Lung Cancer Detection: A Deep Learning Approach

Siddharth Bhatia, Yash Sinha and Lavika Goel

Abstract We present an approach to detect lung cancer from CT scans using deep residual learning. We delineate a pipeline of preprocessing techniques to highlight lung regions vulnerable to cancer and extract features using UNet and ResNet models. The feature set is fed into multiple classifiers, viz. XGBoost and Random Forest, and the individual predictions are ensembled to predict the likelihood of a CT scan being cancerous. The accuracy achieved is 84% on LIDC-IRDI outperforming previous attempts.

Keywords Lung cancer detection · Deep residual networks · XGBoost Random Forests · Ensemble · Deep learning

1 Introduction

Lung cancer is the leading cause of cancer-related deaths all around the world. One of the important steps in detecting early stage cancer is to find out whether there are any pulmonary nodules in the lungs which may grow to a tumor in recent future. This work aims to determine the likelihood of a given CT scan of lungs to be cancerous. In a nutshell, we employ deep residual networks to extract features from preprocessed images which are fed to classifiers, the predictions of which are ensembled for the final output. We explain in this paper the proposed methodology, evaluation, and results using the LIDC-IDRI dataset.

Rest of the paper is organized as follows. Earlier studies on lung cancer detection have been delineated in Sect. 2. Section 3 explains the dataset used. We explain the various background techniques employed in Sect. 4. Section 5 elaborates on the proposed methodology, preprocessing steps, feature extraction, and classification. The results are further described in Sect. 6. We conclude with future directions in Sect. 7.

S. Bhatia · Y. Sinha (✉) · L. Goel
Department of Computer Science and Information Systems, BITS,
Pilani, Pilani Campus, Pilani, Rajasthan, India
e-mail: mail.yash.sinha@gmail.com

J. C. Bansal et al. (eds.), *Soft Computing for Problem Solving*, Advances in Intelligent Systems and Computing 817, https://doi.org/10.1007/978-981-13-1595-4_55

2 Related Work

Lung cancer detection has earlier been studied using image processing techniques [1–3]. With the advent of neural networks and deep learning techniques, these have recently been used in the medical imaging domain [4–6]. Various researchers [7–12] have tried to classify, detect lung cancer using machine learning and neural networks. Not many deep learning techniques have been applied to detect lung cancer. This is because of the lack of a large dataset for medical images especially lung cancer. Shimizu et al. [13] use urine samples to detect lung cancer.

The technique proposed by Hua et al. [14] simplifies the image analysis pipeline of conventional computer-aided diagnosis of lung cancer. Sun et al. [15] experimented using convolutional neural networks (CNN), deep belief networks (DBNs), and stat denoising autoencoder (SDAE) on Lung Image Database Consortium image collection (LIDC-IDRI) [16]. Their accuracies were 79%, 81%, and 79%, respectively.

The National Lung Screening Trial (NLST), a randomized control trial in the USA, including more than 50,000 high-risk subjects, showed that lung cancer screening using annual low-dose computed tomography (CT) reduces lung cancer mortality by 20% in comparison to annual screening with chest radiography [23].

In 2013, the US Preventive Services Task Force (USPSTF) has given low-dose CT screening a grade B recommendation for high-risk individuals [24], and in early 2015, the US Centers for Medicare and Medicaid Services (CMS) has approved CT lung cancer screening for Medicare recipients [25].

The recent challenge launched "LUNA16" [26] aims to predict the position of nodule in the given lung region. Zatloukal et al. [27] present a study of localization of non-small lung cancer cell with chemotherapy techniques. Zhou et al. [27] present a cancer cell identification technique based on neural network ensembles.

3 Dataset

The Lung Image Database Consortium image collection (LIDC-IDRI) [16] contains diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. It consists of more than thousand scans from high-risk patients in the DICOM image format. Each scan contains a series of images with multiple axial slices of the chest cavity. Each scan has a variable number of 2D slices, which can vary based on the machine taking the scan and patient. The DICOM files have a header that contains the details about the patient id, as well as other scan parameters such as the slice thickness. The images are of size (z, 512, 512), where z is the number of slices in the CT scan and varies depending on the resolution of the scanner.

4 Background of Technology Used

4.1 Deep Residual Networks

Deep residual networks [17] have emerged as a family of extremely deep architectures showing compelling accuracy and nice convergence behaviors. Deep residual networks (ResNets) consist of many stacked "Residual Units." The central idea of ResNets is to learn the additive residual function F with respect to $h(x_l)$, with a key choice of using an identity mapping

$$h(x_l) = x_l. \tag{1}$$

Each subsequent layer in a deep neural network is only responsible for, in effect, fine tuning the output from a previous layer by just adding a learned "residual" to the input. This differs from a more traditional approach where each layer had to generate the whole desired output (Fig. 1).

What's happening is that the F(x) + x layer is adding in, element-wise, the input to the F(x) layer. Here, F(x) is the residual.

4.2 XGBoost Regressor

Extreme Gradient Boosting [18] builds on the premise of "boosting" many weak predictive models into a strong one, in the form of ensemble of weak models well known as Gradient Tree Boosting [19]. There are many gradient tree boosting algorithms, but specifically XGBoost uses the second-order method by Friedman et al. [20, 21] and employs a more regularized model formalization to control over-fitting, which gives it better performance.

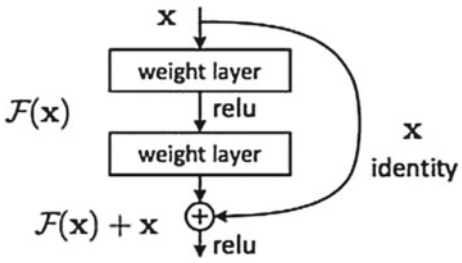


Fig. 1 Residual learning: a building block

4.3 Random Forest Classifier

It is a meta-estimator [22] based on subsampling over many decision trees which controls over-fitting well. The basis of random forest is that randomization over many decision trees can improve the accuracy of the overall classification by boosting the selection rates of features that contribute more toward the classification among others.

5 Proposed Methodology

In a nutshell, we preprocess the CT scan images which are in the DICOM image format to extract the central region of interest of the lungs, which is more likely to have pulmonary nodule. Features are extracted using deep residual networks that are fed into classifiers for supervised learning. For the final output, we ensemble the predictions of various classifiers.

5.1 Preprocessing

The preprocessing step consists of a series of applications of region growing and morphological operations. It identifies and separates the lung structures and nodules to aid the feature extraction. Segmenting lungs from the CT scan aims to identify distinguishing features to aid the classifier and classify the candidates better. This is also important because the CT scan is too huge to be fed into the classifier directly. It will take a lot of time for the classifier to identify differentiating featured from the huge DICOM images.

The segmentation of lung structures is very challenging problem primarily because there is no homogeneity in the lung region. There are similar densities in the pulmonary structures, different scanners, and scanning protocols.

Lung segmentation is followed by normalization and zero centering.

5.2 Feature Extraction

We feed the preprocessed images to ResNet-50 imagenet11k+Places365 feature extractor [17] (Fig. 2).

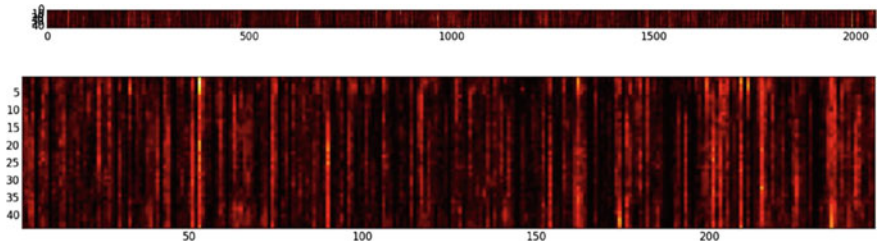
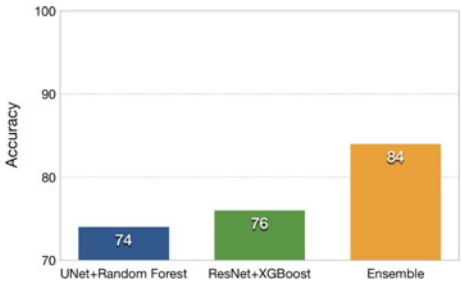


Fig. 2 Visualizing preprocessed features

Fig. 3 % accuracy of various approaches



5.3 Classification and Ensemble

The feature dataset created at the feature extraction stage is fed into a number of classifiers like XGBoost and Random Forest. The results are outlined in Fig. 3. The predictions are ensembled by vote for the final output. The hyperparameters for the classifiers are determined using Grid Search, and the model is tested using 10-fold cross-validation.

6 Results and Inferences

We compare our proposed methodologies in Fig. 3.

We are able to get an accuracy of 84% using an ensemble of UNet+RandomForest and ResNet+XGBoost which individually have accuracies 74% and 76%, respectively.

7 Conclusion

In this paper, we propose an approach to lung cancer detection employing feature extraction using deep residual networks. We compare performance of tree-based classifiers like Random Forest and XGBoost. The highest accuracy we get is 84% using ensemble of Random Forest and XGBoost classifier.

References

1. Palcic, B., et al.: Detection and localization of early lung cancer by imaging techniques. CHEST J. **99**(3) 742–743 (1991)
2. Yamomoto, S., et al.: Image processing for computer-aided diagnosis of lung cancer by CT (LSCT). In: Proceedings 3rd IEEE Workshop on Applications of Computer Vision, 1996. WACV'96. IEEE (1996)
3. Gurcan, M.N., et al.: Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system. Med. Phys. **29**(11) 2552–2558 (2002)
4. Fakoor, R., et al.: Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning (2013)
5. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imag. **35**(5) 1153–1159 (2016)
6. Shen, D., Wu, G., Suk H.-I.: Deep learning in medical image analysis. Ann. Rev. Biomed. Eng. (2017)
7. Cai, Z., et al.: Classification of lung cancer using ensemble-based feature selection and machine learning methods. Molec. BioSyst. **11**(3) 791–800 (2015)
8. Al-Absi Hamada R.H., Belhaouari Samir B., Sulaiman, S.: A computer aided diagnosis system for lung cancer based on statistical and machine learning techniques. JCP **9**(2) 425–431 (2014)
9. Gupta, B., Tiwari, S.: Lung cancer detection using curvelet transform and neural network. Int. J. Comput. Appl. **86**(1) (2014)
10. Penedo, M.G., et al.: Computer-aided diagnosis: a neural-network-based approach to lung nodule detection. IEEE Trans. Med. Imag. **17**(6) 872–880 (1998)
11. Taher, F., Sammouda, R.: Lung cancer detection by using artificial neural network and fuzzy clustering methods. In: GCC Conference and Exhibition (GCC), 2011 IEEE. IEEE (2011)
12. Kuruvilla, J., Gunavathi, K.: Lung cancer classification using neural networks for CT images. Comput. Methods Program. Biomed. **113**(1), 202–209 (2014)
13. Shimizu, R., et al.: Deep learning application trial to lung cancer diagnosis for medical sensor systems. In: SoC Design Conference (ISOCC), 2016 International. IEEE (2016)
14. Hua, K.-L., et al.: Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Therapy **8** 2015–2022 (2014)
15. Sun, W., Zheng, B., Qian, W.: Computer aided lung cancer diagnosis with deep learning algorithms. In: SPIE Medical Imaging. International Society for Optics and Photonics (2016)
16. Armato, S.G., et al.: The lung image database consortium (LIDC) and image data-base resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2) 915–931 (2011)
17. He, K., Zhang, X., Ren, S., Deep, S.J.: residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM (2016)

19. Friedman, J.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001)
20. Friedman, J., Hastie, T., Tibshirani, R., et al.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. **28**(2), 337–407 (2000)
21. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)
22. Liaw, Andy, Wiener, Matthew: Classification and regression by random forest. R news **2**(3), 18–22 (2002)
23. Aberle, D.R., Adams, A.M., Berg, C.D., Black, W.C., Clapp, J.D., Fagerstrom, R.M., Ga-reen, I.F., Gatsonis, C., Marcus, P.M., Sicks, J.D.: Reduced lung-cancer mortality with low-dose computed tomographic screening. N. Engl. J. Med. **365**, 395–409 (2011)
24. Moyer, V.A.: U.S. preventive services task force. Screening for lung cancer: U.S. Preventive services task force recommendation statement. Ann. Int. Med. **160**, 330–338 (2014)
25. Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**, 915–931 (2011)
26. LUng Nodule Analysis (LUNA) Challenge. https://luna16.grand-challenge.org/description/
27. Zatloukal, P., et al.: Concurrent versus sequential chemoradiotherapy with cisplatin and vinorelbine in locally advanced non-small cell lung cancer: a randomized study. Lung Cancer **46**(1) 87–98 (2004)