A hybrid approach of data visualization technique and random forest classifier for binary classification of lung CT images

Ananya Bhattacharjee, P. Stoila Cindy, R. Murugan and Tripti Goel

Biomedical Imaging Laboratory, Department of Electronics and Communication, National Institute of Technology Silchar, Silchar, Assam, 788010, India
murugan.rmn@ece.nits.ac.in

Abstract. Lung cancer is the most common and dangerous cancer worldwide. An automatic detection system is the need of the hour for early diagnosis. Machine learning classifiers often encounter outliers in the extracted features. Motivated by this, the main aim of this study is to develop an outlier free automated computer-aided system for the prediction of pulmonary nodules using various data visualization and machine learning (ML) classification techniques that can help in the decision-making process of radiologists. Data visualization techniques are used for removing outlier values from the extracted features. A comparative analysis using several ML techniques such as Decision Tree, Support Vector Machine, K Nearest Neighbours, and Random Forest classifier has been performed. Random forest is the best-performing classifier, which obtained 92.92% cross-validation accuracy, 96% precision, 90.74% sensitivity, 95.56% specificity, and 93.29% F1 score. Hence, the proposed model can open up new opportunities for radiologists for early lung cancer detection.

Keywords: lung cancer, machine learning, random forest, binary classification, computed tomography.

1 Introduction

Lung cancer is among the top deadly cancers whose survival rate is less than 5 years [1] [2]. The survival rate can be improved if this disease is detected at an early stage and early diagnosis is carried out [3] [4]. Cancerous nodules are the result of abnormal rapid cell growth and their shape varies from irregular, round to polygonal lung nodules that can be either singular or multiple [5]. Lung cancer accounts for roughly 1.8 million new cases each year [6] [7]. It is found that around 70% of patients are diagnosed at an advanced stage because it is hard to detect symptoms of lung cancer at the early stage [8]. Its early diagnosis mainly depends upon lung nodule detection. Computed tomography (CT) scanned images are more often widely preferred by radiologists rather than magnetic resonance imaging (MRI). But manual screening of CT scanned images is a time-consuming task. Real-time detection of malignant nodules by radiologists through the naked eye is a time-consuming act and differences in the opinion of doctors may occur depending on the shape, size, and

R. Murugan et al. (eds.), *Proceedings of the International Conference on Signal Processing and Computer Vision (SIPCOV 2023)*, Advances in Engineering Research 239,
https://doi.org/10.2991/978-94-6463-529-4_21

texture of the nodules [3]. Furthermore, manual detection of small nodules may get missed [9]. To address this issue, computer-aided diagnosis (CAD) systems are widely used to improve the decision-making process of radiologists for early diagnosis [10].

The enthusiasm toward biomedical image processing, Machine Learning (ML), data science, and Artificial Intelligence is the key factor behind this study. Medical imaging is the field that mainly deals with the application of AI and ML and it is making a remarkable change for the betterment of society. Many people are losing their lives due to lung cancer disease \cite{kukreja2023heuristic}. Data science techniques are apt for healthcare applications because of the amount of availability of data. Thus, all the above factors pushed us hard to propose a data visualization-based random forest classifier which can effectively distinguish between malignant and non-malignant cases and hence, is the novelty of the study.

The contributions of the paper are as follows:

1. A unique data science combined approach consisting of data visualization techniques and a random forest classifier has been proposed that can effectively discriminate between malignant and non-malignant classes.
2. Outlier removal of all the features extracted ensured better performance of the proposed model.
3. The proposed model showed the highest performance compared to the existing state-of-the-art techniques.

The remaining sections are arranged as follows. Section II discusses the literature survey. Section III presents the materials and methodology. Section IV depicts the results and Section V depicts the discussion. Lastly, Section VI presents the conclusion.

2 Related Works

Over the years, many researchers have proposed many ML-based techniques for lung cancer detection. The authors in [12], extracted the cancerous nodules by Delta Radiomics. To predict malignant nodules, Support Vector Machine (SVM) was used, and it achieved 90.9% accuracy. In [13], the authors developed an efficient lung nodule detection scheme that performed classification using a SVM classifier and achieved an accuracy of 80.36%. A comparison of four classical ML methods with deep learning methods was performed to classify lymph nodes of lung cancer in [14]. The best classical method achieved an accuracy of 80.00%. The authors in [15] used a technique based on an AI application using the ensemble ML algorithm to predict the different types of lung cancer and obtained an accuracy of 90.74%. A dynamic time series-based algorithm was proposed by [16] using image processing techniques to focus on pulmonary nodules boundary. Recurrence plot was also visualized in order to view the similarity degree of the boundaries. The average processing speed of this model was observed to be 0.58 sec. A computer-aided decision system was proposed by authors in [17] by using a Deep Fully Convolutional Neural Network (DFCNet) and Convolutional Neural Network (CNN) for detecting lung cancer.

In summary, the methods discussed above are purely based on ML techniques. However, the combination of both data visualization techniques and the Random Forest classifier has not been incorporated yet. To the best of our knowledge, this hybrid approach will be the first work that can effectively distinguish between malignant and non-malignant classes.

3 Materials and Methods

This section describes the database of the study and the comparison of different U-Net models.

3.1 Materials

This section describes the database of the study and the comparison of different U-Net models. The database used in this study is ”Lung Image Database Consortium(LIDC)-Image Database Resource Initiative (IDRI) public dataset [18]. This dataset is one of the biggest online databases available with an approx of 125GB, consisting of 1018 helical CT scans. The CT-scanned images are in DICOM format which stands for Digital Imaging and Communications in Medicine. Mostly, CT images of each patient vary from 240 to 550 at various angles. This database can be found on the cancer imaging archive website. It can be downloaded from the National Biomedical Imaging Archive (NBIA). The nodules have been annotated by four radiologists and are categorized into three categories- (1) nodules that are greater or equal to 3 mm are considered nodules and they range from 3-30mm, (2) nodules that are less than 3mm are considered as non-benign nodules, and (3) those non-nodules greater or equal than 3mm does not have any features of being a nodule. 7371 is the total number of nodules present in the database. The thickness of the slice varies for different companies. For instance, Siemens follows 0.75mm of slice thickness whereas, GE Medical System follows 1.25mm. Fig. 1 shows a malignant CT image of the given dataset.

3.2 Methodology

Fig. 2 shows the block diagram of the proposed model. First, the CT images are fed as input and then the lung nodule is segmented from the other parts of the lungs through adaptive thresholding ranging from 1000-1300 HU. The nodules are detected through label-connected components and then nine features such as area, convex area, solidity, major axis, minor axis, circumference, circularity, perimeter and radii are extracted from the detected nodules. Detailed analysis of the extracted features are performed through data visualization techniques such as histogram, boxplot and normal distribution curves. The extracted features are observed to have some outlier values. These outlier values are first visualized through boxplot and normal distribution curves. Then the outliers are removed through python's clipping function. The training and testing data are split in the ratio of 80:20. The outliers-free features are then fed to Machine Learning (ML) classifiers such as Support Vector Machine (SVM), K Nearest Neighbor (KNN), Decision Tree (DT) and Random Forest (RF) classifiers. The proposed model is built with the best classifier which can effectively distinguish between the malignant and non-malignant classes.

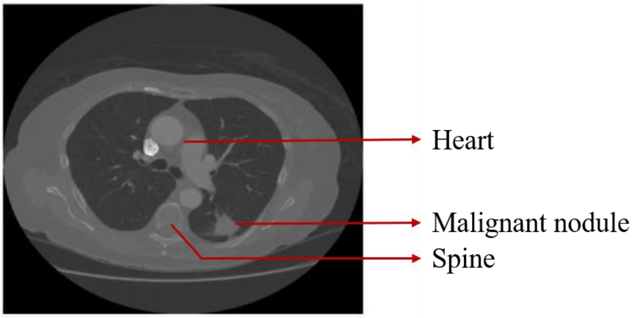


Fig. 1. A sample CT scan image from LIDC-IDRI database [18]

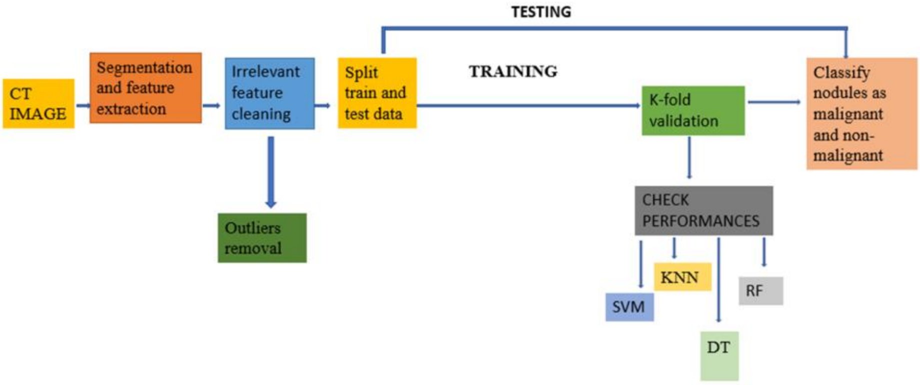

TESTING
TRAINING
CT IMAGE
Segmentation and feature extraction
Irrelevant feature cleaning
Outliers removal
Split train and test data
K-fold validation
Classify nodules as malignant and non-malignant
CHECK PERFORMANCES
SVM
KNN
DT
RF

Fig. 2. Block diagram representation of the proposed methodology

4 Results

This section presents the environment used, the performance metrics considered for evaluating the model, the experimental setup and the experimental results.

4.1 Implementation details

Python code was used to execute the proposed model on an Acer system with an Intel Core i7-8700 processor at 3.20 GHz and 16 GB of RAM. Python 3.8.5 and the scikit-learn Python-based ML library were utilized.

4.2 Experimental Setup

The experimental setup of this study mainly consists of data visualization techniques and ML classifiers. The extracted features are visualized through data visualization techniques such as histogram analysis, box plot and normal distribution curves. The nodules ≥ 3 mm are considered in this paper because of their high chances of being malignant. A total of 495 features are extracted from the detected nodules. An experiment consisting of a comparison of four ML classifiers, namely, DT, SVM, KNN and RF have been performed and then, the proposed model is built with the best-performing ML classifier out of these four. The best model then classifies the malignant and non-malignant cases.

4.3 Experimental results

The pulmonary nodules from the other parts of the lungs are first segmented and then the nodules detected are separated from the segmented output so that feature can be extracted from it. Adaptive thresholding ranging from 1000 to 1300 HU has been used to segment the nodules. Then, the nodules are detected through eight label-connected components from the segmented binary images. Fig. 3 shows the input image, the segmented output and the detected nodule result.

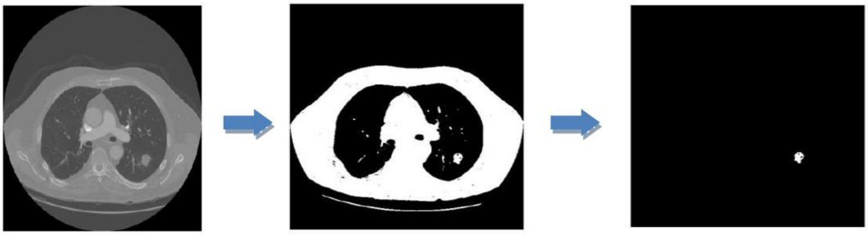


Fig. 3. Input image, segmented output and the detected nodule

Various features such as area, solidity, convex area, minor axis, major axis, circularity, radius, perimeter and circumference are extracted from the detected nodules. The area is determined by assigning ’1’ to all the white pixels of the suspected nodule. The perimeter calculates the nodule’s boundary length. In a convex image, the number of pixels is determined by the convex area. The ratio between the area and the convex area is determined by solidity. The lengths of the ellipse’s main and minor axes are indicated by the major and minor axes, respectively. Circularity provides the degree of circularity of a nodule. The closer its value is to 1, the more circular the shape of the nodule. The radius of the suspicious nodule is determined by radii. The circumference is the outermost extent of the circular path of the nodule. If the nodule is malignant, ’1’ is allocated; otherwise, ’0’ is assigned. The features extracted are converted into Comma Separated Values (CSV) file, which is then fed to the RF classifier after outlier removal through data visualization techniques such as histogram analysis, box plot and normal distribution curves. Fig. 4 shows the histogram analysis of all the extracted features to analyze all the features in depth. It is found that the histogram of some features such as area, major axis, convex area and perimeter shows some discontinuity along the x-axis, thus showing zero values present in between the histogram bars. Thus, the empty values of these features are replaced with their respective median values to fill the empty spaces.

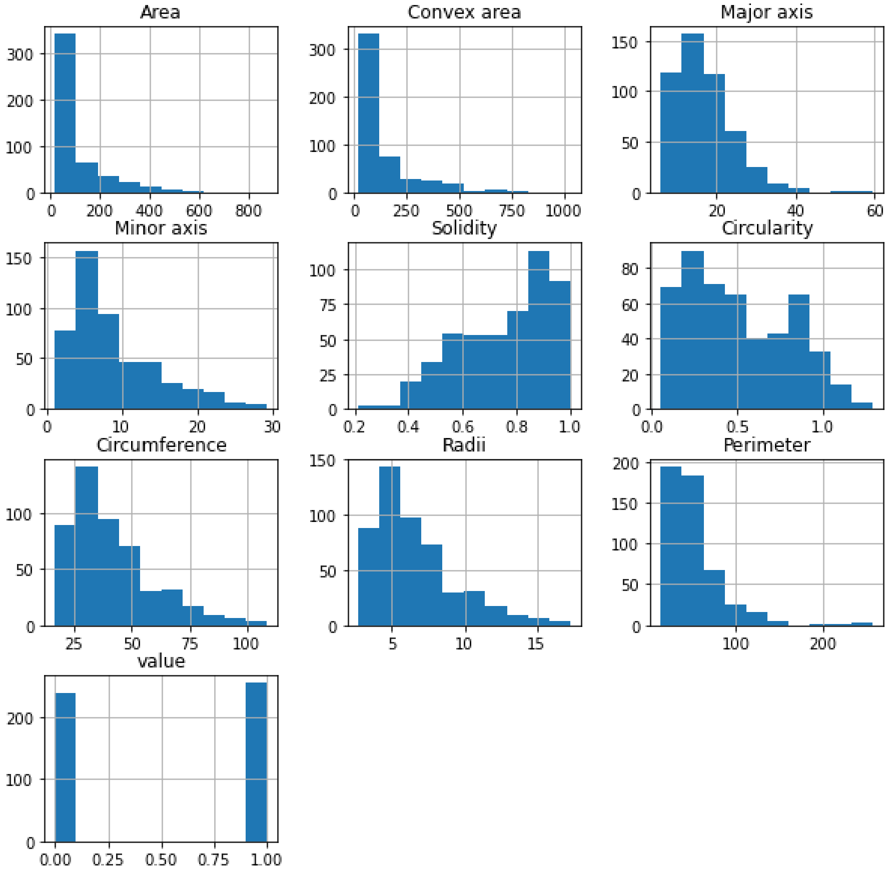


Fig. 4. Histogram analysis of all extracted features

Then, a box plot is plotted to visualize the range of the outliers, as shown in Fig. 5. It is found that Area (Ar) and Convex Area (CAr) have huge outlier values while

Perimeter (P) has medium outlier values and features like Circumference (Ci) and Major Axis (MA) have fewer outlier values. Circularity (C) shows no outlier values, while Minor Axis (MiA), Solidity (S) and Radii (R) show the least outlier values. Values (V) have values either 0 or 1. 0 stand for normal and 1 stand for malignant.

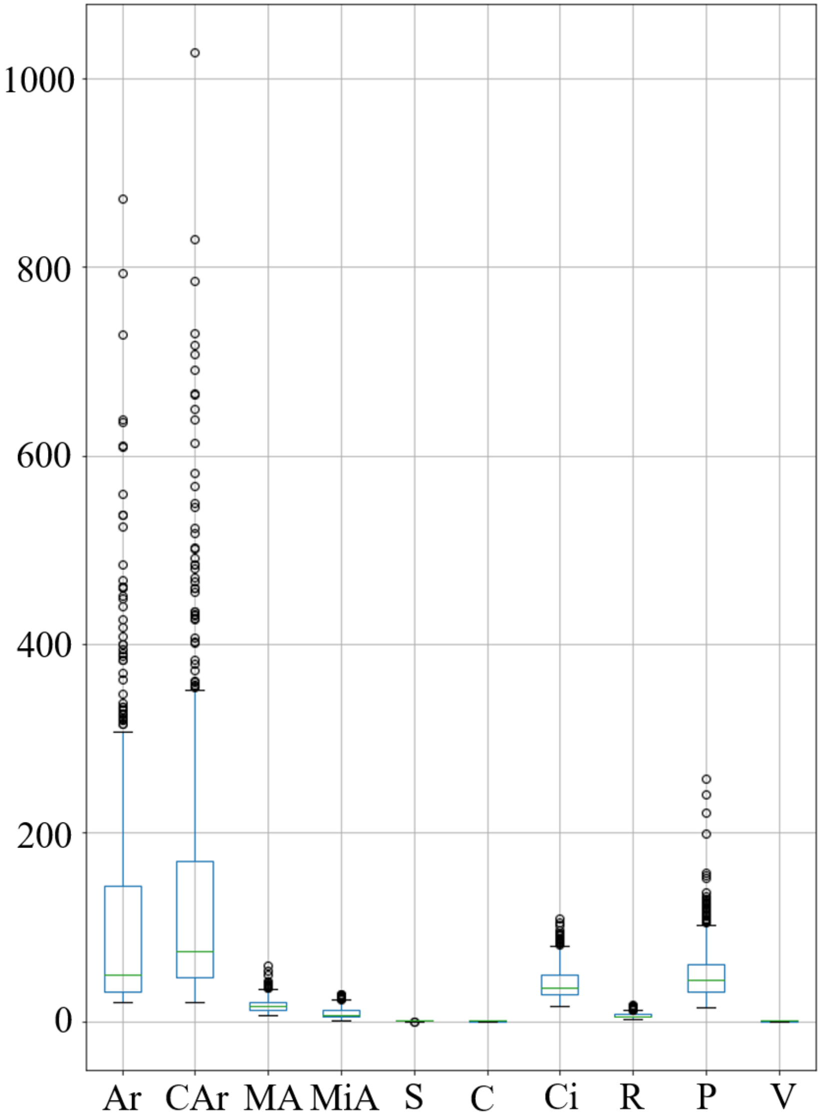


Fig. 5. Boxplot of all extracted features to detect outliers

Fig. 6 shows the normal distribution curves of the Area, Convex area, Circumference, Major axis and Perimeter (ACoCiMP) before removing the outliers. It is observed that ACoCiMP exhibits outlier values far away from the mean of the normal distribution curve that needs to be removed before feeding the features to the RF classifier. Thus, the excess outliers values are clipped using python's numpy clip function to the percentile range of 10 to 90 so that standard deviation values of these features are improved, as shown in Fig. 7. Table 1 compares the mean and Standard Deviation (Std) of ACoCiMP before and after removing the outliers. It is observed that the post-outlier removal ensures better Std values compared to pre-outlier removal.

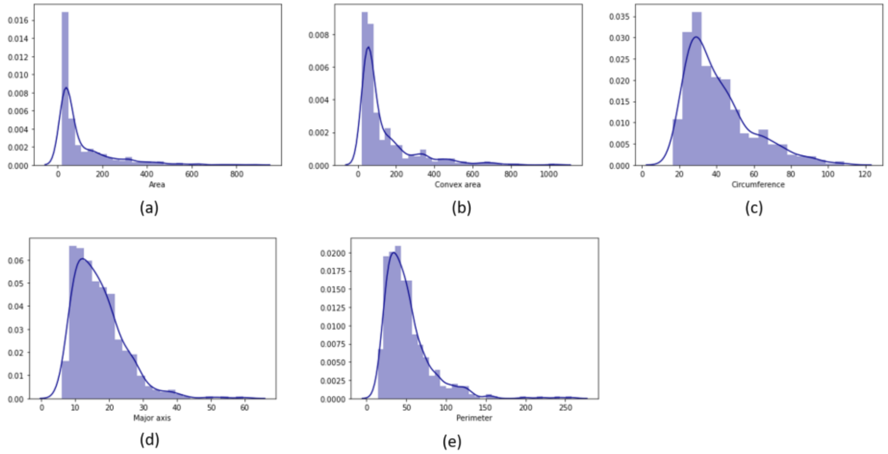


Fig. 6. Normal distribution curve before removing outliers (a) Area (b) Convex area (c) Circumference (d) Major axis (e) Perimeter

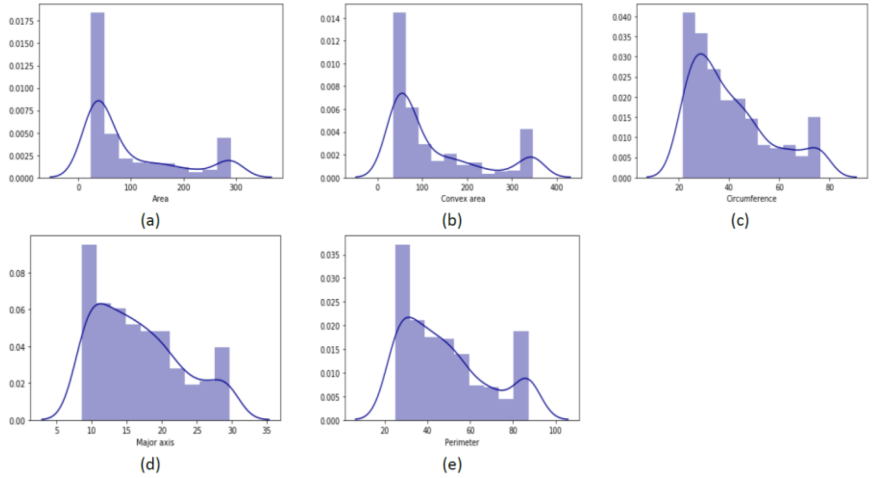


Fig. 7. Normal distribution curve after removing outliers (a) Area (b) Convex area (c) Circumference (d) Major axis (e) Perimeter

After the feature outlier cleaning, all the nine features are fed to the different Machine Learning (ML) classifiers such as Decision Tree (DT) Support Vector Machine

(SVM), K Nearest Neighbor (KNN) and the proposed Random Forest (RF) model. Table 2 shows the experimental results of the proposed model with the ML classifiers with respect to 10 Fold Cross Validation Accuracy (CAcc), Precision (Prec), Sensitivity (Sen), Specificity (Spec) and F1 score (F1). The DT classifier performed the poorest among all ML classifiers by obtaining 88.85% CAcc, 92.16% Prec, 87.04% Sen, 91.11% Spec and 89.53% F1. SVM and KNN performed better than DT. Although KNN and RF have same Prec, Sen, Spec and F1, the proposed RF model has higher CAcc than KNN. The proposed RF model obtained the highest CAcc of 92.66%, Prec of 96%, Sen of 90.74%, Spec of 95.56% and F1 of 93.29% and hence, is the best model among all the ML classifiers.

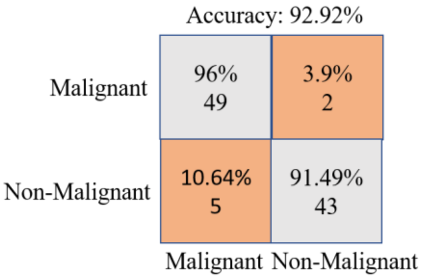
Table 1. Mean and Std of ACoCiMP features

Type	Before removing outliers	After removing outliers
Area	111.13 ± 130.82	97.11 ± 89.65
Convex area	139.30 ± 150.35	123 ± 103.29
Circumference	40.99 ± 17.50	40.52 ± 15.86
Major axis	17.01 ± 7.40	16.68 ± 6.186
Perimeter	51.62 ± 30.47	48.61 ± 20.11

Table 2. Performance analysis of the proposed model and other ML classifiers

Models	CAcc (%)	Prec (%)	Sen (%)	Spec (%)	F1 (%)
DT	88.85	92.16	87.04	91.11	89.53
SVM	89.61	94.12	90.56	93.48	92.30
KNN	91.38	96	90.74	95.56	93.29
RF	**92.66**	**96**	**90.74**	**95.56**	**93.29**

Fig. 8 shows the confusion matrix of the proposed RF model. The number of TP and TN cases are 49 and 43, respectively. Whereas the misdiagnosis cases (FP and FN) are 2 and 5, respectively. The proposed model detected 96% malignant cases and 91.49% non-malignant cases. Fig. 9 shows the ROC curve of the proposed model. The proposed model obtained 99.71% Area under Curve (AUC) value.



Accuracy: 92.92%

	Malignant	Non-Malignant
Malignant	96% 49	3.9% 2
Non-Malignant	10.64% 5	91.49% 43

Fig. 8. Confusion matrix of the proposed model

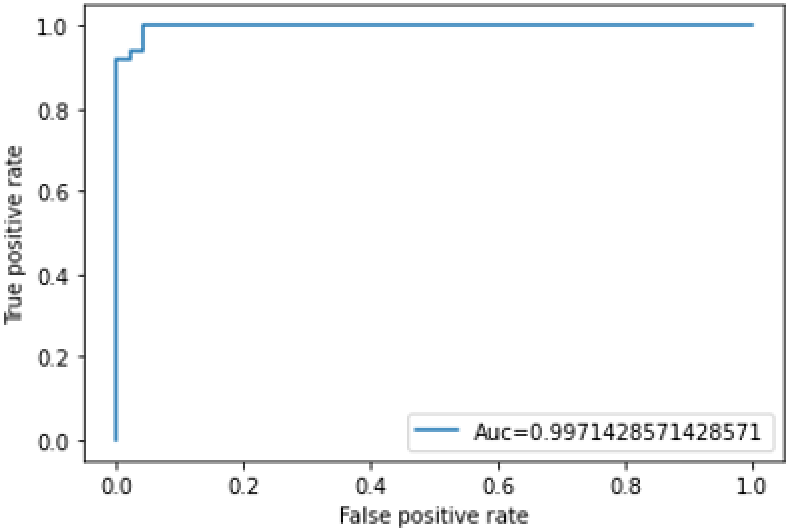


Fig. 9. ROC curve of the proposed model

5 Discussion

In this section, the proposed model is compared with the model without outlier removal and the state-of-the-art methods to prove its robustness.

5.1 Comparison of the proposed model and without outlier removal model

The proposed model is compared with the model Without Outlier Removal (WOR) through data visualization techniques, as shown in Table 3. It is found that the WOR model obtained 91.91% Acc, 92.16% Prec, 92.16% Sen, 91.67% Spec and 92.16% F1. On the other hand, the proposed model obtained the highest Acc, Prec, Spec and F1 of 92.92%, 96%, 95.56% Spec and 93.29%, respectively. Thus, the proposed model outperformed the WOR model.

Table 3. Comparison of the proposed model with WOR

Models	Acc (%)	Prec (%)	Sen (%)	Spec (%)	F1 (%)
WOR	91.91	92.16	92.16	91.67	92.16
Proposed	**92.92**	**96**	**90.74**	**95.56**	**93.29**

5.2 Comparison of the proposed model with state-of-the-art techniques

Table 4 compares the proposed model with state-of-the-art techniques. It is observed that the proposed model outperformed every existing model except the sensitivity of the Time series [16] model because of less proportions of positives being predicted correctly. Apart from this, the proposed model is suitable for lung cancer detection.

Table 4. Comparison with state-of-the-art methods

Models	Acc (%)	Sen (%)	Spec (%)
CNN [17]	77.76	75.35	80.59
Times Series [16]	89.40	91.40	88.20
AdaBoost [15]	90.74	81.80	93.99
DFCNet [17]	86.02	80.91	83.22
CNN [14]	86	84	88
Proposed	**92.92**	**90.74**	**95.56**

6 Conclusion

In this paper, a unique combined approach of data visualization and ML technique has been incorporated. The segmentation, nodule detection and feature extraction were performed through conventional approaches such as adaptive thresholding, label-connected components and nine extracted features, respectively. Then, data visualization techniques were incorporated to visualize and remove the outliers. Four ML classifiers were compared and the proposed RF model showed the best results. It obtained 92.92% accuracy, 96% precision, 90.74% recall, 95.56% specificity and 93.29% F1 score. The proposed model outperformed the non-data visualization-based RF model and the state-of-the-art techniques. Hence, the proposed model can aid radiologists in the early detection of lung cancer.

References

1. Huang, Q., Lv, W., Zhou, Z., Tan, S., Lin, X., Bo, Z., Fu, R., Jin, X., Guo, Y., Wang, H., et al.: Machine learning system for lung neoplasms distinguished based on scleral data. Diagnostics 13(4), 648 (2023)

2. Sebastian, A.M., Peter, D.: Identifying the predictors from lung cancer data using machine learning. In: Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022, pp. 691–701. Springer, (2023)
3. Cao, W., Wu, R., Cao, G., He, Z.: A comprehensive review of computer-aided diagnosis of pulmonary nodules based on computed tomography scans. IEEE Access 8, 154007–154023 (2020)
4. Huang, S., Yang, J., Shen, N., Xu, Q., Zhao, Q.: Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective. In: Seminars in Cancer Biology (2023). Elsevier
5. Yuan, F., Lu, L., Zou, Q.: Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1866(8), 165822 (2020)
6. Xia, C., Dong, X., Li, H., Cao, M., Sun, D., He, S., Yang, F., Yan, X., Zhang, S., Li, N., et al.: Cancer statistics in china and united states, 2022: profiles, trends, and determinants. Chinese medical journal 135(05), 584–590 (2022)
7. Manoharan, S., et al.: Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of lung ct scan images. Journal of Innovative Image Processing (JIIP) 2(04), 175–186 (2020)
8. Li, J., Li, Z., Wei, L., Zhang, X.: Machine learning in lung cancer radiomics. Machine Intelligence Research, 1–30 (2023)
9. Makaju, S., Prasad, P., Alsadoon, A., Singh, A., Elchouemi, A.: Lung cancer detection using ct scan images. Procedia Computer Science 125, 107–114 (2018)
10. Nageswaran, S., Arunkumar, G., Bisht, A.K., Mewada, S., Kumar, J., Jawarneh, M., Asenso, E.: Lung cancer classification and prediction using machine learning and image processing. BioMed Research International 2022 (2022)
11. Kukreja, S., Sabharwal, M., Shah, M.A., Gill, D., et al.: A heuristic machine learning-based optimization technique to predict lung cancer patient survival. Computational Intelligence and Neuroscience 2023 (2023)
12. Baskar, S., Shakeel, P.M., Sridhar, K., Kanimozhi, R.: Classification system for lung cancer nodule using machine learning technique and ct images. In: 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 1957–1962 (2019). IEEE
13. Sivakumar, S., Chandrasekar, C.: Lung nodule detection using fuzzy clustering and support vector machines. International Journal of Engineering and Technology 5(1), 179–185 (2013)
14. Wang, H., Zhou, Z., Li, Y., Chen, Z., Lu, P., Wang, W., Liu, W., Yu, L.: Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18 f-fdg pet/ct images. EJNMMI research 7, 1–11 (2017)
15. Ingle, K., Chaskar, U., Rathod, S.: Lung cancer types prediction using machine learning approach. In: 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp. 01–06 (2021). IEEE
16. Qiu, S., Guo, Q., Zhou, D., Jin, Y., Zhou, T., et al.: Isolated pulmonary nodules characteristics detection based on ct images. IEEE Access 7, 165597–165606 (2019)
17. Masood, A., Sheng, B., Li, P., Hou, X., Wei, X., Qin, J., Feng, D.: Computer-assisted decision support system in pulmonary cancer detection and stage classification on ct images. Journal of biomedical informatics 79, 117–128 (2018)
18. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38(2), 915–931 (2011)