e-ISSN 2320 –7876 www.ijfans.org
Vol.11, S Iss 1, 2022

Machine Learning Methods for Detecting Radiation-Induced Tissue Inflammation in Patients with Lung Cancer

Prashant Agrawal[1*], Keshav Kumar K[2], Pallavi sagar Deshpande[3], Parthiban K G[4], Abdul Rahman Mohammed ALAnsari[5], J.Sundararajan[6], Mohammed Siddique[7]

[1]Department of Computer Applications, KIET Group of Institutions, Delhi NCR Ghaziabad, Uttar Pradesh, India

[2]Department of Humanities and Mathematics, G.Narayanamma Institute of Technology and Science (for Women), Shaikpet, Hyderabad, Telangana, India

[3]Department of Electronics and telecommunication, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune,Maharashtra, India

[4]Department of Biomedical Engineering, Dhaanish Ahmed Institute of Technology Coimbatore, Tamil Nadu, India

[5]Department of Surgery, Salmanyia Hospital, Manama, Bahrian

[6] Department of Electronics and Communications Engineering, NPR College of Engineering and Technology, Natham, Dindugul, Tamil Nadu, India

[7]Department of Mathematics, Centurion University of Technology and Management, Odisha, India

***Corresponding mail** id: prashant.agraw@gmail.com

ABSTRACT

Patients undergoes radiation therapy as measure of their lung cancer are at hazard of developing radiation pneumonitis, a condition caused by lung radiation damage (RP). RP is a theoretically terminal adverse influence of medicine. As a result, new strategies for guiding clinicians in administering personalized treatment doses to individuals at high risk of RP are necessary. Several prediction models have been constructed utilizing machine learning and traditional statistical processes, but no explanation for performance variances has been provided. In this study, we analyze a variety of well-known organization algorithms in the field of deep learning in order to identify several RP risk categories. The usefulness of these cataloguing algorithms is evaluated in combination with different segment assortment approaches, and the influence of technique collection on routine is then estimated further.

Keyword: Lung radiation, machine learning, radiation therapy, Algorithm

1. Introduction

Cancer is one of the worst sicknesses for human being, with just a 15% five-year survival rate, and it is the foremost reason of cancer death. Non-small cell lung cancer accounts for around 8 to 9% of all lung cancer cases (NSCLC) [1]. For around 50% of patients, radiation therapy is utilized in totaling to or as a substitute of operation, and it is the focal remedies for induival with advanced and incurable phases [2]. The deadly side effects of radiation in instances of cancer is RP, which consequences beginning dose-reducing damage to neighboring tissues Dose distribution optimization is critical for providing the greatest doses to tumour tissues while preserving normal tissues from radiation overexposure [3]. Progressive 3D treatment planning technologies, together with precise assessments of tumour local controller likelihood and difficulty risk to neighboring normal tissues, have enabled advancements in tumour localization and dose distribution [4]. Individualised and patient-specific treatment planning choices are also possible with these systems. We apply technique selection and classification as machine learning approaches in this work to uncover RP-related techniques [5].

The two primary kinds of technique selection approaches established utilising unique evaluation criteria are the filter and the wrapper strategy [6]. Distance measurements, dependance and steadiness measures and data measures are some of the criteria used in these techniques [7]. The filter technique selects the appropriate technique subsets based on the qualities of the data without the need of a classification system. The wrapper approach, on the other hand, uses a predefined categorising algorithm to assess the quality of attributes. In most circumstances, the wrapper strategy outperforms the filter method, but it requires more computations. Hybrid solutions have also been created to reap the compensations of both the filter and wrapper techniques. These techniques speed up the technique selection process while improving performance [7]–[9].

The issue of categorising a sample using conditional characteristics is referred to as classification. Many well-known classification algorithms, including as decision trees, ANN, sustenance vector machines, k-nearest neighbour (k NN), and classifier of the function Bayesian, was been presented in a variability of usages [10].

2. Methodology

SVM-RFE, a consecutive regressive technique removal strategy built on SVM, was suggested for better accuracy [15]. Existing structures are organised in such a mode that the smallest significant technique is removed after repeatedly training an SVM classifier with them. A correlation-dependent technique assortment technique examines technique connection and seeks to identify the optimal technique subset by using a heuristic approach similar to the advancing best search. The technique's core idea is that exceptional characteristics are strongly connected to the class but unrelated to one another.

The chi-square approach is a straightforward strategy that use the 2 statistic to discretize techniques on a regular basis until data conflicts are detected. As a consequence of the

discretization, the technique collection is complete. In multi-class scenarios, data theory is utilised to choose techniques using a process known as data gain grounded selection. S is a collection of c1, c2. ck instances from the k classes. The entropy dispersal in S is distinct as below:

$$I(S) = -\sum_{i=1}^{k} \frac{ci}{s} log \frac{ci}{s} \quad (1)$$

The supervised learning of example set S depending on characteristic Fi is then computed as follows:

Gain(Fi)=I(S)-I(S/Fi)

$$=I(S) - \sum_{j=1}^{t} \frac{Sj}{S} I(Sj)$$

(2)

where the set of all potential values for the technique "Fi" is represented by "t." When a particular technique Fi is given, the data gain corresponds to a loss of confidence in the class's overall entropy. In other words, qualities with zero data gain suggest that it is impossible to completely eliminate this uncertainty and that it is best to avoid them.

A supervised learning technique called SVM was developed to address problems with two-class categorisation. Finding the optimum plane for which the assumed training data may be effectively detached is the main notion underlying SVM. After transforming the data x into a larger dimensional interstellar via a plotting function, it is done by maximising the margin between the two classes (x). The following decision function is defined as a result:

F(x1)=(w1,P1(X1))+B
(3)

since w is a wight assigned to the vector function and the scalar is denoted as b

Assume there are labeled training examples (xi,yi),1in, wherein xi is the training is the best sample and yi1,1 is the label for xi. The below optimization technique may be used to represent the job of choosing the optimum hyperplane.

In the hierarchical system of a decision tree, the set of data is iteratively split into divisions, and each division is finally made up completely or almost totally of trials from a single class. Leaf nodes represents the three designate classes, while non-leaf indicate specific sets of rules. Preliminary at the root node, the decision rule assesses one sample at a time. It descends the tree division till it reaches nodes of leaf. J48, a result tree classifier included in the WEKA set, was employed.

A random forest algorithm is a collection of organization trees created in the tree induction process utilising bootstrap of the working out data and a random selection. When a new input is presented, each tree votes, and the class with the most votes is picked. A naïve Bayes classifier assumes that, given a class label, all techniques are independent of one another, i.e., that the class variable is the parent of each technique. Despite its simple assumptions, the

naive Bayes classifier has consistently beaten complicated classification algorithms in a variety of situations.

3. Experimental results

We examined an NSCLC dataset including data from 209 patients treated with radiation at Washington University School of Medicine, with median doses of roughly 70 Gy. Monte Carlo techniques were used to calculate the dosage distribution (MC). The sickness group consisted of 48 persons who had been diagnosed with RP. The residual 161 patients form the another group. Age, gender, race, chemotherapy, stage, smoking, and therapies are all clinical techniques for each patient.

A range of deep learning algorithms for technique were explored for the dataset analysis. For technique selection, SVM-RFE, chi-square selection, correlation-based selection and data gain (IG) based technique selection were utilised. For classification, SVM, naive Bayes, decision trees, and random forests were all used. Throughout the SVM trials, the parameters were adjusted. The following parameters were utilised in this study: Variables in polynomial fluctuate in 1, 2, 3, 4, and 0, 1, correspondingly; for C, 1, 10, 100 are set. Radial function SVM (RBF-SVM) variables are 0.5, 1, 1.5, 2, 2.5, 3. These variables are combined to generate three linear SVMs, twenty-four P-SVMs, and eighteen RBF-SVMs (L-SVMs). Due to the unequal size of the dataset, the illness and control groups got SVM weighting values of 3 and 1, respectively. All of our trials were run using WEKA package. Afterward 30 rounds of 10-fold cross-evaluation for each organization technique, all metrics were summed to provide a fair performance estimate.

In the study of an unnecessary dataset, Matthew's correlation is a frequently used routine assessment statistic. The MCC is calculated as follows:

$$r=\frac{TP*TN-FF*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where FN and FP represent the quantity of patients wrongly off the record in the sickness and control groups, respectively, and TP and TN represent the amount of subjects correctly identified in each group. R supports real values ranging from [1.0, 1.0]. A value of +1 indicates flawless categorisation. In contrast, -1 is an exact opposite prediction. The quantity of an average chance prediction is zero.

e-ISSN 2320 –7876 www.ijfans.org
Vol.11, S Iss 1, 2022
  UGC CARE Listed (Group -I) Journal

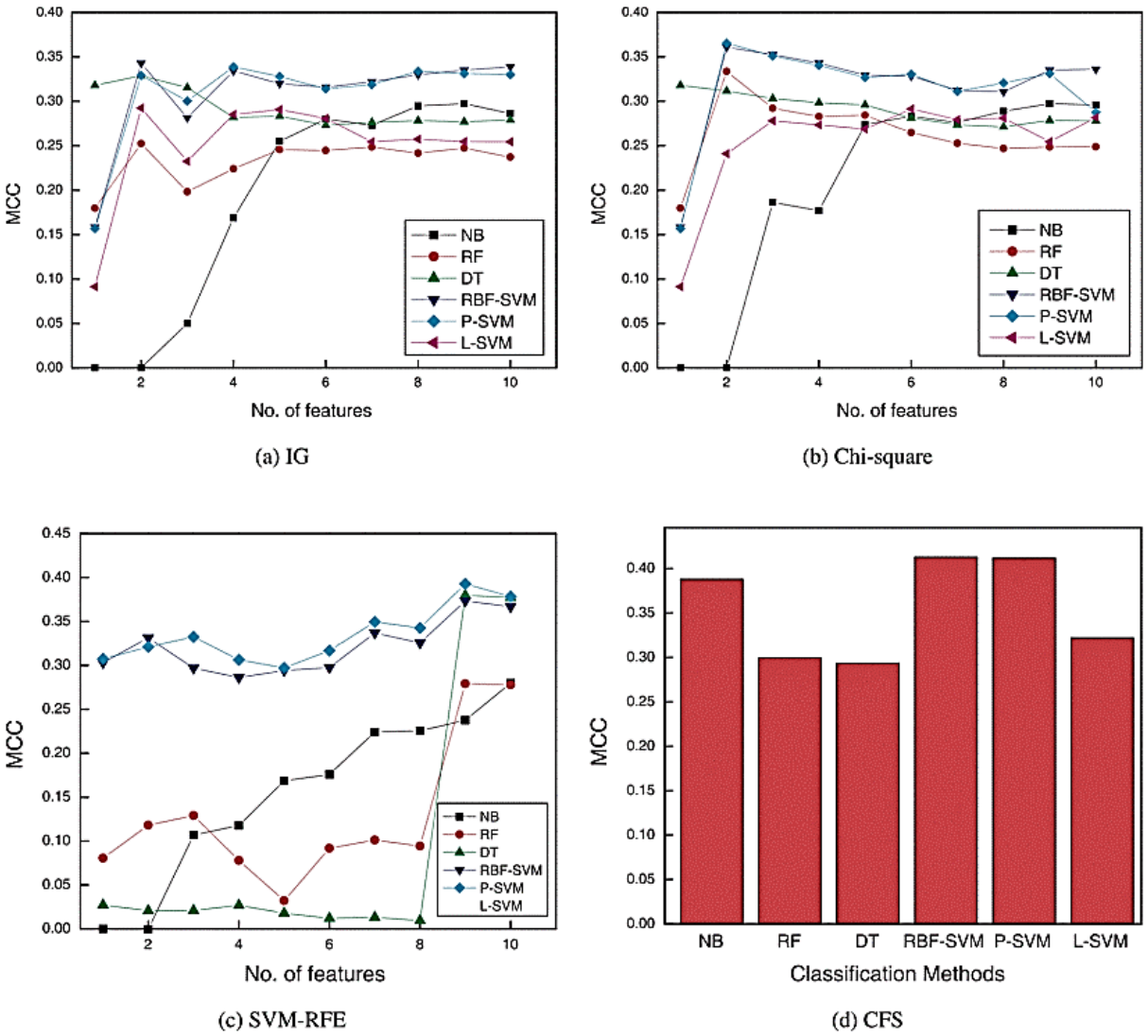


Fig. 1 MCC comparison for four technique selection methods

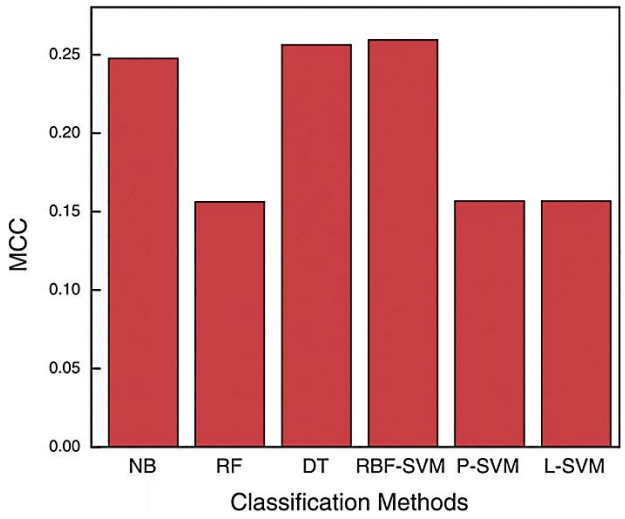


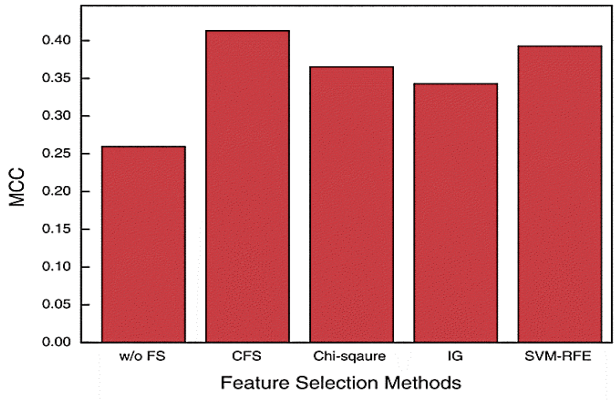
Fig. 2 MCC comparison using all available techniques.

Using the CFS criterion, only five traits were chosen. It is vital to note that these traits were often uncovered via the use of dissimilar technique selection strategies.

Figure 1 depicts the efficacy of four different technique selection techniques for classification algorithms that use the highest one, highest two, etcetera extends to the highest ten techniques. Remember that Figure 1-(d) portrays the results of CFS's search using all five variables. On this dataset, RBF-SVM and P-SVM consistently generated the best MCC. RBF-SVM and P-SVM outperformed other techniques substantially when employing techniques identified by SVM-RFE. As made known in the figure, the finest MCC was attained when RBF-SVM and P-SVM were utilised to use CFS, giving 0.42 and 0.43, correspondingly. Figure 2 demonstrations the MCC standards is lagging when all techniques were used. MCC values were usually lower when just a few critical techniques were employed, as seen in the graphic. It highlights the position of technique selection in algorithms. Figure 3 shows the greatest MCC values for each technique selection methodology for all classification algorithms.

Conclusion

In this research, the machine learning is used to identify major factors associated with RP patients' risk. After correcting for imbalance, kernel SVMs demonstrated much better MCC than not only linear SVMs but too additional challenging classification methods in our classification studies with the given techniques, as predicted. In the future, we want to create more advanced kernel algorithm and include additional quantity such as biotic indicators. We anticipate that this will shed more insight on the underlying processes of RP inception and promote the individuation of radiation in NSCLC patients.



(a) Maximum MCC for each feature selection method

Feature Selection \ Parameters	Max-MCC	Method	No. of features	C	σ	d	e
w/o FS	0.260	RBF-SVM	160	1	5		
CFS	0.413	RBF-SVM	5	100	2		
Chi-sqaure	0.365	P-SVM	2	100		3	1
IG	0.343	RBF-SVM	2	10	2		
SVM-RFE	0.393	P-SVM	9	100		3	1

(b) Parameter values used for the maximum MCC

Fig. 3 The extreme MCC for each technique selection approach is compared across all classification methods.

References

[1] Y. Zhang *et al.*, "Machine learning-based exceptional response prediction of nivolumab monotherapy with circulating microRNAs in non-small cell lung cancer," *Lung Cancer*, vol. 173, no. September, pp. 107–115, 2022, doi: 10.1016/j.lungcan.2022.09.004.

[2] Y. Yang, L. Xu, L. Sun, P. Zhang, and S. S. Farid, "Machine learning application in personalised lung cancer recurrence and survivability prediction," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 1811–1820, 2022, doi: 10.1016/j.csbj.2022.03.035.

[3] M. A. Talukder, M. M. Islam, M. A. Uddin, A. Akhter, K. F. Hasan, and M. A. Moni, "Machine learning-based lung and colon cancer detection using deep technique extraction and ensemble learning," *Expert Syst. Appl.*, vol. 205, no. May, p. 117695, 2022, doi: 10.1016/j.eswa.2022.117695.

[4] S. O. Shim, M. H. Alkinani, L. Hussain, and W. Aziz, "Technique Ranking Importance from Multimodal Radiomic Texture Techniques using Machine Learning Paradigm: A Biomarker to Predict the Lung Cancer," *Big Data Res.*, vol. 29, p. 100331, 2022, doi: 10.1016/j.bdr.2022.100331.

[5] M. Amini *et al.*, "Overall Survival Prognostic Modelling of Non-small Cell Lung Cancer Patients Using Positron Emission Tomography/Computed Tomography Harmonised Radiomics Techniques: The Quest for the Optimal Machine Learning Algorithm," *Clin. Oncol.*, vol. 34, no. 2, pp. 114–127, 2022, doi: 10.1016/j.clon.2021.11.014.

[6] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Transl. Oncol.*, vol. 14, no. 1, 2021, doi: 10.1016/j.tranon.2020.100907.

[7] S. H. Gupta, S. Goel, M. Kumar, A. Rajawat, and B. Singh, "Design of terahertz antenna to detect lung cancer and classify its stages using machine learning," *Optik (Stuttg).*, vol. 249, no. November 2021, p. 168271, 2022, doi: 10.1016/j.ijleo.2021.168271.

[8] S. Hindocha *et al.*, "A comparison of machine learning methods for predicting recurrence and death after curative-intent radiation therapy for non-small cell lung cancer: Development and validation of multivariable clinical prediction models," *eBioMedicine*, vol. 77, no. March, p. 103911, 2022, doi: 10.1016/j.ebiom.2022.103911.

[9] R. S. K. Boddu, P. Karmakar, A. Bhaumik, V. K. Nassa, Vandana, and S. Bhattacharya, "Analyzing the impact of machine learning and artificial intelligence and its effect on management of lung cancer detection in covid-19 pandemic," *Mater. Today Proc.*, vol. 56, pp. 2213–2216, 2022, doi: 10.1016/j.matpr.2021.11.549.

[10] S. H. Huang *et al.*, "How platinum-induced nephrotoxicity occurs? Machine learning

prediction in non-small cell lung cancer patients," *Comput. Methods Programs Biomed.*, vol. 221, p. 106839, 2022, doi: 10.1016/j.cmpb.2022.106839.