LUNGSTAT: IMPROVING LUNG CANCER DIAGNOSTIC ACCURACY THROUGH COMPUTER VISION

Kaustubh Sonawane
Aditya Rai

Plano West Senior High School

Sona.om78@gmail.com; Adyrai05@gmail.com

Abstract: Non-small cell lung cancer (NSCLC) results in over 1.8 million deaths worldwide every year; however, most of these deaths are preventable via early diagnosis, which reduces the mortality rate by more than 50%. Currently, physicians use a CT (Computed Tomography) scan as a preliminary method of identifying cancerous tumors. Unfortunately, this visual process of identifying NSCLC scans becomes time-consuming and inaccurate, leading to high misdiagnosis rates. The goal of this project is to create a cloud-based web application that can take an inputted CT scan and identify regions of potentially cancerous tumors at an accuracy >90%. This is accomplished by first standardizing all inputted scans to a standard size and range of pixel values. A 3D CNN is then trained to classify an inputted scan as either "positive" or "negative" for cancer. Class Activation Mapping (CAM) is then used on scans classified as "positive" in order to identify the location(s) of cancerous tumors. This algorithm core is accessed through a cloud-based user interface on AWS allowing physicians to upload and organize their patient's NSCLC scans as well as receive dynamic results based on patient biometric and cancer history background. Using the LungStat platform, an oncologist can simplify the procedure for lung cancer diagnosis by cutting down the average tumor identification time from about 1 hour to a few minutes. Overall, this project establishes a critical tool needed for the accurate diagnosis and treatment of NSCLC, leading to a severe reduction in the death rates caused by lung cancer.

1. INTRODUCTION AND BACKGROUND

NSCLC is a type of cancer emerging from the inner tracts of the lungs, damaging the airways of the patient and creating long term impacts on oxygen flow systems. This type of lung cancer results in about nearly 3 million deaths globally each year (WHO, n.d.). While NSCLC is traditionally more prominent among smokers, the emergence of more pollutants in the air has caused non-smoker NSCLC rates to match that of the smoking population (Pope et al., 2002). Within the US more specifically, the rise of lung cancer within non-smokers has been startling with growing from 8% of total cases from 1990-1995 to 14.9% of all cases from 2011-2013 (Pelosof et al., 2017). Regardless, the case rate for the general population still continues to rise with more than 200,000 new cases expected in just 2021 within the US (American Cancer Society, 2021). As a direct response to this dramatic increase in overall cases, the US Preventive Services Task Force has increased its recommendation namely for smokers from the ages of 50 to 80 to take an annual CT scan to screen for lung cancer tumors, thus placing the new importance of screening in the spotlight of the medical field.

Unfortunately, a large part of the issue with NSCLC is a lack of effective early diagnosis. Patients diagnosed within the early stages of lung cancer (stages I and II) have a 5-year survival rate of more than 60%; however, this drops down to less than 10% in later stages of cancerous growth, indicating that early detection and diagnosis of NSCLC tumors is essential to maximizing a patient's chance of recovery (Willow, 2020). Due to a lack of specific diagnosis in the early stages, more than 75% of patients are not

aware of the extent of the cancerous growth until a point past potentially curative surgical resection (stages III and IV) (Birring, 2005). Additionally, due to a considerable time cost associated with checking each scan, physician fatigue is spread throughout the NSCLC diagnosis process. As a result of these issues, the diagnosis process is riddled with high false-negative, false-positive, and overdiagnosis rates, leading to only 85-92% of NSCLC cases being accurately diagnosed by a physician (Pinsky, 2014). Thus, new methods and tools of detection are desperately needed to improve the effectiveness and efficiency of NSCLC diagnosis.

Current lung cancer treatments are based on location-specific radiation therapy (for targeting metastases) and chemotherapy to address the overall distribution of cancerous tumors. However, the reason for high mortality with NSCLC lies in the inability to diagnose cancer early and efficiently. 91% of NSCLC diagnosis takes place in the 3rd and 4th stages of the disease. In addition, physicians and other hospital staff must often spend hours (if not days) analyzing the scans for a single patient, leading to physician fatigue. This fact, when combined with the sharp reduction in NSCLC survival rates for patients diagnosed in the later stages (~70% survival when detected in early stages vs. ~25% survival when detected in later stages), indicates that there is significant room for improvement in the NSCLC detection procedure. The detection methodology in analyzing CT scans for tumors has a high prevalence of misdiagnosis and missed tumors, with the current aggregate rate of misdiagnosis at almost 40% (Svoboda, 2020). This project aims to reduce the mortality rate of NSCLC by improving the NSCLC diagnosis procedure via the development of a deep-learning-based software to diagnose NSCLC tumors efficiently and effectively.

Thus, this project aims to help improve the effectiveness and efficiency of the NSCLC diagnosis process by using Machine Learning algorithms to aid physicians in identifying the locations of lung cancer tumors.

2. MACHINE-LEARNING CORE

The Machine Learning section of this project uses the 3D CNN (Convolutional Neural Network) architecture to create a classification model that can accurately identify if a given CT scan has a cancerous tumor present. The backend then utilizes Class Activation Maps in order to determine the regions of the scan that most “excited” the CNN and thus identify the areas of the scan that are likely to contain cancerous tumors.

The classification 3D CNN was trained on the LIDC-IDRI dataset, which includes nearly 1300 full CT scans from over 1000 patients (TCIA, n.d.). However, these scans vary in size and range of pixel values. In order to better standardize the dataset and make it easier for the model to analyze a given scan, a preprocessing script for CT scans was developed. This script takes a given scan and standardizes its size to a 128 by 128 by 128 cube. Depending on the CT scanner used to take the scan, different scans may have different ranges of pixel values; thus, in order to minimize the variance in the dataset and ensure a more accurate end model, scans are then converted from their original values to Hounsfield units of radioactivity, which will have the same range of values regardless of the initial pixel values.

Once the input data had been preprocessed and stored, it was iterated through in 100 epochs. Each epoch trained the model on a randomly selected 70% of the data and “validated” the model on 20% of the data. 10% of the data was withheld from the model entirely to create a completely new set of data upon which to test the model. The model architecture consisted of several Convolution, MaxPooling, and Batch Normalization layers followed by Global Average Pooling and Fully Connected Layers.

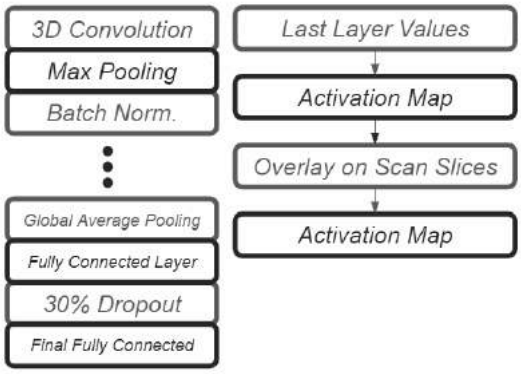

Figure 1: Layers of the Classification CNN

Once the classification network had been trained, a Grad-CAM-based CAM class was developed to detect the regions of highest activation once the model had run on a particular scan. A visualization class was then created to visualize the class maps, the inputted scans, and the overlays of the class maps on the input scans. Using the trained model, the region(s) of highest activation are marked as regions that may contain tumors and displayed through the LungStat Diagnostic Application.

3. DIAGNOSTIC APPLICATION

While previous research has developed a computer vision alternative to the current method of visual identification, the implementation into a clinical setting has been limited with the difficulty of transferring this software into an easy-to-use application. The user interface of the LungStat application aims to complete this task of moving the functionality of the machine-learning algorithm core into a physician-friendly web application.

The development of the LungStat user interface took place over several iterations involving creating an accessible frontend, proper data management, and an efficient structure that could support a large number of queries at a given time. More specifically, the parameters for the created hosting application around the ML algorithm were to remain scalable for individual users without administrative involvement, efficient with a runtime under 3 minutes, and be subjectively simple for physicians to integrate into a standard lung cancer diagnosis. Based on these parameters, the best solution was creating a cloud-based web application using AWS for hosting and data management based on these parameters.

The frontend of the application was created using the React.js framework that allowed quick customization of the UI while maintaining the ability to pull from multiple libraries for stylistic components. The web application structure involved an opening and informational page about LungStat as well as a user-specific section with the physician dashboard, patient entry, and settings pages, all shown in the figure below.

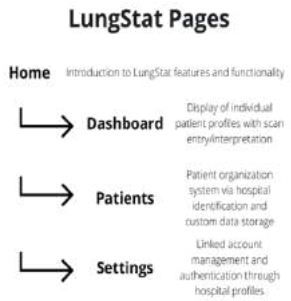

Figure 2: LungStat Application Pages with Purposes

On the backend, AWS Cognito tied with DynamoDB tables were used to set up individual physician accounts and store scans associated along with other patient data. Additionally, AWS S3 buckets were used to store the user-entered scans in ZIP format, thus reducing the direct impact on the backend processing limit. Finally, for the processing and location of the ML algorithm core, AWS Lambda Functions were put into use. Instead of a traditional EC2 instance that would keep the ML algorithm running and ready for processing, the Lambda function instead stores an image of how to boot up the algorithm. At the same time, the majority of the code is hosted within a separate S3 bucket. Overall, the entirety of the application proved effective as a demonstration of serverless architecture, which allows this algorithm to be easily scaled up without significant hits to processing time.

4. RESULTS AND DISCUSSION

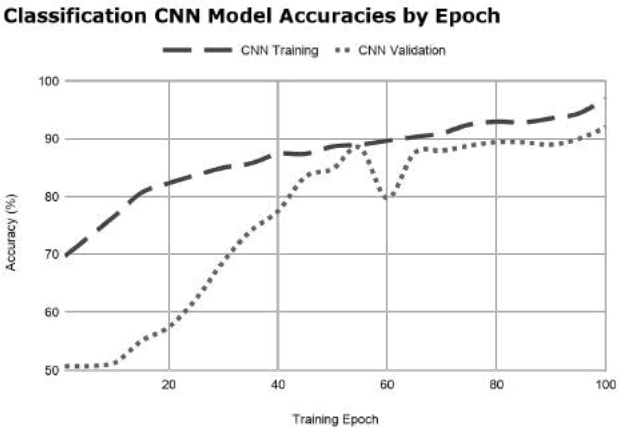

Figure 3: Classification CNN Model Accuracies by Epoch

The usage of this application within a clinical setting effectively reduces issues associated with the current NSCLC diagnosis. With the LungStat application, the patient's traditional process is changed so that when they enter a hospital and a CT scan is taken, the imaging result is uploaded onto the LungStat website, returning a detailed report on the presence and locations of cancerous tumors. This simple change compared to the time-intensive and inaccurate process of visually identifying each of the tumors in the body ensures the reduction of physician fatigue in hundreds of cases each year. As a future step, physicians can recommend a treatment plan based on their own examination of available imaging, thus ensuring that any clinical plans are mapped to the severity and location of all tumors in the body. The high accuracy of 94% and low runtime of 3 minutes throughout this process instead of the alternative of visual inspection at a low accuracy proves beneficial for patient health outcomes with early detection. Additionally, the unique approach of combining the most effective parts of both the current physician process with AI detection ensures the best diagnosis.

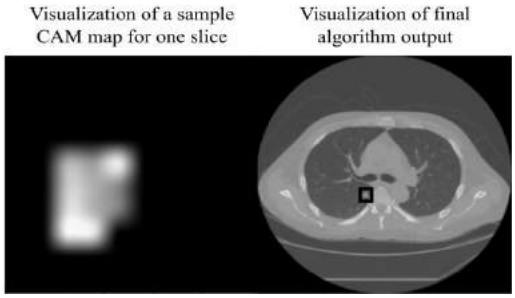

Figure 4: Visualizations of CAM and Final Software Outputs

In the future, the following steps are to expand the capabilities of the machine learning algorithm for tumor detection and optimize this software for physician implementation. While the current 3D CNN in conjunction with CAM can identify the location of the tumor, it falls short of pinpointing the exact boundaries of tumors; thus, for this feature, we plan on supplementing our current model with a 3D Mask RCNN tumor mapping algorithm that can locate the edges of the tumor, ensuring the utmost precision of any diagnostic interpretation. Expanding the current dataset would additionally prove beneficial for the current accuracy of the machine learning algorithm, leading to a more comprehensive report given to the physicians. On the application side, the goals are to expand the ease of integrating this software into a clinical setting through the iterative development of features and user experience. By optimizing the LungStat application to the needs of hospitals and physicians through in-hospital testing, this technology would be much more accessible to patients around the country.

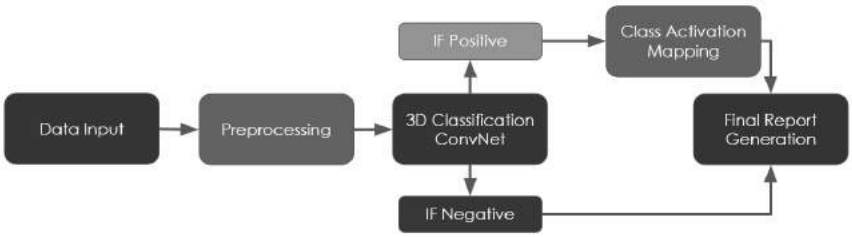

Figure 5: Dataflow for the LungStat Software

5. REFERENCES

[1] Bi, W., Hosny, A., Schabath, M., Giger, M., Birkbak, N., Mehrtash, A., . . . Aerts, H. (2019, February 05). Artificial intelligence in cancer imaging: Clinical challenges and applications. Retrieved December 30, 2020, from https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21552

[2] Birring, S. S. (2005). Symptoms and the early diagnosis of lung cancer. Thorax, 60(4), 268-269. doi:10.1136/thx.2004.032698

[3] Cancer. (n.d.). Retrieved December 10, 2020, from https://www.who.int/news-room/fact-sheets/detail/cancer

[4] Ciello, A. D., Franchi, P., Contegiacomo, A., Cicchetti, G., Bonomo, L., & Larici, A. R. (2017). Missed lung cancer: when, where, and why? Diagnostic and Interventional Radiology, 23(2),118–126. doi: 10.5152/dir.2016.16187

[5] Current Issues in Lung Cancer Screening. (2005, October 31). Retrieved December 20, 2020, from https://www.cancernetwork.com/view/current-issues-lung-cancer-screening

[6] Data From LIDC-IDRI. (n.d.). Retrieved December 16, 2020, from https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[7] Jamal-Hanjani, M., Zenklusen, J. C., McMurray, J. J. V., Celli, B. R., Wedzicha, J. A., Agustí, A., & Hogg, J. C. (2017, June 1). Tracking the Evolution of Non–Small-Cell Lung Cancer: NEJM. Retrieved from: https://www.nejm.org/doi/10.1056/NEJMoa1616288?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub%3Dwww.ncbi.nlm.nih.gov#article_citing_articles

[8] Pope, C. A., Burnett, R. T., Thun, M. J., Thurston, G. D., Ito, K., Krewski, D., & Calle, E. E. (2002). Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. JAMA, 287(9), 1132. https://doi.org/10.1001/jama.287.9.1132

[9] Lung Cancer Statistics: How Common is Lung Cancer. American Cancer Society. (2021, January 12). https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html.

[10] Mid, Rosebrock, A., Iftikhar, A., BELAFDIL, C., Luke, Abu-Abdurrahman, . . . R., R. (2020, April 18). Grad-CAM: Visualize class activation maps with Keras, TensorFlow, and Deep Learning. Retrieved December 18, 2020, from https://www.pyimagesearch.com/2020/03/09/grad-cam-visualize-class-activation-maps-with-keras-tensorflow-and-deep-learning/

[11] Pelosof, L., Ahn, C., Gao, A., Horn, L., Madrigales, A., Cox, J., … Schiller, J. (2017). Proportion of Never-Smoker NON–SMALL cell lung cancer patients at three diverse institutions. Journal of the National Cancer Institute, 109(7). https://doi.org/10.1093/jnci/djw295

[12] Pinsky, P. F. (2014). Assessing the benefits and harms of low-dose computed tomography screening for lung cancer. Lung Cancer Management, 3(6), 491-498. doi:10.2217/lmt.14.41

[13] Svoboda, E. (2020, November 18). Artificial intelligence is improving the detection of lung cancer. Retrieved December 15, 2020, from https://www.nature.com/articles/d41586-020-03157-9

[14] Willow, J. (2020, July 07). Lung Cancer: Types, Survival Rates, and More. Retrieved December 20, 2020, from https://www.healthline.com/health/lung-cancer-stages-survival-rates