Detection of Lung Cancer Using Convolutional Neural Networks

Dosbol Aliev daliev@cougarnet.uh.edu
Kurmanbek Bazarov kbazarov@cougarnet.uh.edu
Michael Moorman mmoorman@cougarnet.uh.edu
Abraar Patel aipatel2@cougarnet.uh.edu
Nouhad Rizk njrizk@uh.edu

ABSTRACT

Cancer detection in the early stages is a significantly useful tool for decreasing cancer mortality and improving outcomes for patients. This research focuses on comparing implementations of machine learning algorithms to classify different types of lung cancer. The model utilizes pre-processed and regularized axial chest CT scans of patients with either one of three non-small-cell lung carcinomas, or no cancer. Traditional supervised learning models Support Vector Classifiers (SVC) and Random Forest Classifiers (RFC) are compared to deep learning model using Convolutional Neural Networks (CNN) by comparing their respective model accuracies and F1 scores, with special care given to false positive and false negatives in cancer detection. The study uses different techniques involving hyperparameter tuning, Dropouts, and regularization techniques to control overfitting in the CNN model in order to get an efficient model which makes accurate predictions of lung cancer detection. The SVC and RFC models are also tuned using k-fold cross-validation to counteract overfitting.The results of the models is compared to that of the current standard of care.

Keywords: Chest CT-Scans, lung cancer, machine learning models, Support Vector Classifiers, Random Forest Classifers, deep learning models, Convolutional Neural Networks, hyperparameter tuning, Dropouts, overfitting.

SUBJECT

This research incorporates the field of machine learning and deep learning to detect lung cancer in patients. Images of the CT scan dataset are classified using CNN methods. Convolutional Neural Networks (CNN) uses image convolution on multiple kernels in order to detect features in the image with a strong correlation to improved model accuracy, reinforcing these through backpropagation.

A comparative analysis of deep learning models is conducted and CNN is explored to classify the chest CT-Scan images and diagnose if the patient has lung cancer or not. This study applies hyperparameter tuning on the models to get the best model accuracy and minimum loss. The study applies machine learning algorithms like Support Vector Classifier(SVC) and Random Forest Classifer (RFC) to discover the most efficient machine learning models by comparing the model accuracies. Overfitting is controlled in the models by using the Dropout layer and applying L2 regularization, The main goal of this research is to classify the chest CT-Scan images into different categories of lung cancer.

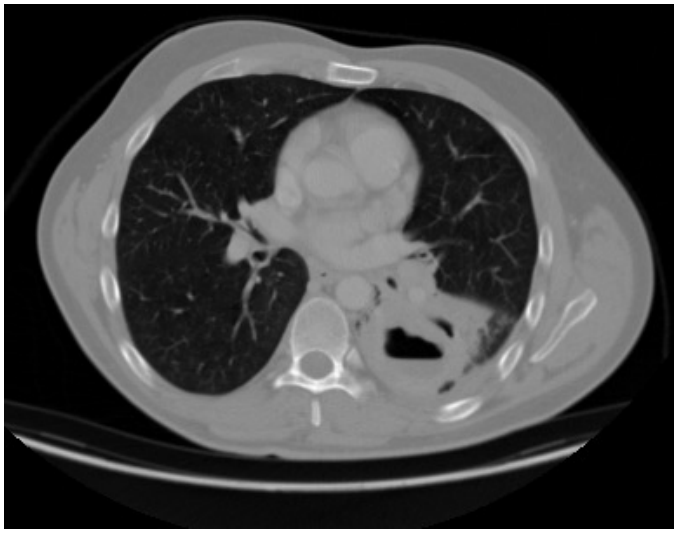

Fig. 1. An example image from the dataset

RELATED WORK

Techniques of convolutional neural networks have been used in medical imaging for quite some time [7]. In 2018, early devices for performing diagnosis of a condition using the output of convolutional neural networks were approved for sale by the FDA [5]. In this case, the device is attempting to use retinal scans to diagnose diabetic retinopathy at a point earlier than doctors typically are able to discern it. It is surely likely that such devices will become more common as neural network technology becomes more accurate. Chest CT is the best imaging modality that detects different parenchymal patterns and disease severity in COVID-19 patients [6] . The most endorsed screening test for detecting lung cancer is CT scan. CT Scan uses X-ray and computer technology to display full images of structures inside the chest and provide more information about chest related injuries. Back in 2008, lung cancer accounted for 18 percent of deaths worldwide. As the lung cancer epidemic has grown and spread, ways of detecting the disease earlier, to improve the cure rate, have been explored. These have mainly been based around imaging using the chest radiograph (CXR) and computed tomography (CT) [3].

MOTIVATION

Cancer detection is traditionally done by trained oncologists interpreting the results of many tests, including those of

medical scans such as CT, MRI, and PET scans. However, the possibility of using deep learning models to recognize the presence or absence of cancer based on training on large datasets of scan images, with accuracy equal or greater than that of medical professionals, would be a significant asset. This would lead to improved patient outcomes as well as freeing up resources of medical professionals, which could have significant benefits in areas of the world where experts in cancer detection are scarcer and their time is more precious.

Aside of cancer detection, the research project also seeks to evaluate the ability to distinguish between three different kinds of lung cancer. In particular, the Chest CT-Scan dataset contains 4 categories of chest CT-scan images. These are adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and non-cancerous. Adenocarcinoma is the most common form of lung cancer accounting for 30 percent of all cases and around 40 percent of all non-small cell lung cancer occurrences. Adenocarcinoma of the lung is found in outer region of the lung. Large cell carcinoma is a type of lung cancer which accounts for 10 to 15 percent of all cases of lung cancer. Large cell carcinoma tends to grow and spread rapidly and can be found anywhere in the lungs. Squamous cell carcinoma is a type of lung cancer which is generally linked to smoking and is found in the center of the lung, where larger bronchi join main airway branches. Normal chest CT scan image shows that no lung cancer is detected. These different types of cancer have different presentations and different prognoses; in particular, adenocarcinomas often have a mutation in ALK proteins that have been targeted with specific cancer medications [1]. Conversely, squamous-cell carcinomas have been harder to treat and require different methods [2]. The utility of being able to distinguish different kinds of cancerous tumors in the class of non-small-cell lung carcinomas is therefore quite useful as well.

Evaluation

The main goal of the research is to compare traditional machine learning to deep learning techniques like Convolutional Neural Networks and analyze whether a Convolutional Neural Network has superior accuracy over non-deep learning techniques. One aspect of a successful outcome for this project overall is for the models to attain higher than 70 percent accuracy; that is, for the model to classify at least 70 percent of previously-unseen testing images correctly. This would demonstrate that these techniques are sufficient to perform at this degree of precision at real world data.

The F1 score is of particular use as the relative frequency of the different cancers is quite disparate both in the dataset and in the larger population, leading to misleading results when considering accuracy alone. In particular, distinguishing the relatively rare squamous cell and large cell carcinomas from adenocarcinoma is of interest, as they favor different treatment routes when discovered early. Conversely, the model should have extremely high accuracy in classifying cancerous versus non-cancerous lungs, as the determination that cancer is present is the most consequential aspect of the model's determination. A model with a high false negative rate for cancer detection would be useless as diagnostic tool.

To get such high accuracy, Dropout Layers are used to minimize training accuracy while maximizing the testing accuracy by introducing bias and reducing variance. The study's use of hyperparameter tuning by randomized search of the parameter space increases the chances of getting the highest precision to predict whether a patient has lung cancer or not.

The comparison against other machine learning tools such as Support Vector Classifiers and Random Forest Classifiers, is to gauge the benefit that deep learning using neural networks bring over other methods and to provide an instructive analysis of techniques in neural network classification.

Resources

The Chest CT-scan image dataset was compiled by Mohamed Hany from public domain sources, and made available under the Open Data Commons Open Database License (ODbL) v1.0 [4]. Dataset preprocessing was done using the Numpy [8] and OpenCV [10] libraries for numerical computing and computer vision. The construction and training of the CNN model utilized Google's Tensorflow framework [12]. Hyperparameter tuning was performed with the KerasTuner framework [9]. The implementations of Support Vector Machines, Random Forest and other classifiers for comparison with the CNN model all were provided through the scikit-learn library [11].

Contributions

Michael Moorman: Was involved in scheduling team meetings regularly and listing out the responsibilities of the team members. Lead data preprocessing by resizing the images. Michael identified a workflow for saving the models on Google Colab so that the models could be shared between team members efficiently without having to retrain them everytime. Trained CNN model using Sequential API. On the writing part, Michael focused on listing out the resources used for the project, describing why the project is interesting and including visuals of the dataset. Included the necessary references and found related work on chest CT scans.

Kurmanbek Bazarov: Was involved in applying hyperparameter tuning on the models and interpreting results in an attempt to determine the best model which could make accurate prediction of lung cancer detection and distinguish the different types of lung cancer correctly. On the writing part, Kurmanbek focused on writing the abstract for the project.

Abraar Patel: Was involved in setting up the Google Colab for the research project. Set up google drive by uploading the dataset and sharing the google drive folder with all the team members so that they could access the drive folder by mounting the google drive on Colab. Loaded the dataset on Google Colab using Kaggle API. Moreover, Abraar focused on building machine learning algorithms like SVC for image classification and interpreting the results . On the writing part, Abraar was involved in describing the dataset and specifying the machine learning and deep learning techniques used to

solve the image classification problem. Came up with the title of the project proposal. Was involved in finding some related work on chest CT scans and included the necessary references based on the related work.

Dosbol Aliev: Was involved in applying regularization techniques and using the Keras Dropout layer to control possible overfitting to improve the overall model accuracy and minimize the loss. On the writing part, Dosbol focused on describing the successful outcomes of the project.

REFERENCES

[1] K. C. Arbour and G. J. Riely. "Diagnosis and Treatment of Anaplastic Lymphoma Kinase-Positive Non-Small Cell Lung Cancer". In: *Hematol Oncol Clin North Am* 31.1 (Feb. 2017), pp. 101–111. DOI: 10.1016/j.hoc.2016.08.012.

[2] B. A. Derman, K. F. Mileham, P. D. Bonomi, et al. "Treatment of advanced squamous cell carcinoma of the lung: a review". In: *Transl Lung Cancer Res* 4.5 (Oct. 2015), pp. 524–532. DOI: 10.3978/j.issn.2218-6751.2015.06.07.

[3] Saeed Mirsadraee Edwin JR van Beek and John T Murchison. "Lung cancer screening: Computed tomography or chest radiographs?" In: *World J Radiol* 7.8 (Aug. 2015), pp. 189–193. DOI: 10.4329/wjr.v7.i8.189.

[4] Mohamed Hany. *Chest CT-Scan images Dataset*. URL: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.

[5] *IDX-dr*. June 2022. URL: https://www.digitaldiagnostics.com/products/eye-disease/idx-dr/.

[6] Mohamed N.E. Kassem and Doaa T. Masallat. "Clinical Application of Chest Computed Tomography (CT) in Detection and Characterization of Coronavirus (Covid-19) Pneumonia in Adults". In: *J Digit Imaging* 34.2 (Feb. 2021), pp. 273–283. DOI: 10.1007/s10278-021-00426-5.

[7] S.-C.B. Lo, S.-L.A. Lou, Jyh-Shyan Lin, et al. "Artificial convolution neural network techniques and applications for lung nodule detection". In: *IEEE Transactions on Medical Imaging* 14.4 (1995), pp. 711–718. DOI: 10.1109/42.476112.

[8] *numpy: Scientific computing with Python*. URL: http://numpy.org.

[9] Tom O'Malley, Elie Bursztein, James Long, et al. *KerasTuner*. https://github.com/keras-team/keras-tuner. 2019.

[10] *OpenCV computer vision library*. URL: http://opencv.org/.

[11] *scikit-learn*. URL: https://scikit-learn.org/stable/.

[12] *Tensorflow Machine Learning Framework*. URL: https://www.tensorflow.org/.