Physician-assist automated AI lung cancer detection: a narrative review

Kaviya Sathyakumar[1], Michael Munoz[1], Snehal Bansod[1], Jaikaran Singh[1], Benson A. Babu[2]

[1]Saint Johns Episcopal Hospital, Far Rockaway, New York, NY, USA; [2]Lenox Hill Northwell Health, New York, NY, USA

Contributions: (I) Conception and design: BA Babu; (II) Administrative support: BA Babu; (III) Provision of study materials or patients: BA Babu; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: BA Babu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Benson A. Babu, MD, MBA. Lenox Hill Northwell Health, New York, NY, USA. Email: bensonbabumd@gmail.com.

Abstract: Lung cancer is the number one cause of cancer-related deaths in the United States as well as worldwide. Radiologists and physicians experience heavy daily workloads thus are at high risk for burn-out. To alleviate this burden, this literature review compares the performance of four different AI models in lung nodule cancer detection, as well as their performance to physicians/radiologists. In total, 648 articles were extracted from the years ranging from 2008 to 2019. Two independent experienced physician reviewers assisted with selecting the article. Four out of the 648 articles were selected. The inclusion criteria are: 18–65 years old, CT chest scans, lung nodule, lung cancer, deep learning, ensemble, and classic methods. The exclusion criteria are: age greater than 65 years old, PET hybrid scans, CXR, and genomics. The model performance outcomes analysis include: sensitivity, specificity, accuracy, receiver operator characteries curve ROC curve, area under the curve (AUC). The search engines used to extract article from are: PubMed/MEDLINE, EMBASE, Cochrane Library, Google Scholar, Web of Science, IEEEXplore, DBLP. This hybrid deep-learning architecture is state-of-the-art architecture, with a high-performance accuracy and low false-positive reports compared to the others reviewed. Future studies, comparing each model's accuracy and outcomes during clinical validation trials, would be valuable. Automated physician-assist systems such as this hybrid architecture may help preserve a high-quality doctor-patient relationship and reduce physician burn out.

Keywords: Machine learning; convolutional neural networks (CNN); deep CNN (DCNN); lung cancer

Received: 15 March 2020; Accepted: 14 August 2020; Published: 30 September 2020.
doi: 10.21037/jmai-20-24
View this article at: http://dx.doi.org/10.21037/jmai-20-24

Introduction

Lung cancer is the number one cause of cancer-related deaths in the United States and worldwide (1). Furthermore, lung cancer has amongst the highest public burden of cost worldwide. Healthcare cost to Medicare beneficiaries were analyzed (2): the highest costs were related to surgery and an estimated $30,000 over a 15-year period. Similarly, patients receiving chemotherapy and radiation therapy faced a cost of $4,000–$8,000 per month, with an average life expectancy of 14 months from the time of diagnosis (2). Europe's incidence of lung cancer is estimated to be 60 per 100,000 inhabitants. Its costs of healthcare and management for the patient post-intervention are estimated to be 17,000 Euros per year (3).

The National Lung Screening Trial (NLST) found that examination with low-dose computed tomography (LDCT) instead of the standard chest X-ray, in a high-risk population, led to a 20% reduction in mortality rate (4). Additionally, the detection rate of lung cancer screening with low-dose CT is 2.6- to 10-fold higher than that with chest radiography (5). The key to reducing lung-cancer related deaths is early diagnosis and this relies on fast and accurate detection of lung nodules and careful examination of chest CT scans to determine malignancy: a process which requires considerable time and effort on behalf of radiologists and physicians.

According to a recent study, physicians spend 75% of each patient visit on activities other than face-to-face patient encounter (6), including working with the EMR. Studies also found that physicians from various specialties spend up to 2 hours on administrative duties for each hour that they see patients in the office, followed by an additional 1 to 2 hours of work after clinic, mostly devoted to the EMR (7). It is likely, although not investigated, that these figures are much higher for physicians screening patients at risk for lung-cancer, due to the time required for the initial examination and evaluation of CT scans.

During the 18th World Conference on Lung Cancer (WCLC), Dr. Flanou confirmed that oncologists were at highest risk from burn-out compared to other physicians as well as other oncology care staff (nurses, psychologists and social workers), with a reported prevalence between 35–60%. Amongst individuals who suffer burn-out there is a risk of mental health issues in 20–35%, moreover in physicians it is associated with a decrease in empathy towards patients and reduced quality of care (8). It is therefore of utmost importance that all ways in which the burden of work on physicians may be reduced, should be explored, for the wellbeing of both the patients and physicians.

One such solution is AI automated CT lung cancer detection, which can be used to assist physicians: thereby reducing their burden of work; optimizing hospital operational workflow; and providing more time to develop a high-quality doctor-patient relationship. A computer-aided detection (CAD) system was first introduced by Niki *et al*. [2001] as a means to extract and analyze data from CT scans, classify benign and malignant lung cancer changes, and for the purpose of screening patients using 3D CT scans (9). Since then, numerous studies have found improved detection of lung nodules on CT scans when examination by a physician/radiologist is combined with the use of a CAD system (5,10). Improved radiologist performance with CAD was noted especially in the detection of small lung nodules, <5 mm in size, which are often easily overlooked by visual inspection alone (11). Thus, CAD and its associated AI models help not only to reduce the burden of work on physicians, and subsequently fatigue-related errors of judgement, but to improve detection of nodules particularly in the early stages of lung cancer, which are more likely to be missed.

We present the following article in accordance with the MDAR reporting checklist (available at http://dx.doi.org/10.21037/jmai-20-24).

Methods

Method outline

We used the PICO framework defined as (I) problem: lung cancer; (II) intervention: machine learning and deep learning comparison; (III) comparision ensemble CNN vs classic machine learning model performance; (IV) outcomes: sensitivity, measures how well the algorithm recognizes the type of nodule correctly, specificity measures the ability of the algorithm to remove the false positives, and a high specificity value means a low rate of misdiagnosis, accuracy measures the proportion of data that was classified correctly. Sensivity-specificity receiver operator characteristics curve ROC curve and area under the curve AUC were analyzed.

Data bases: PubMed/MEDLINE, EMBASE (or Scopus), Cochrane Library, Google Scholar, Web of Science, IEEEXplore, DBLP. Searched terms strategies used are Boolean and fuzzy logic, truncated terms, and wild card.

In total, 648 articles were extracted. Two independent reviewers selected 4/648 studies: article year range 2008–2019.

Inclusion criteria: 18–65 years old, CT chest scans, lung nodule, lung cancer, deep learning, ensemble and classic methods. Exclusion criteria: greater than 65 years old, PET hybrid scans, CXR, genomics.

Description of the research experiment (12)

In this experiment, a hybrid model was proposed: for this specific task, LeNet, AlexNet, and VGG-16 were used. In addition, the features obtained from the last fully-connected layer of CNNs were applied as input for the following machine learning/classification models: linear regression (LR), linear discriminate analysis (LDA), decision tree (DT), support vector machine (SVM), k-nearest neighbor (kNN) and softmax. All the machine learning classifiers were tested at the end and examined separately by comparing their performance (12). In order to increase the classification accuracy, image augmentation techniques were used during the training of the models (12). In this scope, approximately 20 additional images were obtained from each original sample in the dataset. Lastly, the mRMR feature selection method was used to find the most efficient features, which were then applied as the input in the above-mentioned method (12).

The internal structure of the architecture was composed of convolutional and average pool layers (12). This was followed by a straightener convolutional layer, two fully connected layers and finally a softmax classifier (12). LeNet includes a 5×5 filter (12). Image sizes vary from 32×32×1 to 28×28×6 (12). AlexNet consisted of five convolutional, three pooling and three fully connected layers (12). The convolutional layer was based on the process of circulating a filter over the entire image (12). Filters can be different sizes, such as 3×3, 5×5 and 11×11 (12).

The VGG-16 architecture consisted of a 16-layers network structure (12). The most important feature of this architecture was having an increased network structure. In the VGG-16, the size of the input layer was 224×224 px and the filter size was 3×3 (12). The internal structure of this architecture was composed of five convolutional layers, a pooling layer, and three fully connected layers (12). The final layer was the softmax that was used in the classification tasks (12).

In this experiment, ADAM and RMSProp optimization methods were used for LeNet architecture. In AlexNet and VGG-16 architectures, the SGD method was used (12).

Results and discussion

Analysis of the increase in model performance (12)

The main reason the minimum redundancy, maximum relevance feature selection method with CNN performed better than the methods described in the three other papers, is the use of additional techniques such as image augmentation, principal component analysis (PCA), mRMR and appropriate feature selection (12).

In this method, during the last couple of iterations, the dimensions of the feature set obtained using image augmentation techniques were reduced using PCA before the classification task (12). The KNN classifier was then fed with the reduced feature set, resulting in an accuracy of 97.92% (12). Then, the KNN classifier was fed using the mRMR algorithm with the 1,000 features obtained from the fc8 layer of AlexNet architecture; 33, 50, 100, 150 and 200 of the most efficient features were determined and ranked, respectively (12). The extracted features were reclassified with KNN (12). A 10-fold cross-validation method was used for testing (12).

PCA decreases the classification accuracy from 98.74% to 97.92% (12). The PCA method obtained this level of success with only 33 features and consumed less time when training the model, due to the use of fewer features. In addition, the performance results of the KNN with and without PCA method were close (12).

Next, the most efficient features were selected by the mRMR method of 1,000 features, obtained from the last layer of AlexNet without using the PCA method. The best rate of success obtained was 99.51% with 200 features provided by mRMR (12). It was found that using 100, 150 or 200 features from the mRMR algorithm, was more successful than using all 1,000 features obtained from the fc8 layer of AlexNet (12).

After this point, the experiment was extended by focusing on the KNN classifier. In this scope, the k value corresponding to the number of the nearest neighbors was searched in the range of 100 and 102 considering various distance functions by using the Bayesian optimization method (12). Notably, the classification success decreased relatively and gradually as the k value increased (12). The most efficient results were ensured for KNN when the k was set to 1 and the distance function was adjusted to Correlation. In this experiment, the 10-fold cross-validation was also used for evaluation (12). The model achieved an accuracy of 99.51%, sensitivity of 99.32%, specificity of 99.71% and F-score of 99.51% (12).

Literature on performance comparisons

In a different research team model comparison da Silva *et al.* 2018 (13), their convolutional neural network based particle swarm optimization (PSO) for lung nodule by achieved an a lower accuracy of 97.62% in comparison to the described Toğaçar *et al.* 2020 (12). model, mainly because no image augmentation and feature selection technique was used before the CNN architecture, whereas Toğaçar *et al.* 2020 (12) used the mRMR technique to achieve a model accuracy of 99.51% (12).

Another research group Jung *et al.* 2018 (14) model performs lower in accuracy of 96.30 % in comparison with the Toğaçar *et al.* model accuracy of 99.51%, similarly because no image augmentation and feature selection

Table 1 Az values for performance in detecting all nodules, from Awai *et al*. [2004]

Observer No.	Without CAD output	With CAD output
Board-certified radiologists		
1	0.49	0.52
2	0.64	0.70
3	0.74	0.79
4	0.59	0.61
5	0.66	0.67
Radiology residents		
6	0.65	0.70
7	0.70	0.76
8	0.56	0.57
9	0.75	0.78
10	0.64	0.65

technique, whereas Toğaçar *et al*. 2020 (12) used the mRMR technique with better results. Furthermore, Jung *et al*. 2018 research team developed a 3-dimensional ensemble CNN which required more training data and computational power requirements to run (14).

Lastly, a comparison of the final research group Lyu *et al*. 2018 (15) their multi-level CNN classification of lung nodules had a lower model performance accuracy of 84.41% mainly because no image augmentation nor feature selection method was used in their CNN architecture than compared the mRMR used by Toğaçar *et al*. 2020 (12).

Numerous studies assessing the performance of radiologists in lung nodule detection show low inter-observer agreement, varying sensitivities ranging from 30–97%, and false positive counts of 0.6–2.1 per patient, depending on the input data, method and criteria for identification (16). A study from the NLST, assessed CAD retrospectively in 134 subjects and found an improved inter-observer agreement (kappa increase from 0.53–0.66): results confirmed by similar studies (16). As well as reducing inter-observer variation, one of the greatest advantages of CAD remains the detection of smaller lung nodules that are easily missed by radiologists/physicians (11). The use of CAD by 2 radiologists in an emergency clinic study, did find improved reading time when CAD was used (Radiologist 1: 94.6 *vs.* 102.7 s, P>0.05; Radiologist 2: 61.1 *vs.* 76.5 s, P<0.05). Although this decrease in reading time was not statistically significant for both radiologists, they did get a significantly improved rate of nodule detection: 34% and 27% for Radiologists 1 and 2 respectively when CAD was reviewed after the CT images, but not when it was reviewed before the scans (10).

An observer performance study compared the performances of 10 radiologists without and with the use of CAD, in 50 CT examination cases (5). Alternative free-response ROC curves for each output (with and without CAD) were calculated by plotting the true-positive fraction against the likelihood of obtaining an image with false-positive findings (i.e. with one or more false-positive lesions) at each confidence level. Using the area under each alternative free-response ROC curve (Az) to compare the observers' performances, they found that the performance of all observers was significantly improved with the use of CAD (see *Table 1*).

Routine used of CAD by radiologists and physicians, especially in high-pressure environments, is justified due to improved rates of lung nodule detection, inter-observer agreement, interpretation speed, higher true-positive to false-positive ratios and for detection of small (<5 mm) nodules.

Limitations and further improvement

The experiment conducted here performs well (see *Table 2*), but it uses a very small dataset thus, may not perform well on a large production scale. Ideally, the models should be tested on a larger dataset to ensure they work on large, real production data. Also, the image augmentation method was used here to increase the number of images: these techniques may create very correlated images which can lead to overfitting. Another indication of correlation might be the KNN algorithm, which relies on the nearest neighbor, as it performed best on this dataset. It would be beneficial to further test these models on new dataset, which is relatively large and from a different data source. Also, the test dataset should not undergo image augmentation, but be tested in its original form.

Table 2 Comparison of AI experimental models to detect lung cancer

Experiment	Performance metric	Justification
Minimum redundancy, maximum relevance feature selection method on chest CT images with convolutional neural networks (12)	99.51%	CNN in conjunction with additional techniques: Image augmentation PCA mRMR and appropriate feature selection.
Convolutional neural network (CNN)-based PSO for lung nodule, false positive reduction of CT images (13)	97.62%	No image augmentation No feature selection techniques before CNN
Classification of lung nodules in CT scans using three-dimensional deep CNNs with a checkpoint ensemble method (14)	96.30%	No image augmentation No feature selection techniques before CNN Uses an ensemble with 3D: requiring more training datasets
Multi-level CNN for classification of lung nodules on CT images (15)	84.81%	No image augmentation No feature selection techniques before CNN

PSO, particle swarm optimization; PCA, principal component analysis.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the MDAR reporting checklist. Available at http://dx.doi.org/10.21037/jmai-20-24

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jmai-20-24). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Barta JA, Powell CA, Wisnivesky JP. Global Epidemiology of Lung Cancer. Ann Glob Health 2019;85:8.
2. Sheehan DF, Criss SD, Chen Y, et al. Lung cancer costs by treatment strategy and phase of care among patients enrolled in Medicare. Cancer Med 2019;8:94-103.
3. Ausweger C, Burgschwaiger E, Kugler A, et al. Economic concerns about global healthcare in lung, head and neck cancer: meeting the economic challenge of predictive, preventive and personalized medicine. EPMA J 2010;1:627-31.
4. National Lung Screening Trial Research Team, Aberle DR, Adams AM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011;365:395-409.
5. Awai K, Murao K, Ozawa A, et al. Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance. Radiology 2004;230:347-52.
6. Young RA, Burge SK, Kumar KA, et al. A Time-Motion Study of Primary Care Physicians' Work in the Electronic Health Record Era. Fam Med 2018;50:91-9.
7. Sinsky C, Colligan L, Li L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016;165:753-60.
8. Chustecka Z. Oncology Most Stressful of Specialties: High Risk for Burnout. Medscape Medical News Online, 18th World Conference on Lung Cancer (WCLC),

2017. Available online: https://www.medscape.com/viewarticle/887230#vp_1

9. Niki N, Kawata Y, Kubo MA et al. CAD system for lung cancer based on CT image. International Congress Series 2001;1230:631-8.
10. Mozaffary A, Trabzonlu TA, Lombardi P, et al. Integration of fully automated computer-aided pulmonary nodule detection into CT pulmonary angiography studies in the emergency department: effect on workflow and diagnostic accuracy. Emerg Radiol 2019;26:609-14.
11. Sahiner B, Chan HP, Hadjiiski LM, et al. Effect of CAD on radiologists' detection of lung nodules on thoracic CT scans: analysis of an observer performance study by nodule size. Acad Radiol 2009;16:1518-30.
12. Toğaçar MM, Ergen B, Cömert Z. Detection of lung cancer on chest CT images using minimum redundancy maximum relevance feature selection method with convolutional neural networks. Biocyb Biomed Engineering 2020;40:23-39.
13. da Silva GLF, Valente TLA, Silva AC. Convolutional neural network-based PSO for lung nodule false positive reduction on CT images. Comput Methods Programs Biomed 2018;162:109-18.
14. Jung H, Kim B, Lee I, et al. Classification of lung nodules in CT scans using three-dimensional deep convolutional neural networks with a checkpoint ensemble method. BMC Med Imaging 2018;18:48.
15. Lyu J, Ling SH. Using Multi-level Convolutional Neural Network for Classification of Lung Nodules on CT images. Conf Proc IEEE Eng Med Biol Soc 2018;2018:686-9.
16. Rubin GD. Lung nodule and cancer detection in computed tomography screening. J Thorac Imaging 2015;30:130-8.

doi: 10.21037/jmai-20-24

Cite this article as: Sathyakumar K, Munoz M, Bansod S, Singh J, Babu BA. Physician-assist automated AI lung cancer detection: a narrative review. J Med Artif Intell 2020;3:14.