and E. S. Burnside, "Comparison of logistic regression and artificial neural network models in breast cancer risk estimation 1," *Radiographics*, vol. 30, no. 1, pp. 13-22, 2010.

[5] C. L. Chang and M. Y. Hsu, "The study that applies artificial intelligence and logistic regression for assistance in differential

diagnostic of pancreatic cancer," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10663-10672, 2009.
[6] H. Chen, J. Zhang, Y. Xu, B. Chen, and K. Zhang, "Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans," *Expert Systems with Applications*, vol. 39, no. 13, pp. 11503-11509, 2012.
[7] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, John Wiley & Sons, 2013.
[8] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, 2013.
[9] M. Zięba, J. M. Tomczak, M. Lubicz, and J. Świątek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Applied Soft Computing*, vol. 14, pp. 99-108, 2014.

Kittipat Sriwong is currently a master student with the School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Thailand. He has been fully financial support by grant from Suranaree University of Technology throughout his bachelor and master study. He is a research assistant in the data and knowledge engineering research unit, SUT. His current research of interest includes data mining, support vector machine and deep learning techniques, statistical data mining, and data mining applications in medical science.

Kittisak Kerdprasop is an associate professor and the chair of the School of Computer Engineering, SUT. He received his bachelor degree in mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, Florida, USA., in 1999. His current research includes machine learning and artificial intelligence.

Paradee Chuaybamroong is currently an associate professor in environmental science with the Department of Environmental Science, Thammasat University, Thailand. She received her bachelor degree in public health from Mahidol University, Thailand, in 1990, master degree in environmental science from Colorado School of Mines, USA. in 1997 and the doctoral degree in environmental science from University of Florida, U.S.A., in 2002. Her current research includes environmental science and engineering, geophysics, bioaerosols, and environmental photocatalysis.

Nittaya Kerdprasop is an associate professor and the head of data and knowledge engineering research unit, School of Computer Engineering, SUT. She received her bachelor degree in radiation techniques from Mahidol University, Thailand, in 1985, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, USA., in 1999. Her research of interest includes data mining, artificial intelligence, logic programming, and machine intelligence.