Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms

Animesh Hazra
Assistant Professor,
Computer Science & Engineering Department, Jalpaiguri Govt. Engg. College, Jalpaiguri, West Bengal, India

Nanigopal Bera
Student, Computer Science & Engineering Department, Jalpaiguri Govt. Engg. College, Jalpaiguri, West Bengal,
India

Avijit Mandal
Student, Computer Science & Engineering Department, Jalpaiguri Govt. Engg. College, Jalpaiguri, West Bengal,
India

ABSTRACT

One of the most common and leading cause of cancer death in human beings is lung cancer. The advanced observation of cancer takes the main role to inflate a patient's probability for survival of the disease. This paper inspects the accomplishment of support vector machine (SVM) and logistic regression (LR) algorithms in predicting the survival rate of lung cancer patients and compares the effectiveness of these two algorithms through accuracy, precision, recall, F1 score and confusion matrix. These techniques have been applied to detect the survival possibilities of lung cancer victims and help the physicians to take decisions on the forecast of the disease.

General Terms

Accuracy, Classification, Feature Selection, F1 Score, Precision, Recall.

Keywords

Lung Cancer, Logistic Regression, SVM, Confusion Matrix.

1. INTRODUCTION

One of the major reasons of non-accidental death is cancer. From many surveys, it has been demonstrated that lung cancer is the topmost reason of cancer demise in humans worldwide. The demise pace can be reduced if people go for early diagnosis so that suitable treatment can be administered by the clinicians within specified time. In brief, the cancer is an uncontrolled irregular growth of abnormal cells and invades the surrounding tissues. Lung cancer can be further generalized in two subsections, the first one is non-small cell lung cancer (NSCLC) and the second one is small cell lung cancer (SCLC). In this paper, the work is based on NSCLC patients since it is more complex and hard to cure. There are many contrasts regarding the detection and treatment of SCLC and NSCLC. There are various ways to detect the lung cancer, one of them is to apply its datasets besides SVM and LR algorithms to built and develop the classification and prediction model. Discovering the knowledge from great datasets are frequently worn for data mining techniques. It has found its important hold in every pasture including health care. It has played a major role for extracting the hidden information in the medical databases. The mining process is more than the data analysis which includes classification, clustering, union regulation of mining and prediction. If the cancer has spread, a person may feel symptoms in other places in the anatomy. Its symptoms are used to calculate the risk level of disease. A lot of symptoms are identified at the premature phase. They are a pain in the chest, cough can be persistent, dry with phlegm, frequent respiratory infections, shortness of breath, fatigue or loss of appetite, chest malaise and hoarseness. The major focus of this study is to predict the risk level of lung cancer using SVM and LR algorithms respectively.

2. CAUSES AND DETECTION OF LUNG CANCER

Diagnosis of lung cancer is done medically. In general, the patient will be aware when the cancer is an advanced stage. Smoking (it could be in energetic or unresisting) is undoubtedly the principal basis of lung cancer. In non-smokers, lung cancer can be sourced by the presence of air pollution, radon and used smoke or different elements like diesel exhaust, workplace exposures to asbestos or certain other causes. Sometimes the patient got the existence of malignant cells early by accidentally. For example, it may be incidentally found at the time of testing of some other disease related testings. The chest X-ray can reveal abnormal mass or nodule in the lung. In general, CT scan is used for screening the chest to find its cell. Low-dose computed tomography (LDCT) uses the lesser amount of radiation than a general CT scan. Sputum cytology test can also exhibit lung cancer. If a person has a cough and generating sputum, testing the sputum beneath the microscope can periodically reveal the lung cancer cells. Fine needle aspiration cytology and or biopsy test are used for detecting lung cancer.

The other technique that prefers doctors to detect the spot of lung tumor or lung cancer metastases is MRI. Here, MRI produces images using magnetic fields which show comprehensive images of the body but not like as X-ray. When moving images are considered MRI scan does not work well, like in a case of lungs which expands as we breadth in and contracts when we breadth out. In severe case it can affect brain, bones or distant sites.

3. LITERATURE SURVEY

A procedure to impending appraise and the use of positron emission tomography along with the glucose analog fluorodeoxyglucose (FDG-PET) for forecasting the reaction of chemotherapy in patients with promoting NSCLC is preferred by Wolfgang A. Weber et al. [1]. The detection of pneumoconiosis by using various feel subset of lung disorders based on support vector machine is suggested by Runxuan Zhang et al. [2]. The advent of following batch chained and enabled to study thoroughly genomic alterations is proposed by the Daniel Morgensztern [3]. The results from immune checkpoint avoidance are very inspired. This analysis summarizes late proceed in the region of cancer genomics, targeted therapies and immunotherapy. Whether they could recognize a gene declaration impression in PBMCs that would accurately distinguishable patients with advance phase of lung cancer from noncancer controls with alike possibility parts (age, gender and smoking history) and whether alike a signing value in predicting lung nodules detected by diagnostic X-ray or CT scans were benign is offered by Michael K. Showe et al. [4]. Feature-based classification technique viz. SVM classifier is implemented to classify the lung nodules in LDCT slides into four types i.e. well circumscribed, vascularized, juxta-pleural and pleural-tail. The offered technique was trained with the help of polynomial kernel by C-SVC. The likelihood determine upon the disparate kinds were forecasted with the obtained SVM model which was accustomed to classify the feature legend into four categories, is offered by Fan Zhang et al. [5]. Currently, SVM has obtained great observation as a functional implement for image recognition (Avci E [6]). The application of SVM employs two basic steps as instance training and testing. The initial step required feeding known data to the SVM across with the previously known decision. In the training set a SVM gets its comprehension to categorize unknown data is proposed by Van Belle et al. [7]. A few learning have been already used by presenting the SVM, ANN and Bayesian classifier for distinguishing obstructive lung diseases and SVM achieved by leading presentation for classification are suggested by Lee Y et al. [8]. Jonghyuck Lim et al. [9] proposed a new procedure where SVM provide better overall computation for intermediate lung disease contrast in strong intensity computerized tomography images. The LR technique of using historical information on a certain attribute or event to recognize figures which will oblige predicting a future value of the same with a certain probability attached to it and achieved freshmen enrolments is suggested by Vijayalakshmi Sampath et al. [10].

4. FEATURE ANLYSIS ON DATASET

In this project, the clinical dataset is taken from cancer imaging archive [11] which has 422 observations and 10 variables. Here 9 variables are predictors and one outcome variable. The output variable is either 1(dead) or 0(alive). The dataset containing of 10 features are described in Table 1.

Table 1. Description of attributes of CGD (Clinical and Genomic Data) lung cancer dataset [11]

Sl. No.	Attributes	Description of Attributes
1.	Patient ID	Sample code number.
2.	Age	Actual age when cancer is detected.
3.	Clinical. T. Stage	Clinical tumor stage represents T score ranges between T0 to T4. T0 means there is no primary tumor. Higher values mean the greater extent of the cancer.
4.	Clinical. N. Stage	Clinical nodes stage reflects the extent of cancer within nearby lymph nodes. It ranges between N0 to N3. N0 means that no cancer was found in lymph nodes and for N1 to N3, higher the value greater the extent of a cancer.
5.	Clinical. M. Stage	Clinical metastasis stage tells us if the cancer has spread to distant sites. It ranges between M0 to M3. M0 means there is no cancer in distant sites. Higher value means the greater extent of the cancer.
6.	Overall. Stage	There are four types of stages used in the dataset. They are I, II, IIIa and IIIb.
7.	Histology	Though the data is based on nonsmall cell lung cancer so the data are four types large cell, squamous cell carcinoma, adeno carcinoma and nos.
8.	Gender	It specifies the patient’s sex.
9.	Survival. time	It represented the number of days survived by the patients after diagnosis.
10.	Dead status event	It represents the patient is dead (1) or alive (0).

5. PROPOSED METHODOLOGY

The clinical data taken from CGD (Clinical and Genomic Data) portal is used for further processing to get the desired outcome. The proposed methodology built here containing of three phases. At first pre-processing of data set is done. Then the dataset will be split into two lays, one is used for training phase and another for testing phase. Next, classification on the dataset using SVM and LR algorithms were performed. After training the dataset, finally the model will be evaluated. The workflow diagram and all phases of the proposed algorithm are represented in Figure 1.

5.1 Workflow Diagram of the Proposed Methodology

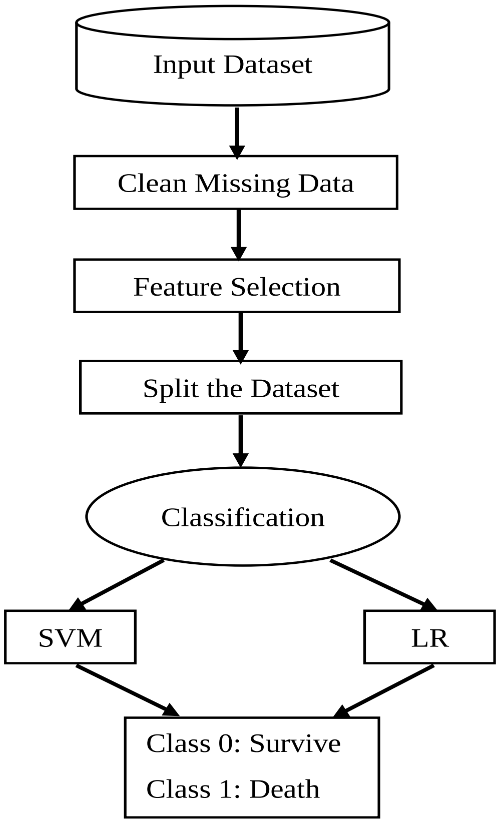


Fig 1: Workflow diagram for predicting lung cancer survivability using support vector machine and logistic regression algorithms respectively

Initially, the CGD dataset is captured as input and then apply the data cleaning technique for cleaning the missing data. Next, feature selection was implemented on the normalized dataset using Pearson Correlation Coefficient (PCC) technique which reveals how much the attributes of the dataset are related to the class attribute and based on that a ranking of the properties has been obtained. Then split that dataset into two segments i.e. train data and test data. Here, test data is 20% and train data is 80%. These test and train data have been taken and applied to two classification techniques viz. SVM and LR respectively. If the output of the classifier is 1 then the patient will die otherwise the patient will survive.

The steps of the proposal algorithm are discussed as follows.

5.2 The Proposed Algorithm for Lung Cancer Prediction

Step 1: Take the lung cancer dataset as input.

Step 2: Clean the missing values from the dataset.

Step 3: Apply the normalization technique on the dataset.

Step 4: Apply Pearson correlation coefficient (PCC) technique for feature selection.

Step 5: Split the dataset into two subsets.

Step 6: Perform the SVM and LR classification techniques on the training dataset.

Step 7: Evaluate the model.

Step 8: Find and compare the accuracies of the SVM and LR classifiers.

The above-mentioned steps are explained below in details.

5.2.1 Pre-processing Phase

5.2.1.1 Data Cleaning

The CGD (Clinical and Genomic Data) dataset taken here consists of 422 test case data. In big data sets, it is common for missing values to occur in several variables. That's why the dataset is cleaned using multiple imputations by chained equation (MICE) method [12]. MICE is an empirical approach for generating imputations(MI Stage 1) build on a put of imputation models, every variable with missing values. Entirely dependent identification and consecutive regression, multivariate imputation is also called MICE. Initially, all missing values are filled in by easy arbitrary sampling with a successor from the detected values. The first variable with missing values say x_1, is regressed on all different variables x_2,....,x_k confined to originals with the observed x_1. Missing values in x_1 are exchanged by simulated draws from the corresponding posterior predictive spread of x_1. Then, the upcoming variable with missing values say x_2, is regressed on all other variables x_1,x_3,......,x_k confined to originals with the observed x_2 and using the imputed values of x_1. Again, missing values in x_2 are exchanged by draws from the posterior predictive issuance of x_2. This procedure is restated for all different variables with missing values in turn which is called a cycle. In order to stabilize the results, the procedure is usually repeated for several cycles to build a single assigned data set and the whole procedure is restated m times to give m assigned data sets. MICE has an ability to handle different variable types (continuous, binary, disordered absolute and ordered absolute) because each variable is assigned by utilizing its unique assigned model. Every variable is assigned using its assigned model in MICE.

5.2.1.2 Normalize the Dataset

It is a procedure which is used to normalize the range of distinct variables or features of data. It is generally accomplished throughout the data preliminary processing phase. Normalization can be performed at the elevation of the input features or at the elevation of the kernel. In many applications, the available features are continuous values, where each feature is measured in a different scale and has a different scale of feasible values. In such cases, it is often beneficial to scale all features to a common range by standardizing the data.

5.2.2 Feature Selection

The main objective of feature selection policy is to identify those imputes or features which are correlated with output values where the values depend upon a specific input which is collected by applying some useful test. The correctness and effectiveness of categorization can be potentially improved by taking the good features. Pearson correlation coefficient (PCC) technique is used here to select the most dominant features. PCC is very well known as statistical model with r value. This r value demonstrates the toughness of the correlation between two variables [13]. The coefficient is not affected by changing of scale in the two variables.

5.2.3 Dataset Partitioning

Here, the lung cancer dataset has 422 test cases where the dataset is split into train and test data. Many combinations of train and test data have been done. But maximum accuracy can be obtained when the test data is 20% (84 test cases) and train data is 80% (338 test cases) of the total test cases.

5.2.4 Classification Techniques Applied on the Training Dataset

5.2.4.1 Support Vector Machine (SVM)

Here, it is used for classification purpose. They are built on the thought of the conclusion level that defines conclusion bordered between groups of instances. A decision plane of SVM is used for separation between a set of items having a different group of membership and distinct a few support vectors in the training set.

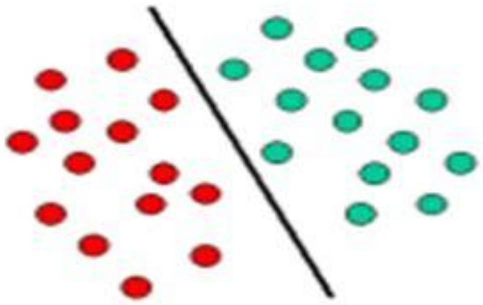


Fig 2: Example of linear support vector machine

Figure 2 gives a superior example of a linear classifier that divides a set of items into their corresponding groups (green and red in this case) with a line. Most classification jobs although are not that easy and frequently more composite shapes are required in form to make an optimal separation i.e. accurately classify new test cases on the inception of the samples that train cases [14]. This situation is illustrated in

Figure 3. Classification jobs build on representation dividing lines to differentiate between items of different group memberships is also called hyper plane classifiers. SVM are especially adapted to handle such jobs.

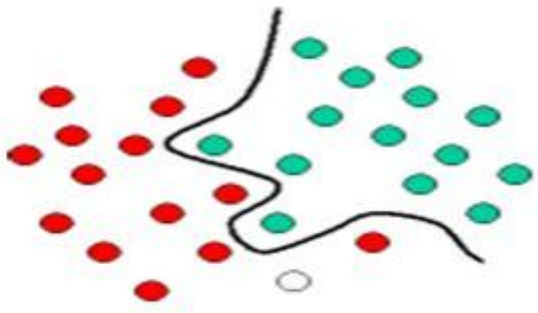


Fig 3: Example of non-linear support vector machine

Figure 4 represents the primary layout of the kernel SVM. Here we distinguish the primitive items (left part is diagrammatic) mapped i.e. reposition using a place of mathematical functions, called kernels [14]. The procedure of repositioning the items is called mapping. Annotation in this new surroundings, the mapped items (right side of the diagrammatic) is linearly distinct and thus alternatively building the composite curve (left diagrammatic). All we have to find an optimal line that can separate the green and the red items.

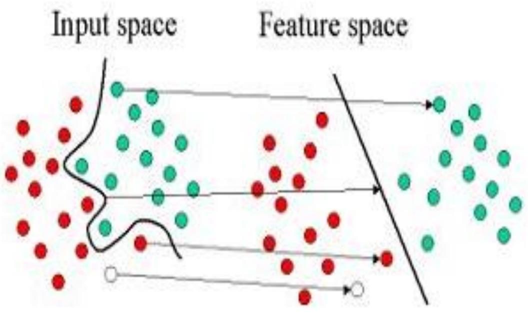


Fig 4: Example of kernel support vector machine

5.2.4.2 Logistic Regression

It is the generalized form of linear regression [7]. Primarily, it is worn for computing binary or multi-class dependent variables. Because the reply variable is discrete, it cannot be modelled directly by linear regression. For building a model, it forecast the odds of its event instead of forecasting the point estimate of the occurrence. In two class problem when the result of odds greater than 50% then the class is designated by assigned value 1 otherwise it is 0. However it is a very ably accepted kit, it allows that the reply variable is linear in the coefficients of the forecast variables. Then, using the experience of data analysis the experimenter must choose the original inputs and decide their functional relationship to the reply variable.

5.2.5 Evaluate the Model

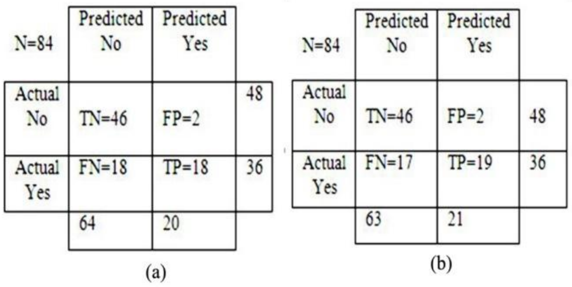
Evaluate model needs a dataset as an input. Here it is needed to train the model and build forecasts on some dataset using the trained model before evaluating the results. It is built on the trained labels or probabilities along with the true labels, all of which are output by the training model. Evaluation is excellence process to appraise the presentation of the model. Both labels create evaluation metrics that can test or contrast across those of distinct models. In binary classification model the above metrics used are accuracy, precision, recall and F1 score. Here the confusion matrix represents the number of true positives, false negatives, false positives and true negatives values along with ROC, precision-recall and lift curves to measure the performance of the model.

5.2.6 Find and Compare the Accuracies

In this study SVM and LR classification techniques are used to find out the accuracies of the proposed method. After getting the accuracies, comparison is made and the highest one is selected. The number of all correct forecasts split by the total number of samples is called as accuracy.

6. RESULT AND DISCUSSION

In this project, a mass study on two classification procedure have been managed and provided a basis for comparisons among them in terms of accuracy percentage, precision, recall and F1 score. The level of effectiveness of the classification model is calculated by using confusion matrix [15][16]. In confusion matrix the real and forecasted classifications are done by a classifier where the true data are known. The classifier's representing it appraise by utilizing the data in a matrix. Figure 5 represents the confusion matrices for SVM and LR classifiers respectively.



N=84	Predicted No	Predicted Yes	
Actual No	TN=46	FP=2	48
Actual Yes	FN=18	TP=18	36
	64	20	

(a)

N=84	Predicted No	Predicted Yes	
Actual No	TN=46	FP=2	48
Actual Yes	FN=17	TP=19	36
	63	21	

(b)

Fig 5: Confusion matrices for (a) support vector machine and (b) logistic regression classifiers respectively

Two feasible forecasted classes i.e. yes and no are present there inside the matrix. If the cancer patient dies, it will be represented as yes and if it has survived then it will be represented as no. The classifiers build on entire 84 predictions. Out of those 84 cases, the SVM classifier predicted yes 21 times and no 63 times. Whereas the logistic regression classifier predicted yes 20 times and no 64 times. In reality, 36 patients have died and 48 patients not died.

The four basic terms of the confusion matrix are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). They are discussed here in detail. True positives (TP) is where prediction is yes (they have died) and they actually died. True negatives (TN) is where prediction is no and they don't die. False positives (FP) is where prediction is yes but they actually don't die. False negatives (FN) is where prediction is no but they actually died. The evaluation metrics accessible for binary classification representations are accuracy, precision, recall, F1 score and there are also three curves i.e. LIFT curve, ROC curve and precision-recall curve. All the evaluation metrics are discussed and calculated below in detail where SVMC denotes support vector machine classifier and LRC represents logistic regression classifier.

Accuracy: It represents how often is the classifier correct. Accuracy is calculated as the number of all correct predictions divided by the total number of the dataset.

Accuracy (SVMC) = (TP+TN)/Total = (18+46)/84 = 0.762 ≡ 76.20%.
Accuracy (LRC) = (TP+TN)/Total = (19+46)/84 = 0.774 ≡ 77.40%.

Precision: The fraction of appropriate instances among the retrieved instances is defined as precision. Precision is calculated as the amount of TP divided by predicted yes.

Precision (SVMC) = (TP/predicted yes) = 18/20 = 0.900.
Precision (LRC) = (TP/predicted yes) = 19/21 = 0.905.

Recall: It shows the snatch of applicable cases that have been retrieved over total relevant. The Recall is computed as an amount of TP divided by actual yes.

Recall (SVMC) = (TP/actual yes) = 18/36 = 0.500.
Recall (LRC) = (TP/actual yes) = 19/36 = 0.528.

F1 Score: In analytical testing of binary taxonomy, the test's accuracy is measured by an F1 score. Both the precision p and the recall r of the test is considered to compute the F1 score. The weighted average of p and r is regarded as F1 score.

F1 Score(SVMC) = 2*[(precision*recall)/(precission + recall) = 2*[(0.900*0.500)/(0.900+0.500)] = 0.643.

F1 Score (LRC) = 2*[(precision*recall)/(precision + recall) = 2*[(0.905*0.528)/(0.905+0.528)] = 0.667.

A curve is a technique of envisioning a classification model. The presentation of a targeting model at forecasting or classifying events as having an augmented acknowledgment is determined by a lift curve and it is settled across an arbitrary option targeting model. If the acknowledgment within the goal is much better than the average for the dataset as a whole then doing a good job for the targeting model can be achieved. A lift curve basically shows the ratio between positive rate and amount of TP and it is shown in Figure 6.

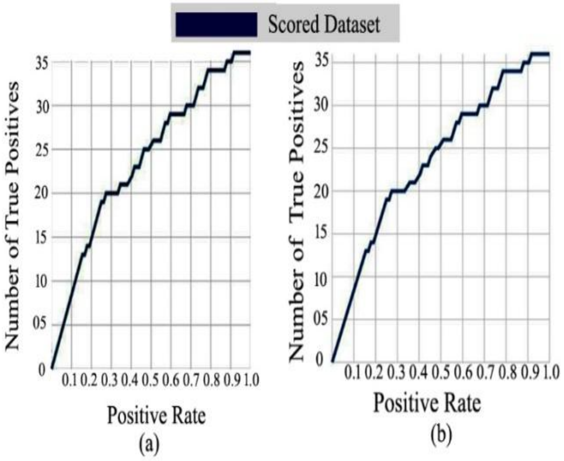


Fig 6: LIFT curves for (a) support vector machine and (b) logistic regression classifiers respectively

ROC curve is a graphical plot that explains the diagnostic facility of a binary classifier system and it primarily displays the ratio in between TP rate and FP rate which is demonstrated in Figure 7.

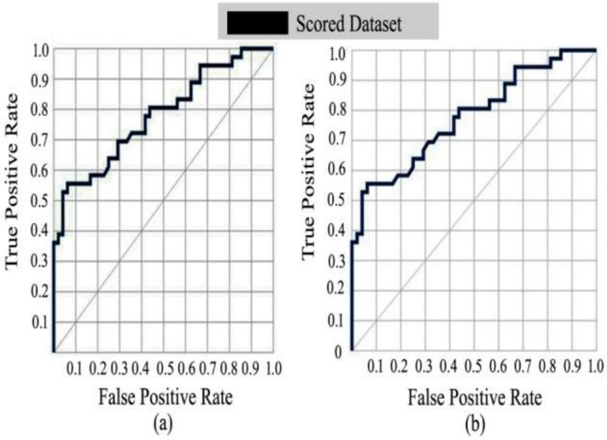


Fig 7: Receiver Operating Characteristic (ROC) curves for (a) support vector machine and (b) logistic regression classifiers respectively

In binary classification, information retrieval and pattern recognition precision and recall are widely used notations. Precision is the snatch of proper samples among the retrieved samples while recall is the snatch of proper samples that own existed fetched over total proper samples in the dataset. Both measures are accordingly established on comprehension and compute the relevance among them. The curve depends on the ratio in between these two which is illustrated in Figure 8.

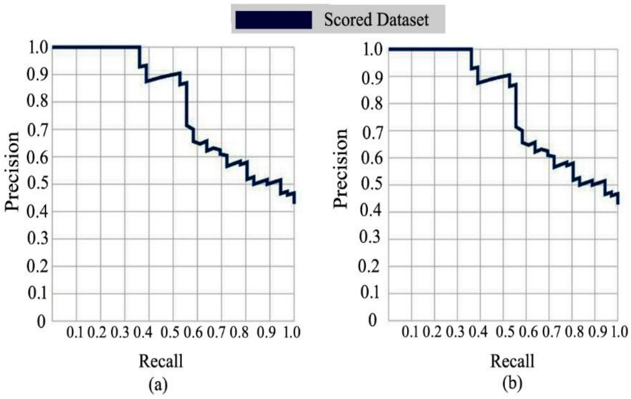


Fig 8: Precision-recall curves for (a) support vector machine and (b) logistic regression classifiers respectively

7. CONCLUSION

One of the major and frequent bases of cancer deaths globally in terms of both instance and transience is lung cancer. The main reason behind the increasing of deaths from it is detecting the disease lately and faults in effective treatment. So, the early detection is needed to save lives from this disease. The survivability rate of lung cancer can be predicted with the help of modern machine learning techniques. Accordingly, it would be clever to determine the survival possibilities among the patients. In this study data cleaning, feature selection, splitting and classification techniques have been applied for predicting survivability of lung cancer as accurately as possible. This project reveals that logistic regression classifier gives the topmost accuracy of 77.40% compared to support vector machine classifier which gives 76.20% accuracy. Also, the logistic regression classifier gives maximum classification accuracy concerning every different classifier. This work can further be enhanced by modifying logistic regression classifier which gives highest accuracy.

With the help of machine learning methods it is really difficult to diagnose the different medical conditions of a lung cancer patient and prediction of conditions are also more critical in nature. It is a challenging task in machine learning and data mining fields to construct a specific and computationally efficient classifier for medical applications. This can be a great future scope of this research. For big datasets how these classification algorithms behave, that is another future scope of this project. Moreover the identification of particular stage of lung cancer can be done in near future. Another prospect of this research is the time and space complexity analysis of different classification algorithms on medical datasets which can be explored in the forthcoming work.

8. REFERENCES

[1] Wolfgang A. Weber, Constantine A. Gatsonis, P. David Mozley, Lucy G. Hanna, Anthony F. Shields, Denise R. Aberle, Ramaswamy Govindan, Drew A. Torigian, Joel S. Karp, Jian Q. (Michael) Yu, Rathan M. Subramaniam, Robert A. Halvorsen, and Barry A. Siegel "Repeatability of18 F-FDG PET/CT in Advanced Non–Small Cell Lung Cancer: Prospective Assessment in 2 Multicenter Trials", The journal of Nuclear Medicine(JNN), DOI: 10.2967/jnumed.114.147728, Apr. 2015.

[2] Runxuan Zhang, Guang-Bin Huang, N. Sundararajan, and P Saratchandran, "Multicategory Classification Using an Extreme Learning Machine for Microarray Gene Expression Cancer Diagnosis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume 4, Issue 3, 2007.

[3] Daniel Morgensztern, Meghan J. Campo, Suzanne E. Dahlberg, Robert C. Doebele, M.D., EdwardGaron, David E. Gerber, Sarah B. Goldberg, Peter S. Hammerman, Rebecca Heist, Thomas Hensing, Leora Horn, Suresh S. Ramalingam, Charles M. Rudin, Ravi Salgia, LeciaSequist, Alice T. Shaw, George R. Simon, Neeta Somaiah, David R. Spigel, John Wrangle, David Johnson, Roy S. Herbst, Paul Bunn, and Ramaswamy Govindan, "Molecularly Targeted Therapies in Non-Small-Cell Lung Cancer Annual Update 2014",J Thorac Oncol. 2015 Jan; Vol. 10, Issue 1, pp. S1-S63,doi: 10.1097/JTO. 0000000000000405, 2015.

[4] Michael K. Showe, Andrew V. Kossenkov, and Louise C. Showe, "The Peripheral Immune Response and Lung Cancer Prognosis", Oncoimmunology,1(8),1414–1416, doi: 10.4161/onci.21093, 2012.

[5] Fan Zhang, Yang Song, Weidongcai, Yun Zhou, Shimin Shan, and Dagan Feng, "Context Curves for Classification of Lung Nodule Images", IEEE/ Digital Image Computing: Techniques and Applications (DICTA), DOI:10.1109/DICTA.2013.6691494, 2013.

[6] Avci E, "A New Expert System for Diagnosis of Lung Cancer: GDA-LS_SVM", J Med Syst, 36(3):2005-9. doi: 10.1007/s10916-011-9660-y, 2012.

[7] Van Belle V, Pelckmans K, Van Huffel S, and Suykens JA, "Support Vector Methods for Survival Analysis: A Comparison Between Ranking and Regression Approaches", Artif Intell Med, 53(2):107-18, doi: 10.1016/j.artmed.2011.06.006, Aug. 2011.

[8] Lee Y, Seo JB, Lee JG, Kim SS, Kim N, and Kang SH, "Performance Testing of Several Classifiers for Differentiating Obstructive Lung Diseases Based on Texture Analysis at High-Resolution Computerized Tomography (HRCT)", Compute Methods Programs Biomed, 93(2):206-15.doi: 10.1016/j.cmpb.2008.10.008, Dec. 2008.

[9] Jonghyuck Lim, Namkug Kim, JoonBeomSeo, Young Kyung Lee, Youngjoo Lee, and Suk-Ho Kang, "Regional Context-Sensitive Support Vector Machine Classifier to Improve Automated Identification of Regional Patterns of Diffuse Interstitial Lung Disease", J Digit Imaging, 24(6): 1133–1140, doi: 10.1007/s10278-011-9367-0, 2011.

[10] Vijayalakshmi Sampath, Andrew Flagel, Carolina Figueroa, "A Logistic Regression Model to Predict Freshmen Enrollments", Paper SD-016.

[11] NSCLC - Radiomics [online] https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics; jsessionid=84D480480E80C9EC195026 ED04D95433 [Dataset].

[12] Melissa J. Azur, Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf, "Multiple Imputation by Chained Equations: What is it and how does it work?", PMC, Int J Methods Psychiatr Res, 20(1): 40–49; Mar. 2011.

[13] MM Mukaka,"Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research", Malawi Med J, 24(3): 69–71, 2012.

[14] Peng Guan, Desheng Huang, Miao He, and Baosen Zhou, "Lung Cancer Gene Expression Database Analysis Incorporating Prior Knowledge with Support Vector Machine-Based Classification Method", J Exp Clin Cancer Res, 28(1): 103, 2009.

[15] A. K. Santra, and C. Josephine Christy, "Genetic Algorithm and Confusion Matrix for Document Clustering", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No. 2, 2012.

[16] Simple guide to confusion matrix terminology - Data School [online] http://www.dataschool.io/simple-guide-to-confusion-mat rixterminology/.