International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Predicting Affect and Survivability of Lung Cancer Using Machine Learning

Jai Agarwal[1], Sooriya Rangan[2], G Merlin Linda[3]

[1,2]UG Student, [3]Assistant Professor
[1,2,3]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Faculty of Engineering and Technology, Vadapalani Campus, Chennai, India

---

Abstract **:** *Lung cancer is caused when cells divide in the lungs uncontrollably. These can reduce a person's ability to breathe with both inhale and exhale. Direct Cigarette smoking and passive smoking are the Main contributor for Lung Cancer as per WHO. The mortality rate is increasing every day in youths as well as in old persons due to lung cancer as compared to other cancers. In this study, Machine learning algorithms are used for Predicting lung cancer survival on Surveillance, Epidemiology, and End Results (SEER) and to predict weather patient is affected or not affected on Kaggle dataset. Several pre processing steps were done on data before applying Classification Algorithms. Classification Algorithms used on both the Datasets are Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine(SVM). Also Neural Network and k-Nearest Neighbor (k-NNN) are two additional algorithms used on SEER Dataset. The efficiency of the algorithm is calculated in terms of accuracy for SEER Dataset and Prediction of Affect is done using accuracy and Specificity for Kaggle Dataset. Logistic Regression and Random Forest showed the best result for Kaggle Dataset and Random Forest and Naïve Bayes for SEER Dataset. These Machine learning-based technique help in predicting the lung cancer using supervised classification machine learning algorithms accurately. And Finally GUI based Interface is used for Kaggle dataset and Orange software is used for SEER dataset.*

Keywords: Dataset, Machine learning, Classification, Python, Lung Cancer, Survivability, Affect.

I. INTRODUCTION

In the US, lung cancer is the second of almost all the common cancers that occur in both men and women. After 5 Years of Diagnosis Survival rate of Lung cancer is approximated to be 15%. One of the most popular topics of medical research is survival analysis. Return of disease for specific time period or death are shown by the use of a predictor variable to predict cancer survival. Predictor models can estimate how long a patient will live after being diagnosed.

One out of every four deaths in 2012 was generally attributed to cancer, with an overall survival rate of 10-15% according to ACS. Lung cancer is the most common type, with high rates of ephemerality due to smoking and pollution. Though both men and women might have prostate and breast cancer, lung cancer has a higher mortality rate. Chemotherapy, radiotherapy, and operations are just a few of the prevention and recovery options available. The majority of patients are diagnosed at an advanced stage around the world. As the Symptoms are not much visible it has become difficult to diagnose even by the doctors in early stage.

For this study, data from SEER and Kaggle were examined for lung cancer survival prediction and weather the patient is affect or not affected. Data is collected from different source and area in USA by SEER. Collection of data with few records began in 1973 and has kept on growing by adding more features and Values. SEER includes survival and cancer incidence data of approximately 30% of US population from population-based cancer registries. Primary tumor location, vital condition follow-up, stage at diagnosis, tumor pathology, and Patient demographics are all regularly collected data in the SEER programme registries. The SEER data consist of nine text files, each containing information about a different anatomical. Each file has 149 attributes, and every record is linked to a specific cancer incidence.

As a field or part of Artificial Intelligence, Machine learning retrieves new conclusion and

results from an already Programmed Algorithm with the help of past experience, with improved accuracy. Out of many learning algorithms in Machine Learning such as association, regression, classification etc. that can be used according to the need for predicting best accuracy as close to human predictions. Type of data with which we are currently working decides the type of algorithm on which data can be operated. There are inbuilt libraries that support the Machine Learning tools and also writing languages to realizing the proposals.

Any models accuracy prediction capabilities depends on its learning capabilities which indeed depends on correctness of the training set.

II. PROPOSED SYSTEM

In this study paper different machine learning algorithms is used for lung cancer survival prediction, on SEER Dataset and weather the patient is affected or not affected on Kaggle dataset. Four Machine learning Algorithms including Logistic Regression, Naïve bayes, Support Vector machine (SVM) and random Forest, were used on both SEER and Kaggle datasets and additionally two extra algorithms including k-Nearest Neighbor(k-NN) and Neural Network were used on SEER dataset. Effect of Lung Cancer is predicted using Kaggle dataset and Survivability of patient is predicted using seer dataset.

A. Data Preprocessing

I. SEER DATASET

Pre-processing Data in Machine learning is an important step as it can increase the performance of classification algorithm notably. In this study, SEER data obtained in 2019 (SEER_1975_2016_TEXT DATA.d04132019) were used. The RESPIR.TXT file which had 149 attributes and 645,682 samples were used for predicting lung cancer survival analysis. Various data processing, cleaning and modifications were done before implementing classification method. Steps below show the preprocessing of data:

• Data of patients diagnosed between the year 1998 and 2001 were selected. There where many key attributes which were included to SEER dataset since 1998 and additionally there were modification applied to SEER dataset after 2002, so the Patients data were limited.

• Patients who died from causes other than cancer had their causes of death removed.

• The class variable is the Survival month attribute.

• Patients with unspecified cause of death were removed.

• Patients with survival months more than 60 are marked as Survived and those with survival month less than 60 are marked as not-survived.

• The patients who survival month is greater than 60 with cause of death other than cancer and the patient whose survived months are less than 60 and are alive are all excluded from data.

• Patient ID number along with Insurance records and many more which are irrelevant attributes are removed.

• Features such as EOD – Extension Proth Path and Breast Subtypes which are not in relation with lung cancer are excluded from dataset.

After preprocessing and data cleaning there are 23 predicting attributes and 45,215 samples left in the dataset. Table I shows selected attributes. Imbalance problem in dataset is shown in Table II. Random Undersampling is used to handle the imbalance data problem. The sample that each class contain after undersampling is 5977.

TABLE I. SELECTED ATTRIBUTES FOR SEER

Selected attributes
Race / Ethnicity
Marital Status at DX
Primary Site
Histologic Type ICD-O-3
Behavior Code ICD-O-3
Grade
EOD—Extension
EOD—Lymph Node Involv
Summary stage 2000 (1998+)
RX Summ—Surg Prim Site
EOD—Tumor Size
Regional Nodes Positive
Regional Nodes Examined
NHIA derived hispanic origin
Reason for no surgery
Diagnostic Confirmation
Sequence Number—Central

Registry ID
Age at diagnosis
Sex
Laterality
AJCC stage 3rd edition (1988-2003)

TABLE II. CLASS DISTRIBUTION

Class	No. of Samples
Not-Survived	45,215
Survived	5,977

- Three kinds of attribute are present in SEER dataset and they are ordinal attributes, numerical attributes and categorical attributes. When the data is in numerical type, Machine Learning algorithms like Logistic Regression can function correctly, so type conversion is used to convert categorical and ordinal attributes to numerical.

II. KAGGLE DATASET.

Pre-processing Data in Machine learning is an important step as it can increase the performance of classification algorithm notably. In this work Kaggle dataset is been used. The File Contain 11 Attributes and 1298 samples. The Preprocessing steps are as follows:

a. The dataset is first analyzed for features visualization.
b. Dataset is then analyzed to find which attributes are Dependent and Independent.
c. Features with no Impact on our data were removed for the dataset and remaining features are taken as attributes for out dataset.
d. Data was analyzed to find any duplicate Samples and all the duplicates were removed.
e. Correlation among the attributes were found.
f. Data was Visualized in many graphs types for better understanding of data.
g. Machine Learning algorithms like Logistic Regression can function with Correctness only when the data are in numerical type so transforming data samples numerical is done by type conversion.
h. A Simple GUI application was build to show weather the patient is Affected or Not Affected.

TABLE III. SELECTED ATTRIBUTES FOR KAGGLE

Selected attributes
Name
Member Id
Diagnosis
Age
Smokes
Smokes (Years)
Smokes (Packs/Year)
AraeQ
Alkhol
Family History
Result

B. Experiments

I. SEER DATASET

Six Machine Learning Algorithms Logistic Regression, Naïve Bayes, Random Forest, k-Nearest Neighbor (KNN), Neural Network, and Support Vector Machine(SVM) were used for Lung Cancer Patients survival prediction. Diversification was generated by using different subsets to train data. All the algorithms which were applied were run using default parameters. Python was used for all the Preprocessing and Orange Software Tool for Machine Learning and Data Mining was used for building models. Datasets received in year 2019 from SEER was used in this study. After preprocessing, the dataset with 149 attributes and 645,682 samples, was reduced to 23 attributes and 11,954 Samples. The tests were carried out on a computer with an Intel Core i5 processor running at 1.80 GHz and 8 gigabytes of RAM. For training and testing, 10-fold cross-validation was used to divide the dataset into ten subsets of almost same size, and then the algorithms were trained on nine subsets, and the last one was used to evaluate the result. The above Process was repeated ten times on various sets to train and test results, and the final performance is evaluated by averaging of all the process that took place..

II. KAGGLE DATASET

Four Machine Learning Algorithms Logistic Regression, Naïve Bayes, Random Forest and Support Vector Machine (SVM) were used for prediction of Lung Cancer Patients Survival. To generate diversity, both of these algorithms use different subsets of training data. The process started from collecting lung cancer dataset from Kaggle. Python, Anaconda's Jupyter Notebook and many more machine Learning Libraries were used for Preprocessing and for building model. There were 11 attributes and 1298 samples. The tests were carried out on a computer with an Intel Core i5 processor running at 1.80 GHz and 8 gigabytes of RAM. The dataset is then split into training and testing using the sklearn's Method then trained data is fit and model is trained, the predicted output is calculated. The model which performed better is used to predict the output and is used for Simple GUI Development to display weather the Patient is Affected or Not Affected.

C. Performance Evaluation Method

To compare and evaluate the Area under ROC curve (AUC) and Accuracy, are used to compare and evaluate the performance of different algorithms.

1) Accuracy: The percentage of accurate predictions by total number of predictions that were made.

$$FREQUENCY = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100$$

- True positive (TP): In True Positive the algorithms output predicts the positive classes correctly.
- True negative (TN): In True Negative the algorithms output predicts the negative classes correctly.
- False positive (FP): In False positive the algorithms output predicts the positive classes incorrectly.
- False negative (FN): In False negative the algorithms output predicts the negative classes incorrectly.

2) Area Under ROC Curve (AUC): Binary Classifier is a common metric for evaluating the AUC, that represents the tradeoff between false positive rate and detection rate by plotting them on a curve with different parameter values.

III. EXPERIMENTAL RESULT

The output was assessed using accuracy and the area under the ROC curve metrics of Machine learning Algorithm on SEER dataset for cancer survival prediction. TABLE IV Shows the result and indicate that Random Forest and Naïve Bayes gives the maximum accuracy compared to other algorithms. The accuracy of Random Forest is 85.1% and accuracy of Naïve Bayes 84.4%. The AUC of Random Forest is found to be 91.8% and AUC of Logistic Regression is found to be 90.1%.

Accuracy and Specificity are used to evaluate the performance of Machine learning Algorithm on Kaggle dataset for cancer survival prediction. TABLE V shows the result and indicates that Random Forest and Logistic Regression gives Maximum Accuracy compared to other algorithms. The accuracy of Random forest and Logistic regression is found to be 100%. The Specificity of both Random Forest and Logistic Regression is also 100% and this is because of small amount dataset.

By Comparing the results from Accuracy and AUC Metric for SEER Dataset we have seen that Random Forest and Naïve Bayes have performed best compared to other Machine Learning algorithms like k-NN with accuracy 88.2%, SVM with accuracy 66.1%, Neural Network with accuracy 84%, Logistic Regression with accuracy 83.7%. Among all of these algorithms we can conclude that SVM has the worst accuracy and AUC along with other metrices as shown in TABLE IV.

By Comparing the results from Accuracy and Specificity Metric for Kaggle Dataset we have seen that Random Forest and Logistic Regression have performed best compared to other Machine Learning algorithms like SVM with accuracy 98.25, Naïve Bayes with accuracy 98.40.

IV. CONCLUSION

Lung cancer, most common cancer, is studied in this paper. For survival prediction, a dataset from the SEER programme for lung cancer was analysed. After preprocessing, the dataset with 149 attributes and 645,682 samples, was reduced to 23 attributes and 11,954 Samples. Random

Undersampling is used to handle the imbalance data problem. The sample that each class contain after undersampling is 5977. Now the data was ready for processing. For predicting survival of lung cancer, six Machine Learning algorithms were tested. For both accuracy and the AUC metrics, Random Forest was better than other algorithms. In contrast to the other Machine Learning methods used in this analysis, the SVM algorithm was the worst. The findings of this research can be used to improve the efficiency of survival prediction systems in the future. Dataset from Kaggle was also used for Predicting weather the patient is affected or not affected with lung cancer. Kaggle data file contained 11 Attributes and 1298 samples. Data was preprocessed and visualized with different graphs to under its attributes. Four Machine Learning Algorithms were evaluated for predicting weather the patient is affected or not affected. Random Forest and Logistic Regression Gave the best accuracy and specificity. The findings of this research can be used to improve the efficiency of survival prediction systems in the future.

V. REFERENCES

[1] "Lung Cancer Statistics," Centers for Disease Control and Prevention, Available: http://www.cdc.gov/cancer/lung.

[2] R. LAG, J. L. Young, G. E. Keel, M. P. Eisner, Y. D. Lin, and M. J. Horner, "SEER Survival Monograph, Cancer Survival Among Adults: US SEER Program, 1988-2001, Patient and Tumor Characteristics," National Cancer Institute NIH Pub., 2007.

[3] "About the SEER Program," National Cancer Institute. [Online]. Available: http://seer.cancer.gov/about/overview.html.

[4] R Javidan, M. A. Masnadi-Shirazi, and Z. Azimifar, "Contourlet-Based acoustic seabed ground discrimination system ," 3rd International Conference on Information and Communication Technologies: From Theory to Applications, IEEE, 2008.

[5] H. M. Zolbanin, D. Delen, and A. Hassan Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," Decis. Support Syst., vol. 74, pp. 150–161, 2015.

[6] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability : a comparison of three data mining methods," Artif. Intell. Med., vol. 34 No.2, pp.113-27, 2005.

[7] B. Zheng, S. W. Yoon, and S. S. Lam, "Expert Systems with Applications Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Syst. Appl., 2013.

[8] J. W. Grzymala-busse, L. K. Goodwin, and X. Zhang, "Handling Missing Attribute Values in Preterm Birth Data Sets," Springer-Verlag Berlin Heidelberg, pp. 342–351, 2005.

[9] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," PhD Thesis, University of Waikato, April 1999.

[10] S Russell, and P. Norvig, Artificial Intelligence a Modern Approach, 3rd Ed., Pearson Edition, 2009.

[11] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, vol. 72, Springer International Publishing, 2015.

[12] Y. Freund, "Boosting a weak learning algortihm by majority," Inf. Comput., vol. 121, no. 2, pp. 256–285, 1995.

[13]T. K. Ho, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 8, pp. 832–844, 1998.

[14] W. W. Cohen, M. Avenue, M. Hill, C. Of, and R. Pruning, "Fast Effective Rule Induction," Proc. Twelfth Int. Conf. Mach. Learn., vol. 3, pp. 115–123, 1995.

[15] I. H. Witten, E. Frank, and M. a. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, vol. 54, no. 2. 2011.

[16] J. C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," Adv. kernel methods, pp. 185 – 208, 1998.

[17] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd Edition, Morgan Kaufmann, 2011.

[18] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[19] https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/loading-your-data/index.html.

TABLE IV. CLASSIFICATION ACCURACY, AUC and MORE

Model	AUC	CA	F1	Precision	Recall
kNN	0.882	0.828	0.828	0.828	0.828
SVM	0.692	0.661	0.658	0.667	0.661
Random Forest	0.918	0.851	0.851	0.851	0.851
Neural Network	0.910	0.842	0.842	0.843	0.842
Naive Bayes	0.907	0.844	0.844	0.844	0.844
Logistic Regression	0.901	0.837	0.837	0.837	0.837
AdaBoost	0.870	0.834	0.834	0.835	0.834

TABLE V. CLASSIFICATION ACCURACY AND SPECIFICITY

Algorithm	Logistic Regression	Naïve Bayes	Random Forest	Support Vector Machine
Accuracy	100	98.40	100	98.25
Specificity	100	95.6	100	0.956