Name(s)

Suraj Anand

Project Number

S0805

Project Title

Machine Learning Ensemble Model for Improved Personalized Lung Cancer Risk Assessment and Malignant Nodule Detection

Abstract

Objectives

Current screening guidelines omit a large number of high-risk candidates that do not fit the traditional screening criteria. Furthermore, malignant lung nodule detection on CT scan is difficult as nodules are often miniscule and often benign/indeterminate. This causes radiologist screening of nodules to be expensive, low-throughput, and often inaccurate. This study develops an algorithm that utilizes machine learning and radiomics to build a complete lung cancer diagnostic pipeline that addresses these issues.

Methods

A large patient history dataset was obtained from Kaiser Electronic Medical Records and a separate large lung-CT scan dataset was compiled and hand-modified from various online sources. In order to assess a patient's risk of lung cancer, a 50-tree Gradient Boosted Machine (GBM) was constructed that employs personalized patient history variables including age, prescriptions, ethnicity, body mass index, blood pressure, and diagnoses to better assess true risk of patients. Once a CT scan is conducted to identify malignant lung nodules, an ensemble of 3D Convolutional Neural Networks (CNNs) of discriminator VGG-like and U-net architectures, trained with multitudinous augmentations and gradient clipping on a hand-engineered dataset, determines nodule morphology (calcification, spiculation, size), position, and malignancy. From these features, a linear classifier predicts lung cancer development in one year.

Results

The GBM significantly surpasses current guideline assessments, capturing omitted patient groups at high risk for lung cancer (sensitivity increased from 23% to 88%). Moreover, the CNN Ensemble obtained statistically comparable predictions to radiologist readings of scans. Further, the Ensemble substantially reduced the false positive rate of Computer Aided Diagnosis models (from on average 15.28 to on average 1.68 false positives per scan).

Conclusions

This model could serve as primary screen for lung cancer nodules to decrease radiologist involvement in screening. The combined automated lung cancer diagnostic system increases early-detection rates and reduces false positive rates of Computer Aided Diagnosis systems, thereby greatly improving the timeliness, accuracy, and affordability of lung cancer detection.

Summary Statement

I developed a machine learning algorithm that acts as a viable fully-automated lung cancer diagnostic pipeline.

Help Received

Dr. Drew Clausen aided in obtaining a patient history dataset and answered various questions regarding exploratory data analysis and model architecture. I accumulated a separate CT dataset from online sources. I further explored data relationships, modified datasets, and developed the algorithms on my own.