JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DETECTION OF LUNG CANCER USING DEEP LEARNING

Erram Ghani, Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

A.Manikanta, Assistant Professor, Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

Dr. Dinesh Kumar Rangarajan, Professor & HoD, Department of CSE, Siddhartha Institute of Technology & Sciences, Telangana.

Dr. JBV Subrahmanyam, Professor & Principal, Siddhartha Institute of Technology & Sciences, Telangana.

Abstract - Lung cancer is one of the leading causes of mortality in every country, affecting both men and women. Lung cancer has a low prognosis, resulting in a high death rate. The computing sector is fully automating it, and the medical industry is also automating itself with the aid of image recognition and data analytics. Lung cancer is one of the most common diseases for human beings everywhere throughout the world. Lung cancer is a disease which arises due to growth of unwanted tissues in the lung and this growth which spreads beyond the lung are named as metastasis which spreads into other parts of the body.

The objective of our project is to inspect accuracy ratio of two classifiers which is Support Vector Machine (SVM), and K Nearest Neighbour (KNN) on common platform that classify lung cancer in early stage so that many lives can be saving. The experimental results show that KNN gives the best result Than SVM. This report discusses the Implementation details of our project.

We have done data preprocessing, data cleaning and implements machine leaning algorithm for prediction of lung cancer at early stages through their symptoms. We have used both classification algorithms to find or predict the accuracy ratio.

Lung cancer is identified as the most common cancer in the world that causes death. Early detection has the ability to reduce deaths by 20%. In the current clinical process, radiologists use Computed Tomography (CT) scans to identify lung cancer in early stages. Radiologists do so by searching for regions called ‘nodules’, which correspond to abnormal cell growths. But identifying process is time consuming, laborious and depends on the experience of the radiologist. Hence an intelligent system to automatically assess whether a patient is prone to have a lung cancer is a need.

This paper presents a novel method which use deep learning, namely convolutional neural networks (CNNs) to identify whether a given CT

scan shows evidence of lung cancer or not. The implementation uses a combination of classical feature-based candidate detection with modern deep-learning architectures to generate excellent results better than either of the methods. The overall implementation consists of two stages. Nodule Regions-of-Interest (ROI) extraction and cancer classification. In nodule ROI extraction stage, we select top most candidate regions as nodules. A combination of rule based image processing method and a 2D CNN was used for this stage. In the cancer classification stage, we estimate the malignancy of each nodule regions and hence label the whole CT scan as cancerous or non-cancerous. A combination of feature based eXtreme Gradient Boosting (XGBoost) classifier and 3D CNN was used for this stage. The LUNA dataset and LIDC dataset were used for both training and testing. The results were clearly demonstrated promising classification performance. The sensitivity, accuracy and specificity values obtained for the nodule ROI segmentation and cancer classification showed to be improved in the combined approach of deep learning with classical feature based classifiers compared to the deep-learning only techniques.

Keywords—Computed Tomography, Lung Nodules, Convolutional Neural Networks, Deep Learning, Segmentation, Classification.

1. INTRODUCTION

In the world, cancer will strike two in every five people in their lifetimes. Lung cancer is the most common and number one cause of cancer deaths in both men and women [1]. Lung cancer related deaths accounts for 27% of all cancer related deaths. Unfortunately, most people who develop lung cancer do not develop symptoms until it has become more advanced. The result is late diagnosis, where treatment can be effective, but rarely curative. Early detection is critical for lung cancer since it opens a range of treatment that is not available at later stages. It has been estimated that early detection has the potential to reduce the lung cancer related deaths by 20%. Typically lung cancers are characterized by pulmonary nodules. Pulmonary nodules are regions corresponding to uncontrolled cell growths. The pulmonary nodules can be attached to the lung wall, bronchus, bronchioles and blood vessels or may be isolated within lung region. Computed Tomography (CT) is breakthrough technology for pulmonary nodule detection that heavily used in current clinical process.

CT imaging has the ability to form three dimensional images of the chest with greater resolution of nodule and tumor pathology. A typical lung CT scan is a 3D image of approximate size 30cm ×30cm×40cm. But the nodules we are interested in malignancy estimation for lung cancer detection are confined to very small regions with diameter ranges from 3mm-40mm. Therefore examining the presence of pulmonary nodule in a lung CT scan is literally similar with finding a needle in haystack problem. Figure 1 depicts a pulmonary nodule identified in a CT scan slice. Those nodule regions looks very similar to tiny bronchioles and blood vessels with their intensities and shapes when viewed in 2D slices.

In current clinical process chest radiologists have to manually inspect through the whole CT scan slice by slice to spot out the presence of a pulmonary nodule. The radiologist has to work with approximately 1:1000000 signal to noise ratio to extract out nodules. And for the malignancy estimation of extracted nodules they have to manually extract the nodule contours and calculate shape and texture parameters like area, growth rate, and shape etc. Therefore lung cancer examining for one CT scan roughly takes 30-50 minutes .This makes manual lung cancer diagnosis process into tedious time consuming task which require the expertise domain knowledge in lung anatomy. Yet manual annotations results in lots of false positives and may miss potential nodules.

Therefore computer processing to assist lung cancer detection is one of most timely need of the world. Here we proposed an intelligent system for

binary classification problem to detect the presence of lung cancer in a given patient-CT-lung scan. The proposed system use combination of deep learning, feature based machine learning and rule based image processing to assess whether a given CT scan is of cancerous patient or not. It will output top candidate regions for nodules, common shape and texture parameters for each nodule region-of-interest (ROI), malignancy estimation for each nodule ROI and overall malignancy of the whole CT scan. Such a system will dramatically reduce the false positive rate that plagues the current detection technology. And help to find nodules missed by human error.

Regardless of the malignancy outcome, automatic nodule detection can be a big help for radiologists since the nodules can easily be overlooked

And it will help to get patients to access life-saving interventions earlier. It will give radiologists more time to spend with their patients, and the system avoids additional follow-up imaging and interventional treatment. It will be advancing the state of the art in future screening, care and prevention.

The proposed system highlighting its novelty and originality since it use combination of classical feature based classifiers and deep learning classifiers namely Convolutional Neural Networks (CNN) to produce results. Convolutional Neural Network is very popular architecture of deep neural networks which has become the state-of-the-art technology of current computer vision domain. Yet applying it in the domain of medical image processing is very recently being grabbed the attention by research community. Possibility of applying deep learning for cancer classification of CT volumes is very recently examined in 2017 Kaggle Data Science Bowl Competition. Other than that there is no significant research efforts which combining both nodule detection and cancer classification outputs with significant improvement in sensitivity number of false positives and level of automation.

2. LITERATURE SURVEY

The research efforts for lung nodule detection and cancer classification of CT volumes can be divided into two categories. Those are feature based machine learning approaches and deep learning based approaches. Both approaches use image processing techniques like thresholding, rescaling, morphological operations, filtering and segmentation methods to perform some preprocessing tasks. Most of classical feature based approaches first segment the lung, extract features from training data and train a classifier to detect nodule regions. Then shape and texture features are extracted from nodule regions to train a classifier for malignancy classification. Support vector machines (SVM), random forest classifier, XGBoost classifiers k-nearest neighbor (kNN), logistic regression like classifiers are engaged as classifiers. But nodule detection classifiers learned using hand-engineered features often give poor generalization to novel test data. This is because lung nodule detection is inherently more challenging due to the high variability of nodule shape, size and texture. More recently advances in deep learning have enabled extraction of high level features without expertise domain knowledge in lieu of hand engineered features.

Deep learning is the branch of machine learning based on deep neural networks which are neural networks composed of more than one hidden layer. Among the most popular architectures of deep neural networks, CNNs are of particular interest which are widely used in computer vision. E.g.: image classification, super-resolution, semantic segmentation. Recent publications report their usage in medical image segmentation and classification with promising results. For example U-net: Convolutional networks for biomedical image segmentation by O. Ronneberger, P. Fischer, and T. Brox outperforms the prior best methods for segmentation of neuronal structures in electron microscopic stacks in ISBI challenge 2015. And K. Komnitsas et al. use 3D CNN for Brain lesion segmentation in multi-channel MRI patient data with top ranking performance on the

public benchmarks BRATS 2015 and ISLES 2015[18].

Deep learning for lung cancer classification very recently grabbed the attention in research community after 2017 Kaggle Data Science Bowl Competition. The winning team of the competition use 3D CNN classifiers for both nodule segmentation and cancer classification. A similar work done by Julio Cesar Mendoza use CNN for lung nodule classification which show that their method is outperforming a base feature-engineering method using the same techniques for other stages. Rushil Anirudh et al proposed 3D CNN for lung nodule detection which works with the availability of weakly labelled data. Their system works with point labels, which specify a single voxel location that indicates the presence of a nodule, and it's largest cross sectional area.

We proposed a method combining both classical feature based classifiers and CNNs to produce more accurate results. To the best of our knowledge we are the first to explore lung nodule detection and cancer classification by combined methods which exploit the advantages of each method to generate more accurate results.

Machine learning involves several algorithms such as k-Nearest Neighbours (KNN), support vector machine (SVM), Naive Bayes (NBs), classification tree (C4.5), gradient boosting machines (GBM), etc. While each of these algorithms processes data differently, in this section, a few recently proposed machine learning candidates in the area of malignant growth finding are reviewed chronologically.

Chen et al. (2013) presented a fuzzy system using KNN (FKNN) for Parkinson's disease (PD) diagnose. Besides, they used the principal component analysis to find the most discriminative features on which the optimal FKNN model was built. They compared their system with the SVM algorithm and found that their proposed method performed better. The best classification accuracy of their FKNN reached to 96.07%.

Odajima & Pawlovsky (2014) declared that the precision of the KNN method changes with the number of neighbours and with the level of information utilized for classification. Meanwhile, they showed details about the variation of the maximum and the minimum values of the accuracy with the classification set sizes and the number of neighbours.

Lynch et al. (2017) applied some supervised learning classification techniques such as linear regression, decision trees, GBM, SVM, and a custom ensemble to the SEER database to order lung cancer patients regarding survival. The outcomes demonstrated that among the five individual models used, the most precise was GBM with a root mean square error (RMSE) value of 15.32.

Septiani et al. (2017) compared the performances of C4.5, NBs, and KNN classification algorithms to detect breast cancer diagnosis on 670 data, each with 9 attributes. They showed that while NBs and kNN have the same accuracy of 98.51%, C4.5 is the worst with the accuracy equal to 91.79%.

3. SYSTEM ANALYSIS:

Our task is a binary classification problem to detect the presence of lung cancer in patient lung CT scans. Here we exploit the use of deep learning, particularly 2D and 3D convolutional neural networks. But at each stage of the procedure, we followed classical feature based machine learning methods and rule based image processing approaches beside the main deep learning architecture and combined the outputs together to generate more accurate results. Below describes the technological approach we have taken.

Usually a CT scan of lung area is a 3D image of size 30cm × 30cm ×40cm which contains about 100-400 slices in transverse view. But the nodules, interested in malignancy estimation is restricted to regions of 3mm- 30mm diameter resulting a very low signal to noise ratio near 1: 1000 000 leading to a literally finding a needle in

haystack problem. Therefore any deep learning architecture will not able to learn something from the raw image data. Given the lung CT scan and the cancer label of that scan, it will unable to find a direct relation between the label and the input scan. Hence our solution consists of two major parts, first extracting the nodule regions of interest thus increasing signal to noise ratio to a great extent, secondly classifying the overall malignancy based on the regions extracted above.

A detailed block diagram explaining the methodology of the system is shown in Figure 3.1. It can be viewed as a combination of several major steps namely image acquisition, pre-processing, nodule ROI extraction, false positive reduction, feature visualization and cancer classification.

For the implementation we used Python 3.5 language, NumPy [1] for N-dimensional array handling, Skimage python library [2] for image processing, SimpleITK [3] for medical image analysis and H5Py [4] for efficiently storing and manipulating huge amount of data.

4. OUTPUT RESULTS:

A) Nodule ROI Extraction

To measure the performance of the nodule extraction stage of our pipeline we calculated several area measures based on number of pixels of interested region. Four parameters; true positive (TP), false positive (FP), true negative (TN) and false negative (FN) are calculated by the logical AND between ground truth mask and predicted binary mask. Then the sensitivity, specificity, precision and F1-score values were calculated based on these parameters as below.

$$Sensitivity = TP/(TP + FN) \quad (8)$$

$$Specificity = TN/(TN + FP) \quad (9)$$

$$Precision = TP/(TP + FP) \quad (10)$$

$$F1 - Score = \frac{2 \times sensitivity \times precision}{sensitivity + precision} \quad (11)$$

For the performance analysis of this stage 29 test scans was used. First we took the results only from the deep learning segmentation approach and followed through the false positive reduction step to calculate the area measures. Then we recalculated these measures for the results obtained from the ensemble of both deep learning and feature based segmentation methods. The mean values obtained are summarized in Table 4.1.

It can be clearly seen that combination of the two approaches results in higher performance in all measures. To test whether this improvement is significant statistically, we followed a t-test. By comparing P-value 3 measures among 4 have that below 0.05. Thus indicating a significant difference.

B) Cancer Classification Stage

For the 3D CNN developed for cancer classification, after iterating 55 times on the training dataset we finally achieved a training accuracy of 77.8% and validation accuracy of 79.3%. The training accuracy and the test accuracy variation at various stages of training procedure is shown in Figure 4.3. Note that reasons behind the validation accuracy leading the test accuracy are applying dropout layers and the accuracy calculation procedure in Keras' APIs.

C) Overall Results

The accuracy values stated above are for per nodule cancer classification. Then we calculated performance measures for per scan cancer classification. For this purpose we calculated cancer labels for each scan in LUNA16 as one if at least one nodule contained in the scan is cancerous and zero otherwise. The results for 222 test scans including 67 cancerous scans are indicated in table 4.2

5. CONCLUSION

We have presented a work for lung cancer classification from lung CT volumes using combined method of classical featured based classifiers and CNNs. While the initial results are

promising there are areas to further improve the system.

The objective of our model is to predict and classify lung cancer in early stages. For this we have applied two classification algorithm SVM and KNN on Common platform or common dataset to compare their accuracy ratio.

We have found that accuracy of KNN is more than the SVM , so we can say that KNN will classify more precisely and accurately lung cancer in early stages. The models were able to retain accurate prediction even with a very small number of features selected.

This is the most effective model to predict patients with lung cancer disease. This project could answer complex queries, each with its own strength with respect to ease of model interpretation, access to detailed information and accuracy.

FUTURE SCOPE

In future the advanced level of algorithm is used to increase the level of prediction while we are in process to include two more Classification Algorithm the to use the data set more effectively.

We are also planning to make our model efficient and more accurate by using feature selection method.

REFERENCES

- G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
- J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
- I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- K. Elissa, “Title of paper if known,” unpublished.
- R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
- Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.