Impact of Various Data Splitting Ratios on the Performance of Machine Learning Models in the Classification of Lung Cancer

Archana Nazarkar[1(✉)], Harish Kuchulakanti[2], Chandra Sekhar Paidimarry[3], and Sravya Kulkarni[2]

[1] Department of Biomedical Engineering, B V Raju Institute of Technology, Narsapur, India
archana.n@bvrit.ac.in
[2] Department of Biomedical Engineering, Osmania University, Hyderabad, India
[3] Department of Electronics and Communication Engineering, Osmania University, Hyderabad, India

Abstract. Owing to revolutionary technological advancements and exceptional experimental data, particularly in the area of image analysis and processing, artificial intelligence (AI) and Machine Learning has lately become widely popular buzzword. This opportunity has been taken by medical specialties where imaging is essential, such as radiology, pathology, or cancer, and significant research and development efforts have been made to translate the promise of AI and ML into therapeutic applications. As these tools are increasingly being used for common medical imaging analytic tasks including diagnosis, segmentation, and classification. The four classifiers Artificial Neural Network (ANN), Support Vector Machine (SVM), Naïve Bayes (NB), and K Nearest Neighbour (KNN) are used in this study to classify lung cancer based on the features that are extracted from lung segmentation Algorithm. The feature data is estimated from 90 image sets and are combined for normalization and divided into training, validation, and testing sets with a ratio of 80:10:10. Different ratios (i.e., 80/20, 70/30, 60/40, 50/50) were used to divide the datasets into the training and the testing datasets to assess the model performance. ANN and KNN were very precise in achieving an accuracy of 99.8% with moderate and high training data.

Keywords: Lung cancer (LC) · Artificial Neural Network (ANN) · Data Splitting Ratios

1 Introduction

Out of all the cancers, the death rate due to lung cancer is more in number. Lung cancer leads to another type of cancer when it is in a metastatic state. The death rate due to lung cancer is estimated by the survival rate. The survival rate can increase by detecting the lung cancer in the early stages.

It is possible to detect cancer in initial stages by segmenting the pulmonary region from the Chest CT image [1–3] and extracting and analysing the small nodules formed

B. Raj et al. (Eds.): ICETE 2023, AER 223, pp. 96–104, 2023.
https://doi.org/10.2991/978-94-6463-252-1_12

in lungs [5]. By developing a CAD system, it is possible to identify the lung nodules and it is easy to determine the stage by calculating the physical properties of the cancerous nodule [6–8]. The size and growth rate of the nodule are the two most typical predictors of nodule malignancy risk. In image processing, texture, geometrical, histogram, spatial, and gradient characteristics are among the features utilized for image classification. Texture features include contrast, energy, entropy, and others. Histogram properties include average, standard deviation, and skewness. The nodule's position is determined by the spatial properties. In the study, the size of the nodule, perimeter, eccentricity, entropy, contrast, correlation, energy, homogeneity, mean, standard deviation, variance, smoothness, kurtosis, and skewness features are extracted from the segmented nodule to know the properties of the nodule to classify them accurately.

Machine Learning (ML) is the field of study that deals with algorithms that are not explicitly programmed but make the computer systems learn from experience and predict the outcome of strange data. Complex and vast data sets can be analysed and patterns can be easily recognized using ML techniques. Thus, ML techniques are used to classify lung nodules in the study. The extracted features of nodules are considered as input elements to algorithms and malignancy of the nodule is identified. The pulmonary nodule extracted from the chest CT dataset is classified into the benign or malignant class. Various classifier is used to train the extracted data in the ratio of 80:10:10 (training: testing: validation) and conclude which classifier best suits the lung cancer classification. Further, the impact of various training and testing data ratios on classification is studied and presented.

1.1 Data Used

An open-source lung image dataset LIDC [4] and CT images acquired from hospitals are used in the present study to extract their features using the proposed nodule extraction algorithm viz., Adaptive Marker Controlled Watershed method discussed [10]. The metrics of the CT Lung scan image nodule features are portrayed for reference below in Fig. 1.

2 Machine Learning Methods

2.1 Artificial Neural Networks with Feed-Forward Network

Artificial Neural Networks (ANNs) are a type of ML algorithm that learns from the data provided and classifies or predicts the responses. They work with the capacity of the human brain and are thus called "artificial neural networks" (ANNs). These are nonlinear statistical models employed in the classification and prediction of the outputs/responses in various applications like medical diagnosis, signature classification, and facial recognition. It can predict new patterns from complex input-output relationships. It can also predict the response by just learning from the sample data.

This study recommends using a feed-forward back propagation neural network configured as a classifier with an input layer, a hidden layer, and the output layer. A back-propagation learning technique for various ANN topologies has also been presented.

Image	Area1	Perimeter	Eccentricity	Entropy	Contrast	Correlation	Energy	Homogene	Mean	Standard_I	Variance	Smoothnes	Kurtosis	Skewness
CT1.jpg	1218	138.12	0.763886	0.988655	0.055224	0.887802	0.455629	0.972388	0.437378	0.496063	0.241228	0.999965	1.063744	0.252476
CT2.jpg	1235	131.347	0.702971	0.989116	0.053156	0.892074	0.457153	0.973422	0.43866	0.496223	0.206031	0.999965	1.061122	0.247229
CT3.jpg	101	56.179	0.972898	0.974905	0.055362	0.88543	0.464492	0.972319	0.407013	0.491277	0.21024	0.999963	1.143302	0.378552
CT4.jpg	213	103.601	0.852697	0.998718	0.062623	0.87457	0.442036	0.968689	0.478928	0.499556	0.199182	0.999968	1.007117	0.084365
CT5.tif	36	25.745	0.377635	0.867962	0.055882	0.864449	0.53498	0.972059	0.289413	0.45349	0.18021	0.999947	1.862553	0.928737
CT6.jpg	31	32.138	0.888104	0.970679	0.083915	0.825028	0.443533	0.958042	0.399536	0.489803	0.200961	0.999962	1.168282	0.410221
CT7.jpg	53	35.454	0.962338	0.778881	0.043873	0.876663	0.602341	0.978064	0.230499	0.421152	0.16625	0.999934	2.637953	1.279825
CT8.jpg	47	34.536	0.978184	0.835049	0.048928	0.874887	0.562398	0.975536	0.265579	0.441641	0.180117	0.999943	2.126972	1.061589
CT9.jpg	36	29.384	0.968896	0.656547	0.0572	0.797459	0.663662	0.9714	0.169495	0.375188	0.1287	0.99991	4.103978	1.761811
CT10.jpg	511	83.091	0.571445	0.934414	0.051961	0.885746	0.496048	0.97402	0.350388	0.477091	0.217106	0.999956	1.393362	0.627186
CT11.jpg	123	90.038	0.964607	0.679159	0.038174	0.870844	0.667719	0.980913	0.179581	0.383838	0.128134	0.999915	3.787416	1.669556
CT12.jpg	67	46.9	0.756774	0.860493	0.048039	0.882089	0.546848	0.97598	0.283737	0.450811	0.17957	0.999946	1.920524	0.959439
CT13.jpg	104	82.909	0.961394	0.918214	0.078891	0.82282	0.482074	0.960555	0.333252	0.471376	0.210272	0.999954	1.50055	0.707495
CT14.jpg	31	32.312	0.978307	0.530267	0.042494	0.799931	0.746916	0.978753	0.120316	0.32533	0.092763	0.999873	6.448249	2.334148
CT15.jpg	34	30.707	0.973144	0.814957	0.053768	0.857869	0.570821	0.973116	0.252335	0.434352	0.161363	0.99994	2.300489	1.14039
CT16.jpg	301	81.975	0.64567	0.804683	0.039384	0.894032	0.590505	0.980308	0.24588	0.430608	0.182445	0.999938	2.393072	1.180285
CT17.jpg	41	29.119	0.973095	0.76206	0.045374	0.868606	0.611358	0.977313	0.221039	0.414947	0.146658	0.999931	2.807853	1.344564
CT18.jpg	70	33.164	0.698047	0.850729	0.050306	0.874586	0.551103	0.974847	0.276566	0.4473	0.179854	0.999945	1.998075	0.999037
CT19.jpg	90	57.511	0.882325	0.875206	0.068658	0.83534	0.519086	0.965671	0.29509	0.456083	0.188891	0.999948	1.80742	0.898566
CT20.jpg	76	58.082	0.924778	0.76825	0.049724	0.85758	0.603611	0.975138	0.224472	0.417234	0.151045	0.999932	2.744342	1.320735
CT21.jpg	286	78.7	0.694	0.771917	0.039553	0.887384	0.610795	0.980224	0.226532	0.418587	0.171479	0.999933	2.707266	1.306624
CT22.jpg	239	93.751	0.901837	0.758383	0.048775	0.857829	0.610535	0.975613	0.219025	0.413585	0.146675	0.99993	2.846146	1.35873
CT23.jpg	34	30.707	0.973144	0.81677	0.053922	0.857895	0.569538	0.973039	0.253494	0.435011	0.161843	0.99994	2.284437	1.13333
CT24.jpg	30	20.138	0.867019	0.89426	0.045129	0.894911	0.527473	0.977436	0.310944	0.46288	0.191423	0.999951	1.667277	0.81687
CT25.jpg	42	27.68	0.958742	0.77827	0.044822	0.873858	0.601854	0.977589	0.230148	0.420928	0.166045	0.999934	2.643976	1.282176
CT26.jpg	36	29.384	0.968896	0.578252	0.0519	0.782207	0.712497	0.97405	0.137726	0.344612	0.109006	0.999889	5.420526	2.102505
CT27.tif	36	25.745	0.377635	0.867962	0.055882	0.864449	0.53498	0.972059	0.289413	0.45349	0.18021	0.999947	1.862553	0.928737

Fig. 1: Metrics of the CT Lung scan images nodule features

Rather than examining a large number of input elements and hidden neurons, we have chosen to simply feed forward the propagation topology with eight significant features as input and 7 hidden layers. This is to ensure the network's capacity to simplify, which may be accomplished by reducing the node's number as much as feasible. If a high number of nodes are used, the network performs perfectly with the trained set. But it utterly fails with the un-trained dataset. Extreme care should be exercised while choosing these hidden layers to avoid over-fitting and under-fitting. We have selected 7 hidden layers in the first instance as the study says that it should be 2/3 of the sum of the input layers and the number of output layers. Furthermore, the number of hidden layers has been increased and checked for performance. The Table 1 shows the parameters used in the design of the ANN classifier.

The results of the study on pulmonary nodule classification with the neural network have shown optimal values and are summarized in the Table 2. The model consists of eight feature elements as inputs, seven hidden neurons, and one output neuron. The ANN model with feed forward and back propagation network has attained the results with minimal error. The proposed classification model has achieved a training accuracy

Table 1: Design parameters of ANN Model

S. No.	Parameters	Method
1	Number of features selected	8
2	Number of Hidden neurons	7
3	Training Function	Trainlm
4	Output layer Activation Function	Purelin
5	Performance of network	MSE

Table 2: ANN classifier results

NN Classifier	No. of Epochs	%MSE	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Sensitivity (%)
8-7-1 (7 hidden neurons)	132	2.431e-1	99.982	99.864	100	100

of 99.98%. Moreover, it also takes less simulation time and has good generalization even with a high number of samples.

Furthermore, the study was carried out to evaluate and compare the performance of different machine learning (ML) algorithms. The performance of the discussed Artificial Neural Network (ANN) with Feed-forward back propagation network architecture and other topologies like ANN-Trainable cascade-forward back propagation network, ANN-Elman back propagation network, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naive Bayes (NB) algorithms is evaluated. Further, the influence of various training-to-testing ratios in predicting cancer and its stage is assessed. Several ratios of training and testing datasets, like 80/20, 70/30, 60/40, and 50/50, were used for the performance assessment of models. The prediction accuracy of the models was evaluated using popular statistical indicators such as Mean Squared Error (MSE) under various training and testing ratios.

2.2 Support Vector Machine (SVM)

Another popularly used supervised learning technique for classification is SVM. It is a discriminative classifier where a hyper plane is created to separate the dataset into classes. A hyper plane is a line that separates different data points in an n-dimensional plane. The dataset can be separated by many lines, but the ideal hyper plane is selected that best maximizes the margin. SVM provides high accuracy and less time for processing as compared to other algorithms. Figure below shows an example of SVM and the separation of data points by the optimal hyper plane. The results of our experiment on pulmonary nodule classification into benign and malignant with the SVM classifier have shown the best performance for an 80% training and 20% testing dataset. The results are shown in Fig. 2. The accuracy of the model is 100%. But the specificity of the model is 92%.

2.3 K-Nearest Neighbors (KNN)

KNN is a lazy learning approach that predicts the new instances by computing the Euclidean distance and assigning the nearest neighbors a higher weight and a lower one to the distant ones in the entire training dataset. KNN is used for data mining as well as machine learning. It searches for the K's nearest instance. In simple words, K is a user-defined constant, and it classifies the new sample with the most recurrent label nearest

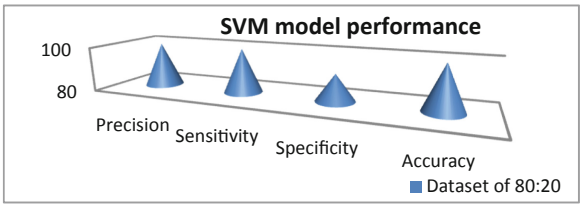

Fig. 2: Performance metrics of SVM model

to it in the training dataset. It is simple and easy to implement. It is used in data mining and classification applications. The performance of the KNN topology is at its best with the present dataset. The accuracy is 100% and the simulation time is very small. The precision, sensitivity of the model is 100%.

2.4 Naive Bayes (NB)

The Naive Bayes classifier is a probabilistic ML approach and is based on Bayes' theorem. It is very easy to implement and can accurately predict the results with a small training dataset.

Using Bayes' Theorem (Eq. 1), the likelihood of A happening is estimated, given that B has occurred. The prior probability of the predictor is P(B), and the posterior probability is P(A|B).

$$P(A|B) = \frac{\mathrm{P(B|A)P(A)}}{P(B)} \tag{1}$$

The performance of the NB model is 100% accurate, and the precision, sensitivity and specificity of the model are also 100% for the 80–20 training and testing dataset.

2.5 Performance Evaluation of Machine Learning Techniques

The extracted significant features from LIDC and a real-time dataset were used in the classification. These are fed as the input elements to the classifiers. Six popular ML techniques ANN-FF, ANN-CF, ANN-EL, NB, KNN, and SVM were applied with training and testing ratio as 80-20 to classify the lung cancer. The performance metrics like accuracy, precision, sensitivity, specificity is obtained for all the classifiers to find the best classifier in classifying lung cancer. Figure 3 shows the plot of accuracy, precision, sensitivity, specificity for the classifier analyzed in the study.

Performance of the ANN-CF, ANN-EL, and SVM models is unsatisfactory with the accuracy of 94%, 90% and 94% respectively. The accuracy is at its best for ANN-FF, KNN, and NB with 99.8%, 99.8% and 99.7% respectively. The sensitivity of ANN-FF, KNN, NB and SVM outperformed other models. The specificity and precision are low for ANN-CL, ANN-EL and SVM models.

Validation and comparison results state that performance of every model is good; however, ANN-FF, KNN, and NB were the best approaches, taking the metrics into

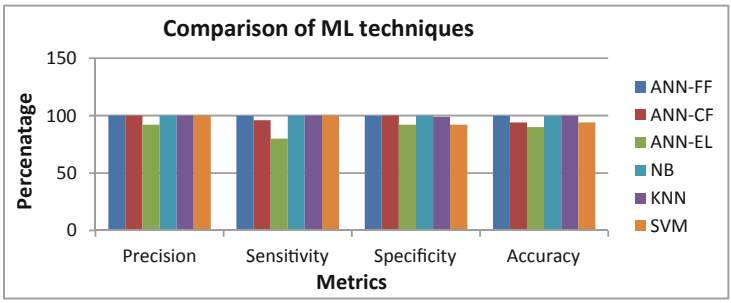

Fig. 3: Comparison of the metrics of various ML models

consideration and compared within the studied models. ANN-FF was an accurate and statistically stable model, taking into account the Mean Square Error (MSE = 0.3205).

Further, it is learnt from the literature [9] that the data splitting ratio will also affect the performance of the model. Thus, a study was carried out to understand the impact of the data training and testing splitting ratio. For training and testing purposes, a ratio of 80:20 was considered for the earlier implementations. Now to understand the effect, 70/30, 60/40, and 50/50 ratios of training and testing datasets were executed for all the models studied.

3 Impact of Data Split Ratios on the Performance of ML Methods

The change in the dataset ratios impacted the performance of the model. The performance of the training is enhanced by increasing the volume of train data, and it even enhanced the model stability. Likewise, the performance of the testing is also improved with an increase in the training dataset. But when the latter has increased from 80% to 90%, the drift is in another direction. Thus, the splitting ratio had a significant impact on the capability of the ML models to predict.

The primary goal of this study is to assess the efficacy of machine learning models for cancer stage prediction using various data-splitting ratios. In this study, four machine learning approaches; ANN, KNN, NB, and SVM were considered to assess the malignancy of a pulmonary nodule based on different training and testing splitting ratios of input data. It is the first time to study the impact of various data splits of training and testing data used in lung cancer prediction models, which is the primary distinction between this study and the prior published works. Mean Squared Error (MSE) and statistical approaches are used to evaluate the results to select the best model for prediction of lung cancer. The dataset was split into 70/30, 60/40, and 50/50 training and testing ratios, in addition to the previously 80/20 dataset, and the metrics were analysed.

The ANN-FF model has an accuracy of 100% with the 80/20 dataset, but as the data split altered, the accuracy decayed. 96% for 70/30 data split, 94% for 60/40 data split, and finally 98% for 50/50 data split. The accuracies of the ANN-CF and ANN-El are poor for all the ratios. Interestingly, the NB and KNN models exhibited an accuracy of 100% for all the data splits executed. SVM has a decay of 70/30 and 60/40 split but achieved an accuracy of 100% for 50/50 training and testing datasets. Figure 4 shows

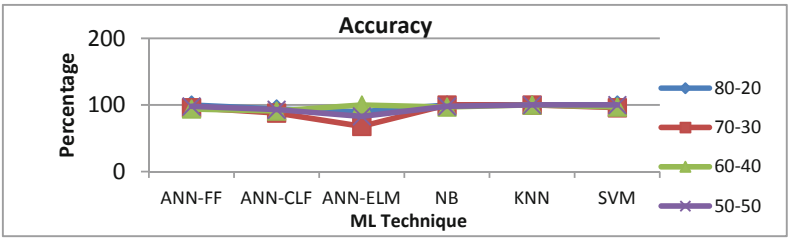

Fig. 4: Accuracy of various ML models with different data splits.

the trends of accuracy for different ML techniques employed in the study with different training and testing datasets. Likewise, the precision is analysed and is presented in Fig. 5. NB and KNN executed the best precision percentage with any data split ratio.

Figure 6 shows that ANN-FF, SVM, and KNN outperformed the sensitivity metric when compared to the performance of other models. The sensitivity of the NB slightly decreased with the 60/40 and 50/50 data splits.

The specificities of the models ANN and NB were 100% with all the splits considered for the study, whereas the other models' specificities decreased with the decay in the training data. Figure 7 shows the trends in the specificity of ML models with various data splits.

The performance of the ML techniques is varying with the training and testing dataset ratios as observed from the above discussion. Taking into consideration the performance of all the metrics, it can be concluded that the ANN-FF and KNN models are robust and

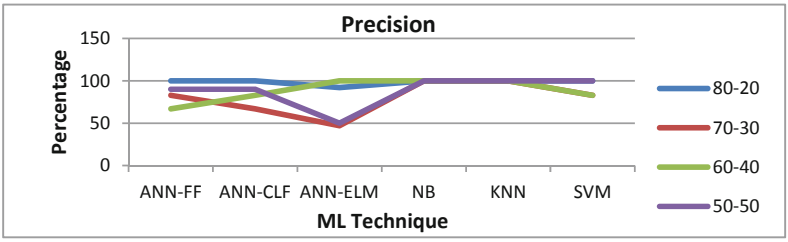

Fig. 5: Precision of various ML models with different data splits.

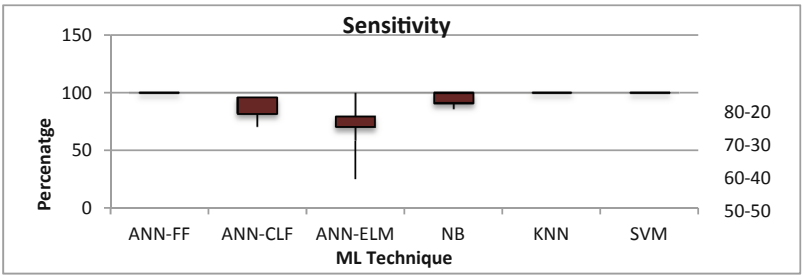

Fig. 6: Sensitivity Trends with Different Dataset Ratios

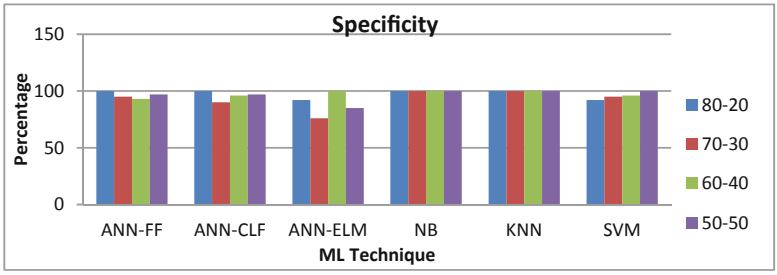

Fig. 7: Trend of Specificity with Different Dataset Ratios

very efficient with moderate training and testing dataset ratio also. The performance of these models is superior irrespective of the dataset ratio.

4 Conclusions

Various texture, geometrical, histogram, spatial, and gradient features were extracted from the proposed CAD system and eight significant features were used to classify lung cancer. Most of the researchers used texture, fractal features in classification using ANN, SVM as the classifier. The diagnostic accuracies of those CAD systems were around 98%. In the current study, with these significant features the CAD system aids a superior detection of cancerous region. ANN and KNN were robust and very precise in achieving the strong accuracy of 99.8% with moderate and high testing data. It can be observed from the above that if the training dataset is of 50% or 80%, the ANN-FF and KNN achieved an enriched accuracy of 99.8%.

References

1. Sluimer I, Prokop M, van Ginneken B: Toward automated segmentation of the pathological lung in CT: IEEE Trans Med Imaging 2005; 24(8):1025–1038.
2. M. L. Giger, N. Ahn, K. Doi, H. MacMahon, and C. E. Metz: Computerized Detection of Pulmonary Nodules in Digital Chest Images: Use of Morphological Filters in Reducing False-Positive Detections:" Med. Phys, vol. 17, pp. 861–865, 1990.
3. K. Suzuki: A supervised 'lesion-enhancement' filter by use of a massive-training artificial neural network (MTANN) in computer-aided diagnosis (CAD): Phys. Med. Biol, vol. 54, pp. S31–S45, 2009.
4. Armato, I., Samuel McLennan, G., McNitt-Gray, F. R., Michael, Charles, Reeves, Anthony P, -- Clarke, Laurenc, (2015): Data from LIDC-IDRI. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX. Lung Image Database Consortium - Image Database Resource Initiative (LIDC-IDRA) database. https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI.
5. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017 by MacMahon et al. Radiology (2017) https://doi.org/10.1148/radiol.2017161659.

6. Manikandan T, Devi B, Helanvidhya T: A Computer-Aided Diagnosis System for Lung Cancer Detection with Automatic Region Growing, Multistage Feature Selection and Neural Network Classifier: International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1S, November 2019.
7. M. C. Lee, L. Boroczky, K. Sungur-Stasik et al.: Computer aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction: Artificial Intelligence in Medicine, vol. 50, no. 1, pp. 43–53, 2010.
8. Aggarwal, T., Furqan, A., & Kalra, K. (2015): Feature extraction and LDA based classification of lung nodules in chest CT scan images: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), DOI:https://doi.org/10.1109/ICACCI.2015. 7275773.
9. Quang Hung Nguyen, Hai-Bang Ly, Lanh Si Ho, Nadhir Al-Ansari, Hiep Van Le, Van Quan Tran, Indra Prakash, Binh Thai Pham: Influence of Data Splitting on Performance of Machine Learning Models in Prediction of Shear Strength of Soil: Hindawi, Mathematical Problems in Engineering, Volume 2021.
10. Kuchulakanti H., Paidimarry C. (2020) Early-Stage Squamous Cell Lung Cancer Detection. In: Satapathy S.C., Raju K.S., Shyamala K., Krishna D.R., Favorskaya M.N. (eds) Advances in Decision Sciences, Image Processing, Security and Computer Vision. Learning and Analytics in Intelligent Systems, vol 3. Springer, Cham. https://doi.org/10.1007/978-3-030-24322-7_15.