Original Article

Exploring Machine Learning Models for Recurrence Prediction in Lung Cancer Patients

Priyanka Ramesh[1], Anika Jain[1,2], Ramanathan Karuppasamy[1], Shanthi Veerappapillai[1,*]

[1]Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, INDIA.
[2]Biological Sciences Graduate Student, Purdue University, West Lafayette, IN, US.

ABSTRACT

Background: A proper assessment for the probability of recurrence in lung cancer is mandatory for a clinician to make an effective treatment-decision. **Materials and Methods:** Here, we employed machine learning algorithms to predict the lung cancer recurrence rate using the Caribbean and few white ethnicities populations. A 100 metastatic record with 15 predictor variables and 1 dependent variable was considered for model development. These models were evaluated using seven performance metrics, including accuracy and F1 score. **Results:** Our study results show that the decision tree outperformed the other models with the highest accuracy and F1 score of about 0.95 and 0.90, respectively. Of note, the *p*-value and correlation matrix show that the most significant features accounting for the tumor recurrence are cancer stage, ethnicity, tumor size, genome doubled and time to recurrence. **Conclusion:** Thus, our study provides insights into implementing machine learning algorithms to evaluate cancer outcomes in a clinical setting.

Key words: Machine learning, Lung cancer, Recurrence, Statistical analysis, Correlation matrix.

INTRODUCTION

Lung Cancer, with an overall survival rate of 5-years at a dismal 20%, is one of the leading causes of cancer-associated deaths worldwide.[1,2] Non-Small Cell Lung Cancer (NSCLC) is the most common type of lung cancer, accounting for 85% of all cases.[3] About 30-55% of people who suffer from NSCLC encounter an eventual increase in recurrence with the rise in stages of cancer (Stage 0 to Stage 4).[4] Recurrence can be defined as the return of cancer after 3 months of remission. It may occur due to the spread of original tumor cells that remained after the initial treatment. Local failure of certain treatment methods due to the histologic type can also lead to recurrence. Upon stereotactic body radiation therapy, a popular treatment method for NSCLC in the early stages with local tumor control rates of more than 90% has the greatest risk of local failure leading to recurrence.[5] Despite employing highly effective treatment methodologies such as the gold standard complete surgical resection on stage I adenocarcinoma, 18-32% of the patients had a recurrence and died within 5 years of the resection treatment.[6] The high possibility of recurrence even when the cancer is detected at an early stage necessitates the development of accurate predictive tools. Earlier, traditional data analysis in a clinical setting involved manual data accessing, processing, analysis and distribution for diagnosis or risk prediction.[7] Effective management of meaningful inferences is one of the challenges in conventional methods of data handling. Moreover, classical regression analysis was generally used which does not account for non-linear relationships between the variables and the expected outcomes.[8]

Recently, machine learning (ML) algorithms can be applied to large clinical datasets to gain extensive insight into the correlation between the different features and risk factors that influence disease progression and recurrence. As a result, the most crucial variables to the possibility of recurrence can

Submission Date: 02-01-2021;
Revision Date: 12-12-2021;
Accepted Date: 04-06-2022.

DOI: 10.5530/ijper.56.3s.147

Correspondence:
Dr. Shanthi Veerappapillai
Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore-632014, Tamil Nadu, INDIA.
E-mail: shanthi.v@vit.ac.in

www.ijper.org

be identified. These variables can be used to set up an accurate predictive classifier to identify the individuals most at risk of recurrence and who require better post-operative care as well as optimized adjuvant therapies. Ultimately, it reduces healthcare costs and burdens with early intervention.[9] Notably, machine learning can improve the accuracy by up to 25% of predicting the recurrence in cancer patients than the traditional data analysis strategy.[10] Hence, advanced technological methods such as machine learning provide an effective alternative method of systemic analysis of clinical data for important inferences, including possible treatment administration, prediction of risk factors, and recurrence.[11] In specific, machine learning can make predictions through the provided clinical factors by developing pattern-recognition criteria. Importantly, in recent years, ML studies were successfully implemented with diverse datasets to assess multiple treatments and post-treatment related procedures. For instance, ML has been used to predict breast cancer survivability using various algorithms with the aid of a dataset containing 8942 patient records distributed over 24 variables.[12] Although recurrence prediction ML models were reported in the recent literature, studies in the Caribbean and few white ethnicities populations are limited. Thus, our present study predicts the recurrence in the Caribbean and a few white ethnicities using different machine learning strategies. We hope analyzing the data with more significant ethnic variability may aid in gaining better insight into the relationship between ethnicity and recurrence.

MATERIALS AND METHODS

Dataset collection

A lung adenocarcinoma dataset of 100 patients was retrieved from the public database, BioStudies with accession number S-EPMC6196259. The dataset contains 12 categorical variables and 4 continuous variables. The dataset includes information on gender, age, smoking status, and ethnicity of the patients. It also included details of cancer status such as stage, genomic doubling status, vascular and pleural invasion levels. In addition, the details of treatment administered were highlighted on resection margins and adjuvant therapies administered. In essence, the dataset also records the recurrence status of patients and the duration it took to manifest post-treatment.

Moreover, this study is a secondary analysis of work submitted in the BioStudies repository. Hence only the following inclusion and exclusion criteria were considered during the selection of metadata. The inclusion criteria include: (i) Pathological confirmation of lung cancer, (ii) Cancer staging data, (iii) Received tumor resection using surgical approach, and (iv) Recurrence time. On the other hand, the exclusion criteria include: (i) Benign tumor and (ii) Tumor stage IV. Moreover, the selection criteria were cross-verified based on the recent articles published on cancer recurrence prediction.[13-15]

Dataset preprocessing

Indeed, data quality is the crucial factor which is directly correlated with the performance and accuracy of the models. This shows that preprocessing data is most important before model development.[16] In the present study, the dataset was cleaned appropriately to handle the 14 missing values that were found in the column "pack years". The missing values were filled with the mean value of each category of a variable.

Machine Learning Model Generation

The 100 patient samples were divided randomly into a training group and a test group in the ratio of 7:3 respectively for model development. The binary feature recurrence was used as the target variable during our analysis. Four ML algorithms were then applied: logistic regression, decision tree, random forest, and support vector machine (SVM). The goal of these models is to afford an accurate earlier prediction of recurrence in lung cancer patients to provide effective treatment.

Logistic Regression

Logistic regression (LR) is the most commonly used model for binary classification in epidemiology and medicine. Here, logistic regression is used to study the effect of multiple predictive features on binary categorical outcomes (recurrence/non-recurrence). This algorithm provides class probabilities describing the target variable 'Y' based on a linear combination of the input variable predictors assigned as 'X'.[17] The algorithm utilizes the given below equation for the model development.

$$\text{Logit}(p) = \beta 0 + \Sigma \beta i Xi \quad (1)$$

Where p represents the probability of binary outcomes.[18] The default parameters were used for logistic regression model generation. In this study, we have employed Logistic Regression sub-package of sklearn for performing this algorithm.

Decision Tree

Decision Tree (DT) algorithm has a tree-like structured scheme of classification. The most important feature is represented at the top root node in the resultant tree and are divided based on the features of the provided data. Each terminal node and leaf in the decision tree represent

the analysis's input features and outcome, respectively. The clear architecture of a decision tree and its ability to handle different types of data make it an easy and accurate classification method.[19] In general, Gini index scoring system was used to measure the distribution of the data based on the given below equation:

$$\text{Gini index} = 1 - \sum_{k=1}^{n} P_k^2 \quad (2)$$

Where P represents ratio of observation in each class.[20] Here, we have implemented Decision Tree Classifier sub-package of sklearn to construct our decision tree. To prevent overfitting, the algorithm was further optimized using tuning parameters: max_depth=10, min_samples_split=3.

Random Forest

Random Forest (RF) has its algorithmic basis same as in the decision tree. It is a classifier consisting of combination of multiple basic algorithms in which each can be analyzed through a decision tree basis. The random forest algorithm combines all the individual decision trees and produces an optimized mean result. Moreover, it can be applied with ease to both categorical and continuous variables.[21] R*f* was performed using RandomForestClassifier of sklearn package. In addition, tuning parameters n_estimators=1000, min_samples_leaf=2 was used to optimize the model results.

Support Vector Machine

Support Vector Machine (SVM) is a relative learning method that maps the vector entered in a higher dimensional feature space and then identifies the hyperplane that provides a separation of the points of data into two classes. The greater this distance, the lower is the expected generalization error. Thus, the classifier can efficiently classify new samples. SVM is especially useful in working with a large number of features as it only includes features that lie on the boundaries of the hyperplane.[22] Hence, SVM model with linear kernel was employed to divide the input space using the given below formulae:

$$W^T X_i + b = 0 \quad (3)$$

Here, the bias is denoted as b and the hyperplane is represented as W. In the current investigation, we employed svm sub-package from sklearn to stratify the dataset into recurrence/non-recurrence of cancer cells in lung cancer patients.

Model assessment

The performance of each model was evaluated in terms of the accuracy, precision, recall and F1 scores. In addition, receiver operating characteristic (ROC) curve was also analyzed for the developed models. Another important tool for performance evaluation of the classifiers for their ability to classify objects is confusion matrices. They examine the consistency between the predicted and actual classification of the models.[23] In addition, the precision, recall, specificity and sensitivity of the models were deduced from the developed confusion matrix using the given below formulae:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (6)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (7)$$

K-Fold Cross Validation

In general, cross validation is used to validate the performance of the developed models with an aim of obtaining unbiased outcome. In the current analysis, the models were validated using 5-fold and 10-fold cross validation strategy by splitting the dataset in the ratio of 4:1 and 8:2 randomly. Then the average of accuracy and F1-score were calculated to identify the best classifier. The analysis was performed using k-fold and cross_val_score subpackages of sklearn.model_selection package in python.[24]

Statistical Analysis

The patient dataset was then processed using statistical analysis to gain a clear understanding of the features. The *p*-value was calculated using a student's *t*-test to deduce significant differences between patients with and without cancer recurrence. In this study, the feature with a *p*-value less than 0.001 is considered statistically significant. Further, the continuous variables were analyzed in terms of their mean and standard deviation. On the other hand, categorical variables were summarized by their counts and percentages with respect to the recurrence in the patient.[23] Here, we have implemented statsmodel.api package of python to perform statistical analysis of the obtained dataset. In addition, correlation matrix was constructed to establish a relationship within the selected variables and the target.

RESULTS AND DISCUSSION

Data characteristic analysis

Data characterization analysis portrays the overall distribution of features and its association with recurrence in lung cancer patients. Table 1 depicts the significance difference between the categories of each feature in the dataset. The dataset of 100 entries belonging to lung cancer patients were considered for model generation and validation. The dataset contained both male and female patients in the ratio of 6:4 respectively. The range of age of patients considered in this study was between 48 and 85, and the mean age of patient cohort was found to be 68 years. Among the dataset, 74 patient cohorts experienced recurrence whereas 26 patients did not have any relapse. Overall, the tumor size of the patients ranges from 10 mm to 110 mm respectively. Interestingly, it is to be noted that patients with tumor size less than 31.97 ± 16.49 mm did find to have relapse of cancer cells. On the other hand, patients with tumor size greater than 52.38 ± 26.57 mm were found to experience relapse of cancer. Moreover, 93% of the patient population was distributed among invasive adenocarcinoma and squamous cell carcinoma histology. On the other hand, the remaining 7% of the population was distributed among carcinosarcoma, adenosquamous carcinoma and large cell carcinoma. Moreover, the packyears in the dataset ranged between 0.05 and 118. In addition, the dataset contained records of 12 never smokers, 7 current smokers, 48 ex-smokers and 33 recent ex-smokers, respectively. Only 20% of the patient population in study had obtained adjuvant treatment. It is to be noted that all the patients who had undertaken adjuvant therapy had survived in the dataset.

Performance Evaluation

Totally 15 predictor variables and 1 target variable were considered for model generation and validation. The performance metrics such as accuracy, ROC curve and area under the curve (AUC) were used to evaluate the classifiers. Accuracy is a reliable evaluation parameter for the ML models as an equalized sample number of each feature is required to determine the model accuracy.[25] In our study, the decision tree algorithm had the greatest accuracy of 0.95 while the other three algorithms had significantly lower accuracies in the range of 0.7. This indicates that the decision tree model as the successful model in stratifying the prediction into relapse or non-relapse during our analysis. On the other hand, the ROC and AUC curve are used to study the ratio of true positive rate to that of false positive rate. It is evident from the Table 2 that all the four models had an AUC value of greater than 0.8. In specific, the AUC of random forest was the highest in the training set (0.96) as well as the test set (0.96). It is worth mentioning that decision tree also demonstrated comparatively equivalent result in AUC to that of random forest. It is evident from Figure 1 that higher AUC of random forest and decision tree model indicates the significance of a good classifier with low false-positive rates and a greater proportion of true positive rates than the other models studied.

Model Evaluation using other parameters

Confusion matrix is used to evaluate the quality and performance of developed models. Figure 2 demonstrates the confusion matrix of each algorithm. The other performance metrics including F1 score, precision, recall, sensitivity and specificity were evaluated using the developed confusion matrix during our analysis. In our study, the precision of the models ranges from 0.4 and 1.0. The lowest precision belonged to logistic regression and the remaining models were found to have equivalent precision value above 0.9. Nevertheless, the recall values were moderate and varies between 0.17 and 0.9 for the test sets with the lowest value and the highest were achieved by random forest and decision tree respectively. In general, F1 score is the harmonic mean of precision and recall which are important evaluators of a ML model and can be used to evaluate the predictive ability of the models.[26] Importantly, the decision tree algorithm had the highest F1 score of 0.9 indicating that the model has high precision and recall than the other models. On the other hand, the F1 score of the other models were 0.5 or lower indicating that they had low values of either precision, recall or both. In terms of specificity, all the models were found to be above 0.95 whereas the specificity raged between 0.16 and 0.75 with the lowest and highest belonging to random forest and decision tree respectively. Hence, analysis of each classifier's confusion matrix displayed an agreeable relationship between a classifier's actual inputted values and the predicted values. Figure 3 represents the nodes and leaves of the best classifier decision tree.

Validation of generated model

In order to validate the outcome and performance of the developed models, 5-fold and 10-fold cross validation strategy was implemented in the current investigation. As accuracy serves as a significant parameter, it was used to compare the model's performance. In addition, F1 score was also calculated as it serves a benchmarking metric by combining precision and recall during evaluation. It is interesting to note from Table 3 that the chronological order of performance of the model

Table 1: Baseline characteristics of Caribbean and few white ethnicities dataset.			
Recurrence	**No**	**Yes**	***P*-value**
N	74	26	
Stage			<0.001
1a	23(31.10%)	3(11.54%)	
1b	32(43.24%)	4(15.38%)	
2a	8(10.81%)	5(19.23%)	
2b	5(6.75%)	6(23.08%)	
3a	6(8.10%)	7(26.92%)	
3b	0(0%)	1(3.85%)	
Age	68.55 +- 7.77*	67.92+- 11.58*	0.769
Gender			0.683
Male	45(60.81%)	17(65.38%)	
Female	29(39.19%)	9(34.62%)	
Ethnicity			0.024
Caribbean	2(2.70%)	1(3.85%)	
White-British	68(91.89%)	21(80.77%)	
White-Irish	4(5.41%)	0(0%)	
White-Other	0(0%)	4(15.38%)	
Histology			0.921
Adenosquamous carcinoma	1(1.35%)	2(7.69%)	
Carcinosarcoma	0(0%)	2(7.69%)	
Invasive adenocarcinoma	49(66.22%)	12(46.15%)	
Large cell carcinoma	1(1.35%)	0(0%)	
Large Cell Neuroendocrine	0(0%)	1(3.85%)	
Squamous cell carcinoma	23(31.08%)	9(34.62%)	
Tumor Size	31.97 +- 16.49*	52.38+-26.57*	<0.001
Resection margins			0.076
R0	72(97.30%)	23(88.46%)	
R1	2(2.70%)	3(11.54%)	
Vascular invasion			0.447
No	42(56.76%)	17(65.38%)	
Yes	32(43.24%)	9(34.61%)	
Pleural invasion			0.992
No	54(72.97%)	19(73.08%)	
Yes	20(27.03%)	7(26.92%)	
Adjuvant therapy			0.718
No adjuvant treatment	54(72.97%)	18(69.23%)	
Adjuvant	20(27.03%)	8(30.77%)	
ECOG			0.907
0	38(51.35%)	13(50%)	
1	36(48.65%)	13(50%)	
Smoking status			0.319
Never Smoked	9(12.16%)	3(11.54%)	
Current Smoker	2(2.70%)	5(19.23%)	
Recent Ex-Smoker	26(35.14%)	7(26.92%)	
Ex-Smoker	37(50%)	11(42.31%)	
Pack years	37.93+-29.19*	34.45+-24.06*	0.586
Genome doubled			0.489
Not GD	18(24.32%)	6(23.08%)	
Clonal GD	55(74.32%)	18(69.23%)	
Subclonal GD	1(1.35%)	2(7.69%)	
Time to recurrence or death (months)	19.18+-6.58*	10.35+-6.54*	<0.001

* - Mean ± Standard deviation for continuous variable; P-value – Probability value (<0.001 are significant)

Table 2: Forecast Results of the developed machine learning models.							
Algorithms	**Accuracy**	**Precision**	**Recall**	**Sensitivity**	**Specificity**	**F1_score**	**AUC**
Logistic Regression	0.77	0.4	0.67	0.33	0.96	0.5	0.73
Decision Tree	0.95	0.9	0.9	0.75	0.95	0.9	0.93
Random forest	0.71	1.0	0.17	0.16	1.0	0.29	0.96
Support vector machine	0.74	1.0	0.25	0.25	1.0	0.4	0.83

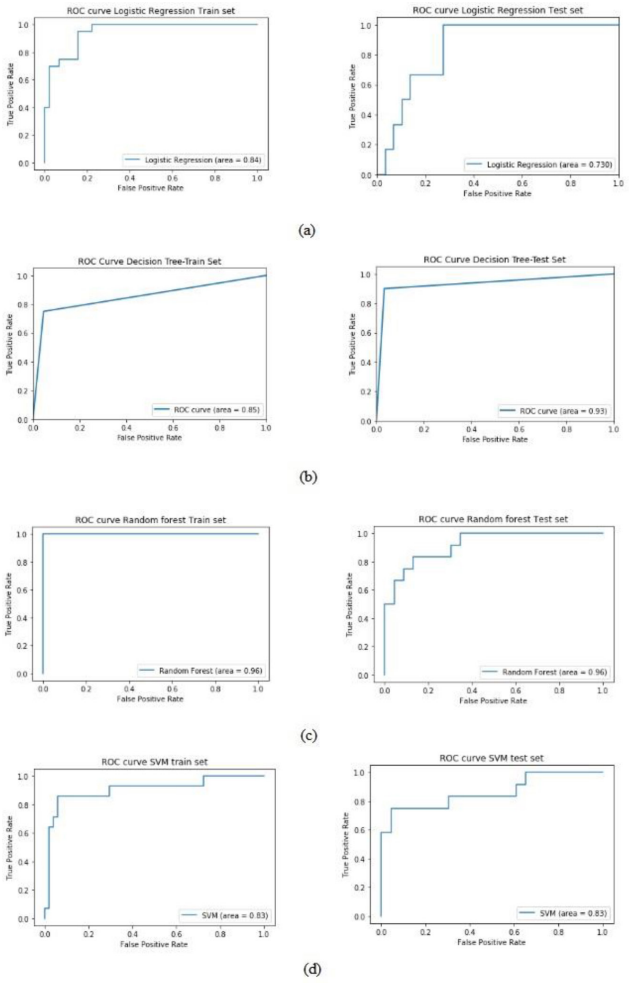


Figure 1: ROC curve of training set and test set of (a) Logistic regression (b) Decision tress (c) Random Forest and (d) Support vector models generated using all the features of the dataset.

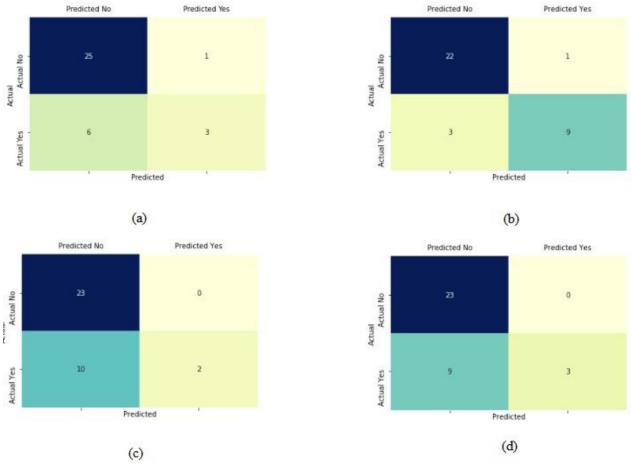


Figure 2: Confusion Matrix for a) Logistic Regression b) Decision Tree c) Random Forest and d) Support Vector Machine.

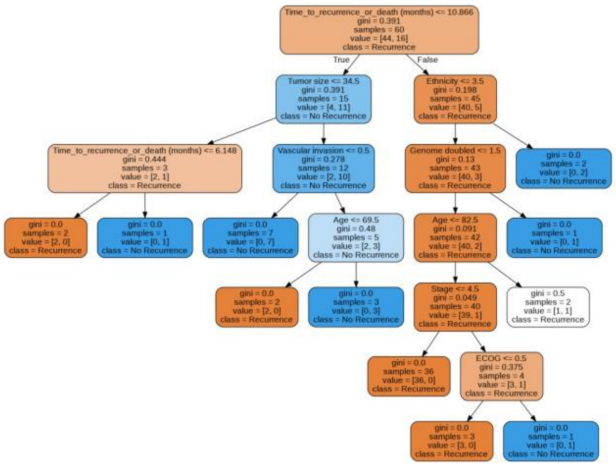


Figure 3: Decision Tree with important nodes and leaves.

during cross validation is similar to that of the test phase. For instance, on comparing the models based on accuracy, decision tree outperforms all other models followed by logistic regression, support vector machine and random forest. It is worth mentioning that decision tree outperformed the other models during 5-fold and 10-fold cross validation with highest accuracy of 0.91 and 0.96 respectively. Thus, our cross-validation results demonstrate the ability of models to predict lung cancer recurrence in patients.

In addition, harness of the models was evaluated and the results are represented in Figure 4. This process involves resampling method for splitting the dataset, machine learning method to evaluate and the performance metric.[27] In the present study, the process resulted the following accuracy (standard deviation): logistic regression – 0.78 (0.107), decision tree – 0.83 (0.09), support vector machine – 0.79 (0.144) and random forest – 0.78 (0.14) respectively. It is interesting to note from the results that the decision tree outperformed

Table 3: Cross validation analysis of the developed machine learning models.

Algorithms	5-Fold Cross Validation		10-Fold Cross Validation	
	Accuracy	F1_score	Accuracy	F1_score
Logistic Regression	0.79	0.54	0.84	0.62
Decision Tree	0.91	0.88	0.96	0.90
Random forest	0.65	0.37	0.79	0.54
Support vector machine	0.78	0.45	0.81	0.62

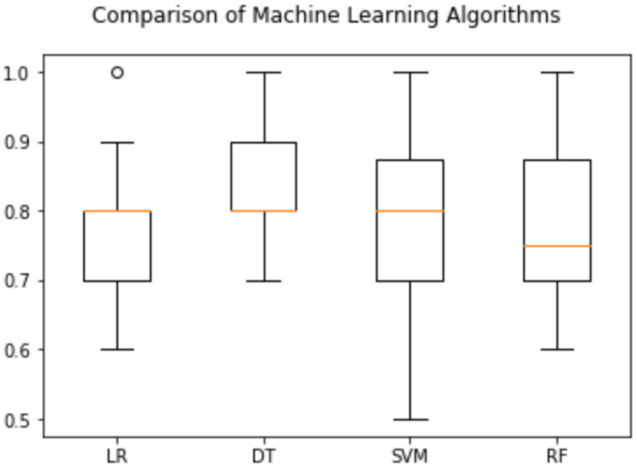


Figure 4: Analysis of test harness of all the machine learning models.

among all the other models with highest accuracy and minimal standard deviation.

Statistical analysis to predict significant features

From the correlation matrix, Figure 5, it can be noted that Stage, Ethnicity, Tumor Size and Resection Margins had a positive correlation to recurrence. Adjuvant therapy, Genome Doubled and ECOG are seen to have a weak positive correlation with recurrence. While, Age, Gender, Vascular invasion, Smoking status, Pack years and the time to recurrence (months) had a negative correlation to Recurrence instances. Pleural Invasion and Histology are shown to have a weak negative correlation with recurrence calculated in our analysis. By considering the features with positive correlation in the correlation matrix as well as *p*-values, five features can be concluded as being important and influential for the evaluation of recurrence namely, Stage, Ethnicity, Tumor Size, Resection Margins and Genome Doubled.

Considering the experience of recurrence as the key factor in the field of biomedical, a classifier with higher sensitivity and accuracy is preferred. So, a successful classifier must be capable of forecasting a potentially

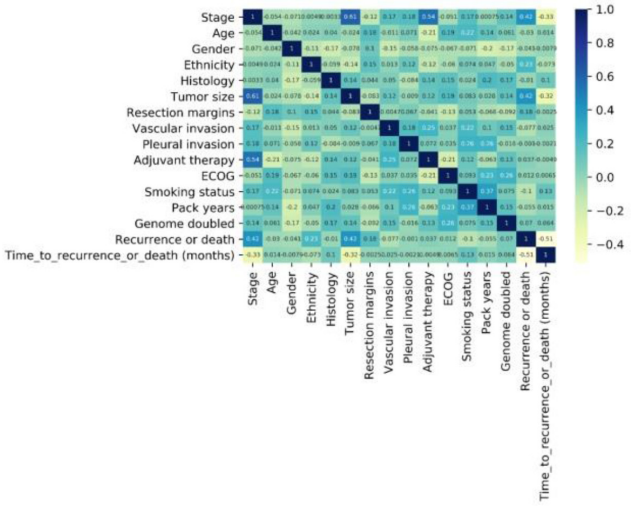


Figure 5: Correlation analysis of factors influencing recurrence.

future metastatic patient using the independent variables. Several studies have analyzed the reliability of different classifiers to estimate the outcome of interest. For instance, Alabi *et al.* performed a study to predict recurrence in oral tongue cancer patients using different supervised machine learning algorithms. They compared the performance of four different algorithms, including support vector machine, naïve bayes, boosted decision tree, and decision forest, to predict the relapse in oral cancer patients.[28] Our finding showed that decision tree outperformed than other machine learning models in predicting the metastasis of the lung cancer patients in our investigation.

CONCLUSION

The present investigation focuses on assessing the performance of four machine learning techniques in predicting relapse in patients with lung cancer. Our finding achieved highest accuracy of 95% using decision tree classifier in stratifying the recurrence status of lung cancer patients belonging to Caribbean and few white ethnicities. Moreover, cancer stage, ethnicity, tumor size, genome doubled and time to recurrence are the statistically significant features identified during analysis. Importantly, it is to be noted that the time and cost for collecting the patient's data is comparatively higher. Thus, the key features proposed in our analysis certainly helpful not only to manage cost, time, and resource but also, assist the clinicians in terms of decision management.

ACKNOWLEDGEMENT

The authors thank VIT for providing 'VIT SEED GRANT' for carrying out this research work.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

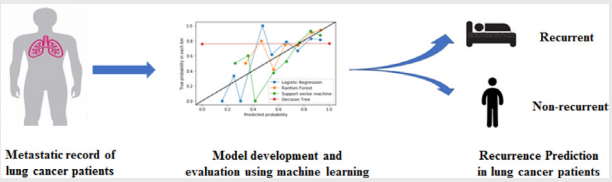
ABBREVIATIONS

AUC: Area Under the Curve; **ML:** Machine Learning; **NSCLC:** Non-Small Cell Lung Cancer; **ROC:** Receiver Operating Curve; **SVM:** Support Vector Machine.

REFERENCES

1. Suresh R, Ali S, Ahmad A, Philip PA, Sarkar FH. The role of cancer stem cells in recurrent and drug-resistant lung cancer. Adv Exp Med Biol. 2016;890:(57-74). doi: 10.1007/978-3-319-24932-2_4, PMID 26703799.
2. Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, *et al*. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell. 2012 Sep 14;150(6):1107-20. doi: 10.1016/j.cell.2012.08.029, PMID 22980975.
3. Duma N, Santana-Davila R, Molina JR. Non–small cell lung cancer: Epidemiology, screening, diagnosis and treatment. InMayo Clinic Proceedings 2019 Aug 1. Elsevier.
4. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. Transl Lung Cancer Res. 2014 Aug;3(4):242-9. doi: 10.3978/j.issn.2218-6751.2013.12.05, PMID 25806307.
5. Ricardi U, Badellino S, Filippi AR. Stereotactic body radiotherapy for early-stage lung cancer: history and updated role. Lung Cancer. 2015 Dec 1; 90(3):388-96. doi: 10.1016/j.lungcan.2015.10.016, PMID 26791797.
6. Qian J, Xu J, Wang S, Qian F, Yang W, Zhang B, *et al*. Adjuvant chemotherapy candidates in stage I lung adenocarcinomas following complete lobectomy. Ann Surg Oncol. 2019 Aug 15;26(8):2392-400. doi: 10.1245/s10434-019-07366-z, PMID 31011907.
7. Soto SV, Luna J, Cano A, editors. Big data on real-world applications. BoD–books on demand; 2016 Jul 20.
8. Lei L, Wang Y, Xue Q, Tong J, Zhou CM, Yang JJ. A comparative study of machine learning algorithms for predicting acute kidney injury after liver cancer resection. PeerJ. 2020 Feb 25;8:e8583. doi: 10.7717/peerj.8583, PMID 32140301.
9. Hasnain Z, Mason J, Gill K, Miranda G, Gill IS, Kuhn P, *et al*. Machine learning models for predicting post-cystectomy recurrence and survival in bladder cancer patients. PLOS ONE. 2019 Feb 20;14(2):e0210976. doi: 10.1371/journal.pone.0210976, PMID 30785915.
10. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, *et al*. Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation. Cancer Med. 2020 May;9(9):3234-43. doi: 10.1002/cam4.2811, PMID 32154669.
11. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: Management, analysis and future prospects. J Big Data. 2019 Dec 1;6(1):54. doi: 10.1186/s40537-019-0217-0.
12. Ganggayah MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Med Inform Decis Mak. 2019 Dec 1;19(1):48. doi: 10.1186/s12911-019-0801-4, PMID 30902088.
13. Tseng YJ, Wang HY, Lin TW, Lu JJ, Hsieh CH, Liao CT. Development of a machine learning model for survival risk stratification of patients with advanced oral cancer. JAMA Network Open. 2020 Aug 3;3(8):e2011768-. doi: 10.1001/jamanetworkopen.2020.11768, PMID 32821921.
14. Huang Y, Chen H, Zeng Y, Liu Z, Ma H, Liu J. Development and validation of a machine learning prognostic model for hepatocellular carcinoma recurrence after surgical resection. Front Oncol. 2020;10:3327. doi: 10.3389/fonc.2020.593741, PMID 33598425.
15. Lou SJ, Hou MF, Chang HT, Chiu CC, Lee HH, Yeh SJ, *et al*. Machine learning algorithms to predict recurrence within 10 years after breast cancer surgery: A prospective cohort study. Cancers. 2020 Dec;12(12):3817. doi: 10.3390/cancers12123817, PMID 33348826.
16. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, *et al*. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep. 2017 Aug 7;7(1):7402. doi: 10.1038/s41598-017-07408-0, PMID 28784991.
17. Xu Y, Ju L, Tong J, Zhou CM, Yang JJ. Machine Learning Algorithms for predicting the Recurrence of Stage iV colorectal cancer After tumor Resection. Sci Rep. 2020 Feb 13;10(1):2519. doi: 10.1038/s41598-020-59115-y, PMID 32054897.
18. Halabi A, Kenett RS, Sacerdote L. Modeling the relationship between reliability assessment and risk predictors using Bayesian networks and a multiple logistic regression model. Qual Eng. 2018 Oct 2;30(4):663-75. doi: 10.1080/08982112.2017.1368556.
19. Paredes AZ, Hyer JM, Tsilimigras DI, Moro A, Bagante F, Guglielmi A, *et al*. A novel machine-learning approach to predict recurrence after resection of colorectal liver metastases. Ann Surg Oncol. 2020 Dec;27(13):5139-47. doi: 10.1245/s10434-020-08991-9, PMID 32779049.
20. Gastwirth JL. Is the Gini index of inequality overly sensitive to changes in the middle of the income distribution? Stat Public Policy. 2017 Jan 1;4(1):1-. doi: 10.1080/2330443X.2017.1360813.
21. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clinical Epidemiology and Global Health. 2019 Sep 1;7(3):293-9.
22. Monferrer E, Burgos-Panadero R, Blanquer-Maceiras M, Cañete A, Navarro S, Noguera R. High Oct4 expression: Implications in the pathogenesis of neuroblastic tumours. BMC Cancer. 2019 Dec;19(1):1. doi: 10.1186/s12885-018-5219-3, PMID 30606139.
23. Wang J, Yu H, Hua Q, Jing S, Liu Z, Peng X, *et al*. A descriptive study of random forest algorithm for predicting COVID-19 patients outcome. PeerJ. 2020 Sep 9;8:e9945. doi: 10.7717/peerj.9945, PMID 32974109.
24. Ramesh P, Veerappapillai S. Prediction of micronucleus assay outcome using *in vivo* activity data and Molecular Structure features. Appl Biochem Biotechnol. 2021 Dec;193(12):4018-34. doi: 10.1007/s12010-021-03720-8, PMID 34669110.
25. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015 Jan 1;13:8-17. doi: 10.1016/j.csbj.2014.11.005, PMID 25750696.
26. Li YM, Jiang LC, He JJ, Jia KY, Peng Y, Chen M. Machine learning to predict the 1-year mortality rate after acute anterior myocardial infarction in Chinese patients. Ther Clin Risk Manag. 2020;16:1-6. doi: 10.2147/TCRM.S236498, PMID 32021220.
27. Crumpei-Tanasă I, Crumpei I. A machine learning approach to predict stress hormones and inflammatory markers using illness perception and quality of life in breast cancer patients. Curr Oncol. 2021 Aug;28(4):3150-71. doi: 10.3390/curroncol28040275, PMID 34436041.
28. Alabi RO, Elmusrati M, Sawazaki-Calone I, Kowalski LP, Haglund C, Coletta RD, *et al*. Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. Int J Med Inform. 2020 Apr 1;136:104068. doi: 10.1016/j.ijmedinf.2019.104068.

PICTORIAL ABSTRACT



SUMMARY

Recently, ML has increasingly begun to take root in the oncology world to build models for forecasting cancer progression and survivability. This study explored the use of four ML classifiers to predict the recurrence in lung cancer patients. Among them, decision tree classifier was accurate in stratifying the recurrence status of the patients. Importantly, as the time taken and cost for collecting the patient's data is high, the features used in our analysis is certainly helpful not only to manage cost, time and resource but also assist the clinicians in terms of decision management. Overall, we conclude that the predictive models, based on the combination of scientific evidence and personal experience, may support but will not substitute the physician's recommendations.

Cite this article: Ramesh P, Jain A, Karuppasamy R, Veerappapillai S. Exploring Machine Learning Models for Recurrence Prediction in Lung Cancer Patients. Indian J of Pharmaceutical Education and Research. 2022;56(3s):s398-s406.