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

Suraj Anand

2018-2019

Science Research

Table of Contents

ANAND, Suraj
Computer Science
2019

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

Lung cancer is the deadliest cancer, causing 1.4 million deaths annually due to late diagnosis and limited access to screening specialists. Currently, high-risk patients (guideline of >50 years and >30 smoking-pack-years) are advised to undergo a CT scan, omitting other high-risk candidates that do not fit the screening criteria. Furthermore, malignant lung nodule detection is similar to finding a needle in a haystack; nodules are often less than 3x3x3mm in a 400x400x400mm CT scan 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. For preliminary lung cancer risk assessment, a 50-tree Gradient Boosted Machine (GBM) employs personalized statistics including age, prescriptions, ethnicity, body mass index, blood pressure, and diagnoses to better assess true risk of patients. Once a CT scan is conducted, 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, determine nodule morphology (luminosity, spiculation, size), position, and malignancy; from these features, a linear classifier predicts lung cancer development in one year.

The GBM significantly surpasses current high-risk guideline assessments, capturing omitted patient groups (sensitivity increased from 0.23 to 0.88). Moreover, the CNN Ensemble obtained statistically comparable predictions to radiologist readings of scans. The combined system increases early-detection rates and decreases radiologist involvement in screening, thereby greatly improving the timeliness, accuracy, and affordability of lung cancer detection.

Introduction

Lung Cancer:

Human cells require an adequate supply of oxygen to function properly. The location of this oxygen inflow as well as carbon dioxide release is the lungs. A mucus layer on the cilia acts as a defensive mechanism and protects the lungs from bacteria, viruses, and other toxins from the outside environment. However, the lungs still experience a plethora of diseases including chronic obstructive pulmonary disease, asthma, bronchitis, pneumonia, sarcoidosis, tuberculosis, pulmonary fibrosis, and lung cancer due to constant interaction with foreign air. These diseases are often difficult to differentiate and require specialized radiologists to diagnose. This study focuses on lung cancer, which develops as a result of genetic damage to oncogenes and tumor suppressing genes, engendering rapid and uncontrolled cell proliferation.

Lung cancer is one of the deadliest forms of cancer, resulting in over 1.4 million deaths annually (Yu, et al. 2016). Smoking, in addition to other lung carcinogens such as asbestos, radon, and arsenic, is responsible for over 85% of lung cancer cases (Adetiba, Olugbara 2015). There are two types of lung cancer: small cell lung cancer (SCLC), making up 13% of cases and non-small cell lung cancer (NSCLC), making up the remaining 87% of cases. NSCLC is classified as more dangerous due to its slow moving nature; NSCLC prognosis statistics display that Stage IA diagnosis five-year survival is 67% , Stage IIA diagnosis five-year survival is 55%, Stage IIIA diagnosis five-year survival is 23%,

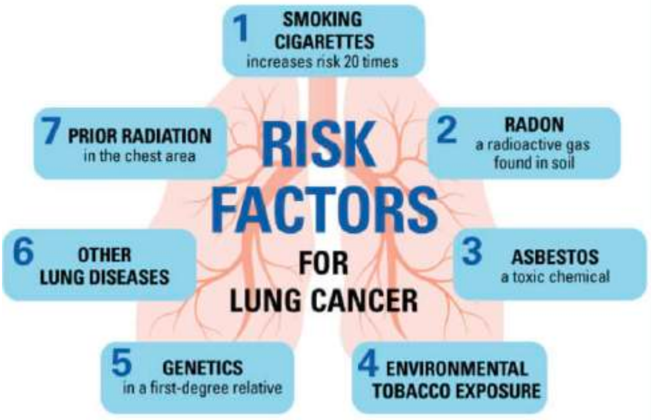

1 SMOKING CIGARETTES increases risk 20 times
2 RADON a radioactive gas found in soil
3 ASBESTOS a toxic chemical
4 ENVIRONMENTAL TOBACCO EXPOSURE
5 GENETICS in a first-degree relative
6 OTHER LUNG DISEASES
7 PRIOR RADIATION in the chest area
RISK FACTORS FOR LUNG CANCER

Figure 1: Known Lung Cancer Risk Factors (National Jewish Health 2012)

and Stage IVA diagnosis five-year survival is 1% (Adetiba, Olugbara 2015). Lung cancer manifests as pulmonary nodules, which are defined as “round opacities, at least moderately well marginated and no greater than 3 cm in maximum diameter”; larger masses are called pulmonary masses and have a higher probability of representing cancer (Girvin, Ko 2008).

Current Detection Techniques:

Presently at most medical centers, patients above the age of 50 with a smoking history of at least a pack a day are advised to conduct a chest tomography (CT) scan if feeling symptoms such as chest pain. Recent advances in CT scanning render it significantly better at small nodule detection than other tests; therefore, the US Preventive Task Force (USPTF) recommends yearly screening and tomography computation for high risk patients (Girvin, Ko 2008). Other detection imaging such as magnetic resonance imaging (MRI), and Positron emission tomography (PET) is used less often.

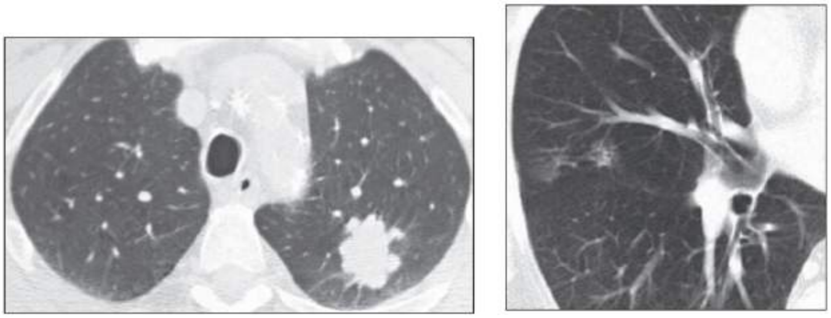


Figure 2: Large Malignant Mass, left and Smaller Pulmonary Nodule, right

From scans, radiologists attempt to find malignant nodules. As displayed in Figure 2, malignant lung cancer masses are easily visible in CT scans. However, small lung cancer nodules are much more difficult to identify, and are often confused with pneumonia or pleural effusion. Moreover, the “overall identification of lung cancer images using this radiological equipment is very low at the early stage of the disease [because] radiologists… do not sometimes differentiate accurately between malignant and benign and forms of lesions” within the lungs (Adetiba, Olugbara 2015). The tedious and subjective analysis of elderly and large-smoking-history patients prevents high patient throughput and more importantly, omits a large proportion of possible candidates that may not fit the testing age or smoking history criteria. In addition, low-dosage CT scans are expensive, ranging from $270 to $5000 and increase the risk for cancer. It would be beneficial to develop a Computer-Assisted Diagnosis (CAD) model to consider all possible patients, ease the load of the radiologist, and possibly identify features not perceived by radiologists.

Computer-Assisted Diagnosis Intelligent Techniques:

Presently, two primary CAD approaches have been developed for general malignant nodule detection: a radiomics approach employing radiological quantitative features (QIF) and a deep learning approach using techniques such as a 2D convolutional neural network (CNN) (Causey 2018). The deep learning approach shows great potential; however, its feature computation remains enigmatic to physicians due to a "blackbox effect," and therefore people are skeptical for generalized use. While the radiomics approach requires properly pre-segregated nodules, CNNs can perform analysis without significant preprocessing. However, CNNs necessitate a much larger dataset than radiomics analysis; still, once trained, direct CNN prediction is much more efficient than the feature extraction prediction pipeline of radiomics approaches.

Convolutional Neural Networks possess input, hidden, and output layers; the hidden layers often are convolutional, Rectified Linear (ReLU), and pooling layers. These layers apply a nonlinear filter to the image, compressing the visual field to multiple overlapping receptive fields that possess a smaller representation of the spatial imagery. Each convolutional unit applies a convolution operation (mathematically learned operation) to mimic the visual ability of a neuron for a particular region. The region then passes through a nonlinear activation function to model the data, a pooling function (compressed), and then is passed to another convolutional unit. After this process repeats, the low representation of the data is mapped to some non-spatial numeric feature by a dense unit. On the contrary, radiomics requires individual nodule discrimination (often manual), nodule rendering, feature extraction, and then analysis. This model lacks the holistic pattern recognition that accompanies CNNs, as it simply considers known features such as diameter and number of nodules.

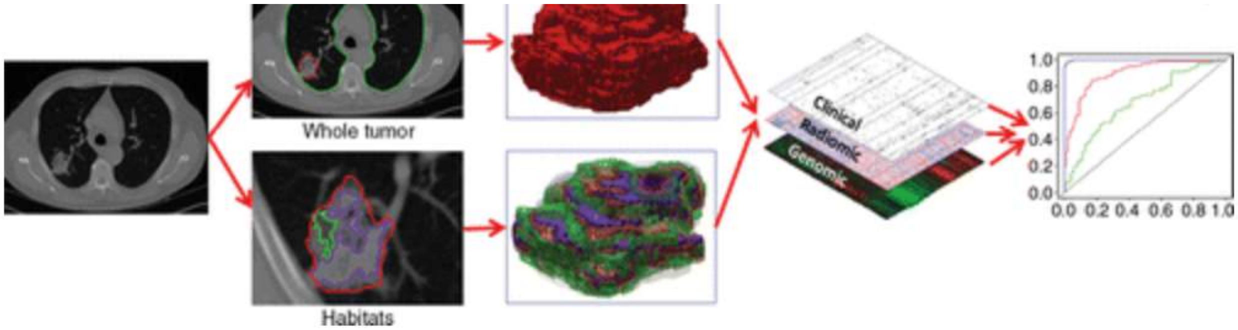


Figure 3: Above is the radiomics pipeline: a region of interest is identified, the region is rendered in 3D, features are extracted from the feature, and some statistical model is applied to the feature. (Giles 2008)

While there has been significant progress in the development of both procedures, the results remain poor compared to experienced radiologists. Over the last few years, models have been assembled that attempt to combine the deep learning and radiomics fields. One such model, NoduleX, has reached an astonishing sensitivity level of 88.5% only after identifying possible benign or malignant nodules; this result is promising, but not to the level of very experienced radiologists, who exhibited sensitivity levels of between 94.4-96.4% in the NLST and NELSON trials with an average false positive rate of 0.6-2.1 (Rubin 2016). However, radiologist sensitivity levels vary greatly, varying from 30-97% depending on the nature of the input. Currently, CAD systems display a sensitivity of around 60-80% with a very high average of 28 false positives per read (Rubin 2016). Present models require excessive preprocessing, demand distinct 2D scan and nodule splitting, and predict possible malignant nodules, not expected cancer development risk. The proposed system effectively overcomes current CAD analysis limitations by making use of individual patient data in addition to the CT scan. This novel systematic approach will employ a UNET architecture applied to a 3D scan in addition to a fully-connected convolutional network and discriminator CNN. Finally, a gradient boosted machine (algorithm that intelligently weights malignant nodules and patient features) is utilized to predict pulmonary cancer development risk. This model is superior to traditional models because it considers all possible patients (not just patients in the supposed risk population), assesses individual patient features (age, smoking history, etc) in addition to the CT scan, uses an image-size agnostic architecture, splits segmentation and analysis into distinct and automated stages, and computes a final, single-digit numeric predictor of malignant pulmonary cancer development within the next year.

The objective of this research was to address inefficiencies with the current pipeline for malignant lung cancer development risk—mainly its low throughput, sometimes inaccurate assessment, and omission of a large proportion of possible candidates. Malignant lung nodule prediction is similar to finding a needle in a haystack as in a 400x400x400mm CT scan, nodules are often less than 3x3x3mm in volume and often benign or indeterminate; this causes the brutally tedious task of radiologist analysis is subjective, sluggish, often incomplete, and expensive. Furthermore, the current high risk candidate selection imprecisely defines patients, omitting a large proportion of possible candidates that may not fit the testing age or smoking history criteria**.** Finally, this research confronts impediments to reliance on CAD models (mainly high false positives and negatives) by hand modifying an accumulation of datasets and integrating several aspects of forefront CT-analysis models with creative modifications. A preliminary intelligent survey will be constructed to assess initial malignant lung cancer risk. The survey will indicate predicted risk and whether or not the patient should conduct a CT scan. In order to better predict possible high-risk lung cancer candidates, this project employs age, prescriptions, ethnicity, BMI, blood pressure, and diagnoses to assess the likelihood of malignant cancer development. Moreover, once a low-dosage CT scan is conducted, multiple 3D convolutional neural network models with accumulated and hand-modified datasets would be utilized to observe malignant nodules. The use of this model would enable a faster and more accessible assessment of lung cancer.

Methods

This research was split into two distinct phases: intelligent assessment survey construction, and CT scan malignant nodule detection. Each phase of the project required problem analysis, exploratory data analysis, data preprocessing, model construction, model evaluation and revision, and result prediction. First, a preliminary literature review was conducted to accumulate information about lung cancer, current detection methods, and possible model structures. This stage of the study allowed a full immersion into the present field of research, which better enables feature engineering and analysis when constructing a model.

Current research focuses on malignant pulmonary nodule CT scan detection. While showing promising results, false positive nodule counts are excessively high, rendering this software secondary to radiologist analysis. In addition, these studies limit the possible patient prediction to people that receive low-dosage CT scan analysis, an expensive, tedious, and often inaccessible test which exposes the patients to radiation risk. The preliminary survey more intelligently assesses initial patient risk.

In order to grasp a better understanding of the problem and acquire a baseline for improvements (and whether or not there was a problem present), the Kaiser Pulmonology Unit was reached out to for possible problem analysis. After several interactions, the Kaiser Permanente Innovation Center compiled a strictly de-identified portion of the Kaiser Permanente Electronic Medical Record (EMR) for in-house supervised use. This dataset possessed 23,000 patients (11,700 of which had contracted lung cancer), and variables researched to be associated. Metrics for evaluations including precision, recall, accuracy, and an f1 score were calculated based on the current screening candidate technique (patients >50 and with 30+ pack years). These were the standard of comparison for this study.

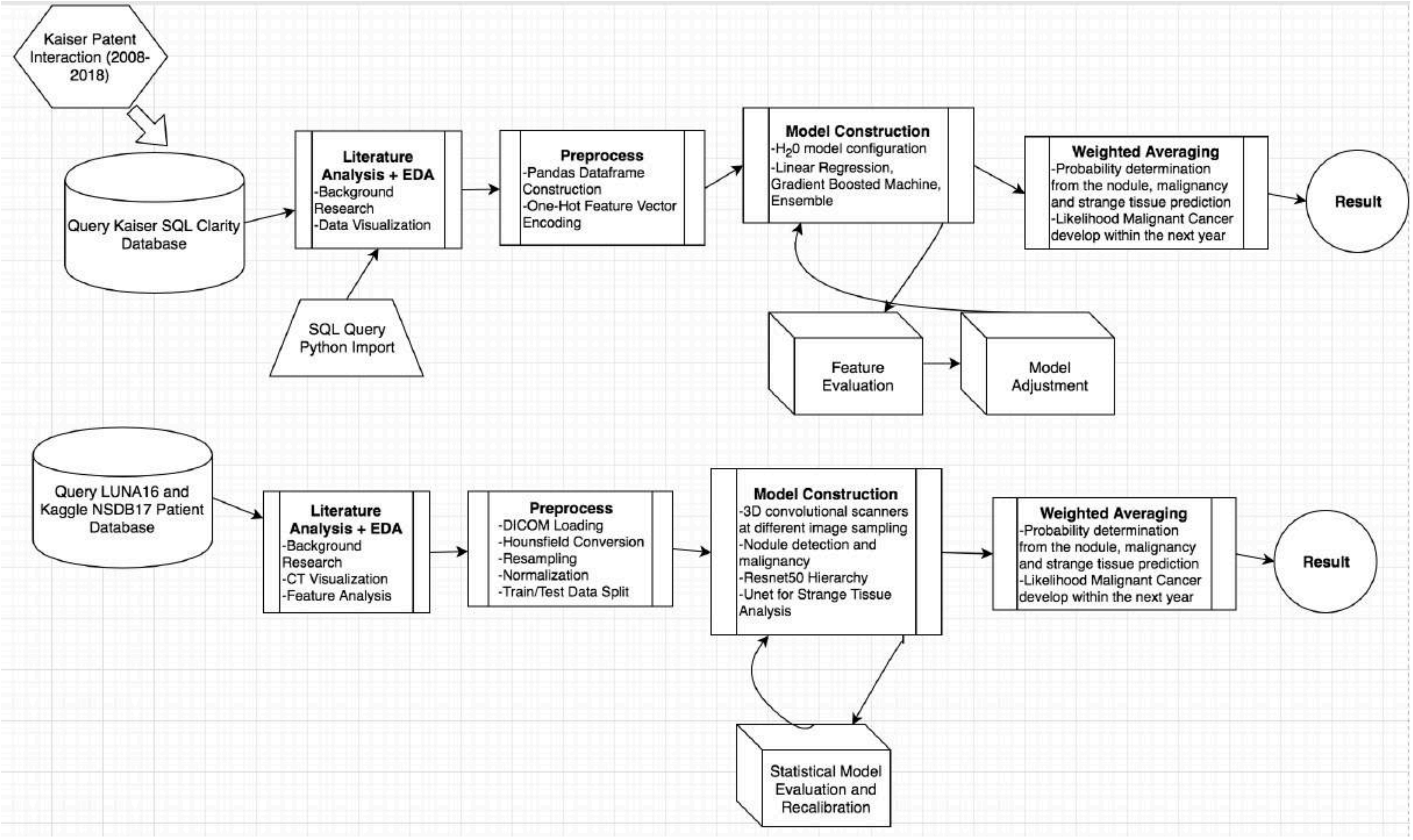

Kaiser Patent Interaction (2008-2018)
Query Kaiser SQL Clarity Database
Literature Analysis + EDA
-Background Research
-Data Visualization
SQL Query Python Import
Preprocess
-Pandas Dataframe Construction
-One-Hot Feature Vector Encoding
Model Construction
-H2O model configuration
-Linear Regression, Gradient Boosted Machine, Ensemble
Feature Evaluation
Model Adjustment
Weighted Averaging
-Probability determination from the nodule, malignancy and strange tissue prediction
-Likelihood Malignant Cancer develop within the next year
Result
Query LUNA16 and Kaggle NSDB17 Patient Database
Literature Analysis + EDA
-Background Research
-CT Visualization
-Feature Analysis
Preprocess
-DICOM Loading
-Hounsfield Conversion
-Resampling
-Normalization
-Train/Test Data Split
Model Construction
-3D convolutional scanners at different image sampling
-Nodule detection and malignancy
-Resnet50 Hierarchy
-Unet for Strange Tissue Analysis
Statistical Model Evaluation and Recalibration
Weighted Averaging
-Probability determination from the nodule, malignancy and strange tissue prediction
-Likelihood Malignant Cancer develop within the next year
Result

Figure 4: Pipeline of the Two-Part Study split into intelligent survey construction and CT scan malignant nodule detection and assessment

Stage 1:

The primary candidate assessment machine was created by constructing an intelligent classifier based on survey data. The patient diagnosis history during the 2-years before lung cancer diagnosis was extracted from the EMR's SQL Patient Database. From previous literature, a list of variables that may correlate with increased lung cancer risk was procured. The queried data from the EMR included age, smoking history, ethnicity, Body Mass Index (BMI), blood pressure, number and medical category of prescriptions, diagnoses, and Hierarchical Condition Categories (HCCs). Age and smoking history are important, established identifiers of lung cancer risk. In addition, ethnicity may help the model decipher any genetic risk of the cancer. BMI and blood pressure display the fitness and any other risk habits of the individual. The features of prescriptions and diagnoses allow the model to analyze the affected medical health of the individual. Finally, the HCCs serve as dimensionality reduction of other diagnoses information for the model, allowing a better indication of the patient's physical health and influenced tissues. These features were queried from the SQL database, and arranged into a tab-separated value (tsv) file. In addition, control random patient data is compiled to diversify the patient dataset. Next, the positive patient information and control patient information were read into Python 3.6 Pandas DataFrames.

To acquire a better understanding of the features and their relationship, exploratory data analysis was conducted. Histograms are generated of each of the numeric features, exhibiting correlations, frequency distributions, and the importance of certain features. Violinplots, scatterplots, and a correlation matrix are generated using the Python Library Seaborn, determining associations between different features. These plots are displayed in Figures 5 and 6.

From the distribution plots in Figure 5, age and pack-year history were determined to be important predictors of risk. BMI and Blood pressure, while not as ostensibly determinant, were included to give the model a more holistic view of the patient, allowing it to make multivariate connections. Age and smoking history are highly correlated (p-value = 5.4e-224) and positive malignant patients almost exclusively were above 40 years old. However, these statistics show serious flaws with the current predictive risk factor system—more than one third of the cancerous patients had less than 30 pack-years and over five percent of cancerous patients were less than 50 years old.

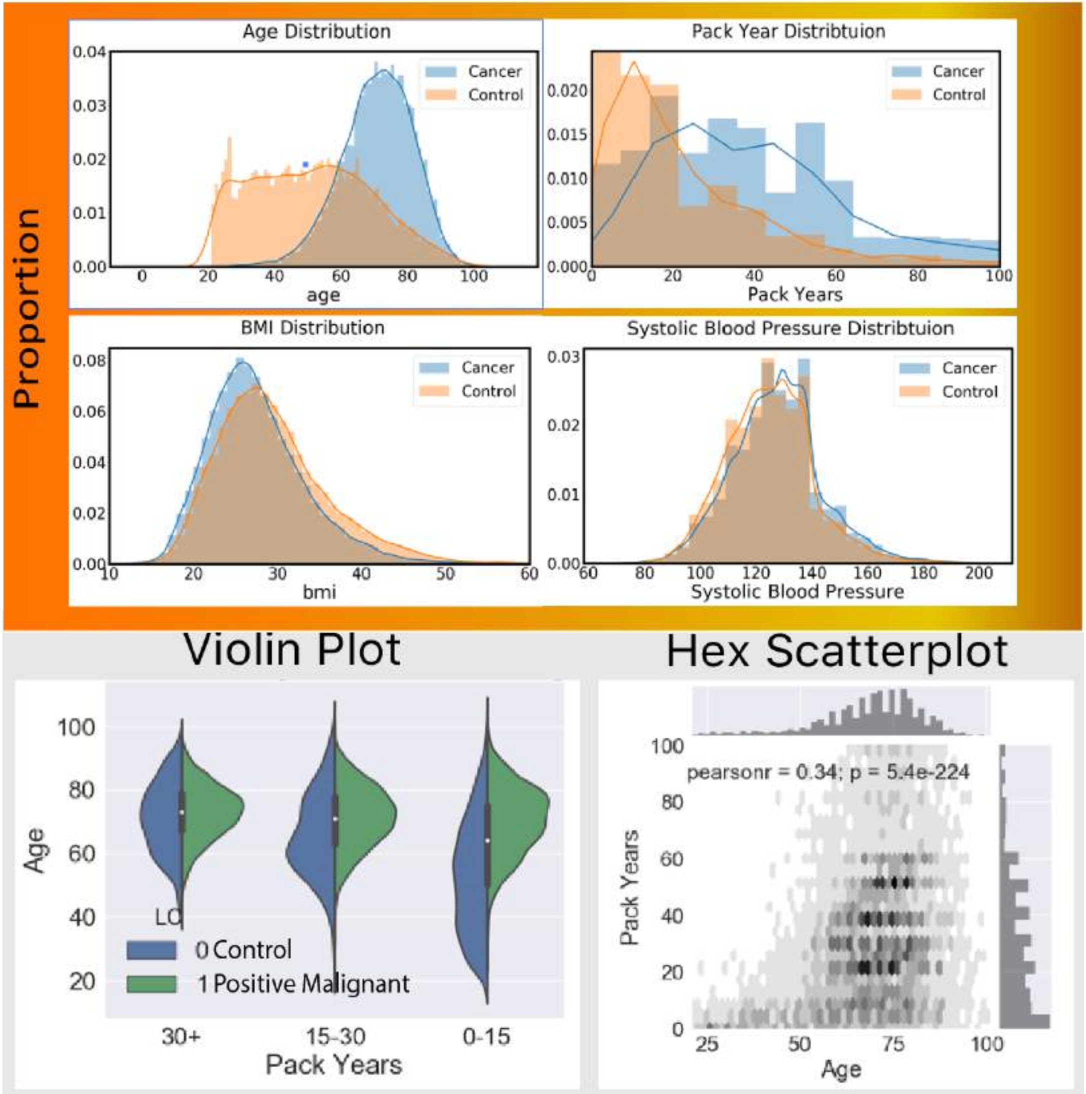


Figure 5: Distribution Plots of Age, Pack-Years, BMI, and Systolic Blood Pressure Features above. Violin plot and Hex Scatterplot between Age and Pack-Years Below (created with seaborn in python)

Next, histograms of the diagnoses, and prescription features were created as seen in Figure 6. From these plots, an array of features is generated—binary one-hot vector features were coded for all ethnicities as well as the top twenty diagnoses (top 10 shown below) and numerical one-hot vector features were coded for the number received of the top twenty prescriptions (top 10 shown below) in the last two years. The plots show that diagnoses such as CKD Stage 3 and Medications in the Cardiovascular and EENT Preps categories are likely indicators of increased lung cancer risk.

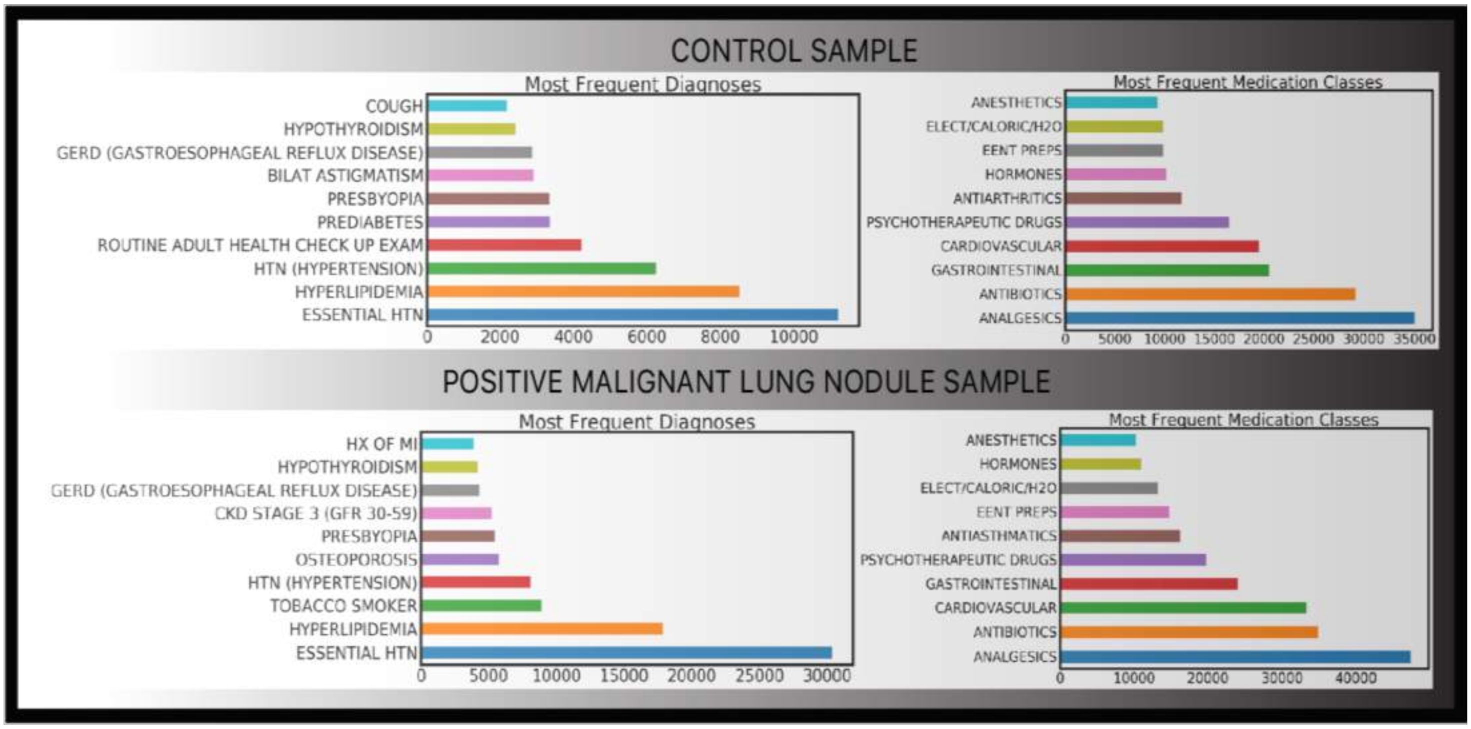
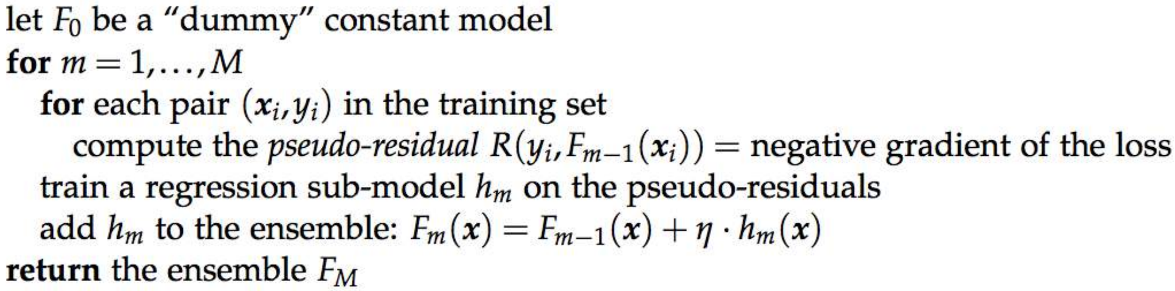


Figure 6: Histograms of top twenty Control and Positive Sample Diagnoses (Left) and Prescriptions (Right)—these are encoded as one-hot vectors to be used in malignancy prediction

After feature engineering, the data is split 80-20% into a training set and a validation set. The validation set acts as an unbiased indicator of the model's true performance of unseen data. A gradient boosted model and logistic regression model were constructed using H_2O.

Boosting performs significantly better than the logistic regression model during initial stages; consequently, logistic regression is removed from the model. This coherent response was expected due to the nature of the training set. While logistic regression employs one strong learner to determine a boundary line between the multi-dimensional space, a gradient boosted machine utilizes an ensemble of weak learners to formulate a strong hypothesis. This approach is more flexible. An initial constant tree is constructed, and fit on a subset of the dataset; the gradients of the loss function with respect to every variable of each data patient are computed. Then a subsequent tree is added to reduce the loss of the previous cumulative trees (taking a step towards the gradient). This process is repeated until proper model approximation occurs. To prevent overfitting, the number of trees was limited to 50 and partial dependency plots analyzed and L1 gradient descent regularization employed. Final output scores are converted to probabilities by a sigmoid function.



```
let $F_0$ be a “dummy” constant model
for $m = 1, \ldots, M$
  for each pair $(\boldsymbol{x}_i, y_i)$ in the training set
    compute the pseudo-residual $R(y_i, F_{m-1}(\boldsymbol{x}_i))$ = negative gradient of the loss
  train a regression sub-model $h_m$ on the pseudo-residuals
  add $h_m$ to the ensemble: $F_m(\boldsymbol{x}) = F_{m-1}(\boldsymbol{x}) + \eta \cdot h_m(\boldsymbol{x})$
return the ensemble $F_M$
```

Figure 7: Pseudocode of the general gradient boosting algorithm (Johansson 2016).

The gradient boosted decision trees are trained to optimize first the cost function of accuracy ($\frac{number\ correct}{total\ number}$) on the whole dataset and then the f1 score on combinations of a positive patient down sampled dataset ($2 * \frac{Precision*Recall}{Precision+Recall}$); this produced a generalized, accurate model for predicting lung cancer candidacy. The algorithm’s hyperparameters are modified to improve results. The results are manually sorted through to find any error patterns; the boosted tree model is altered to account for these errors.

Once the H_2O model was evaluated, partial dependency plots were produced of each of the variables. An extremely high weight (possible overfitting) had been set to age and the control vs. positive datasets possessed a vast imbalance (control had many more people under 40). Thus, the control dataset was revised to possess only patients above the age of the 40. The model was retrained and reevaluated.

Finally, the model was employed to predict the candidacy of the validation data to determine the true accuracy.

Stage 2:

During the second stage, low dosage CT scans were compiled from various data sources including the Lung Image Database Consortium (LIDC) IDRI image collection, LUNA16, and National Data Science Bowl 2017 datasets in the DICOM format. Building off past research from the University of Pennsylvania, the DICOM files were preprocessed in a multi-step process: first, the pixel values were translated to Hounsfield Units (HU), allowing densities of bones, tissues, water, and air to be calibrated; next, the DICOM files were resampled such that each pixel represents 1mm x 1mm x 1mm in the CT scan to remove variance in the scanner resolution; finally, the scans were bounded by the HU values of -1000 to 400 to remove unnecessary data (leave only the tissues and some parts of bone), as well as normalized and zero-centered to reduce the complexity of values. These steps are fully automated.

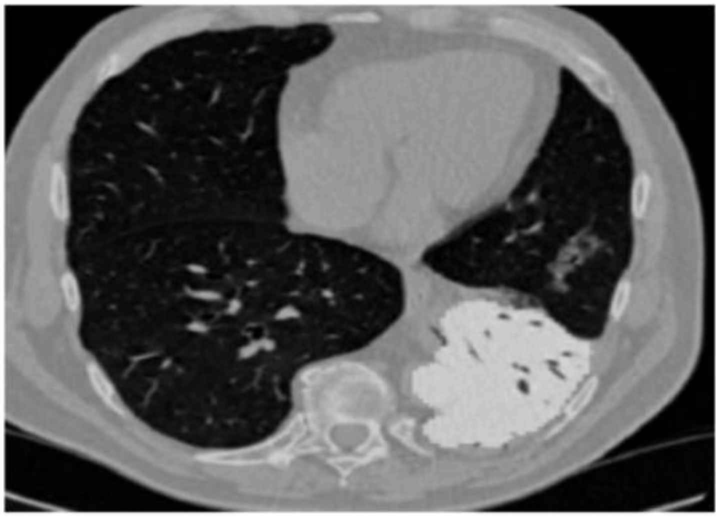


Figure 8: Anomalous Tissue Example located in Luna16 dataset

Next, exploratory data analysis was conducted to visualize the frequency distribution, image of true positive nodules, image of false positive nodules, and anomalous orientation/tissue photos. This found a high frequency of malignant pulmonary masses that were unmarked—due to their straightforward appearance, these were hand labeled. This significantly aided the CNN in training. In addition, some of the scans possessed anomalous tissue (not nodules); these samples had a higher chance of incurring lung cancer. A separate network would be trained to detect anomalous tissue.

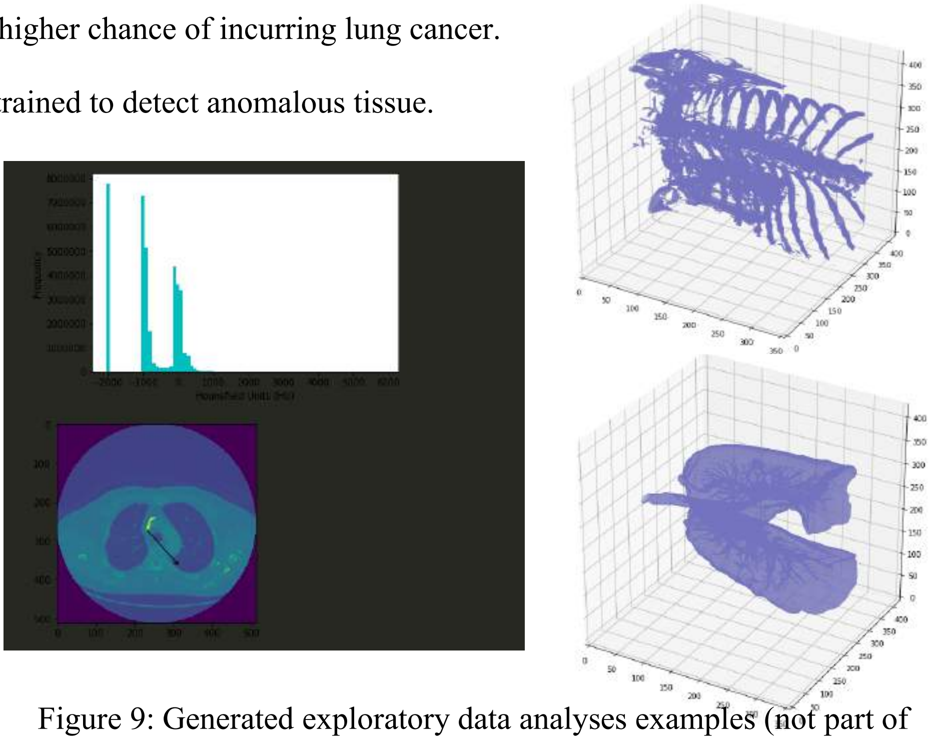


Figure 9: Generated exploratory data analyses examples (not part of algorithm). 3D visualizations using HU values—Kaggle notebook approach.

An ensemble of convolutional networks was trained to assess patient risk of lung cancer development within the next year from detected malignant nodules. Due to a low positive sampling rate (low proportion of patients with malignant nodules), various augmentation techniques (translations, reflections, and blurs) of scans were generated. These techniques provide the model with novel orientations of a slightly modified

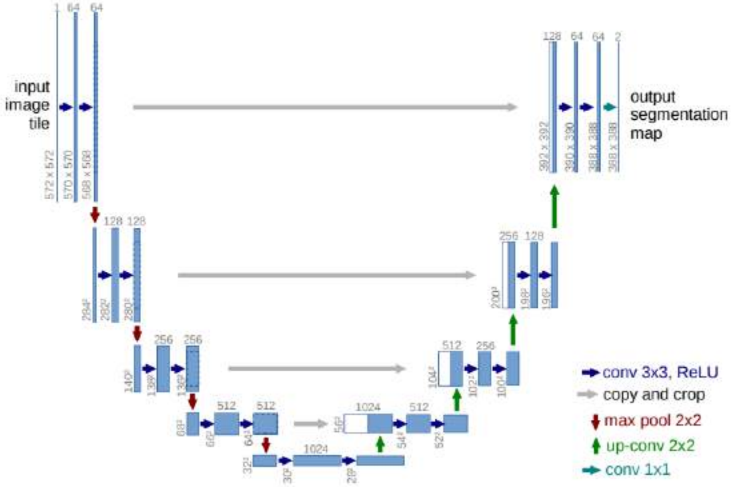
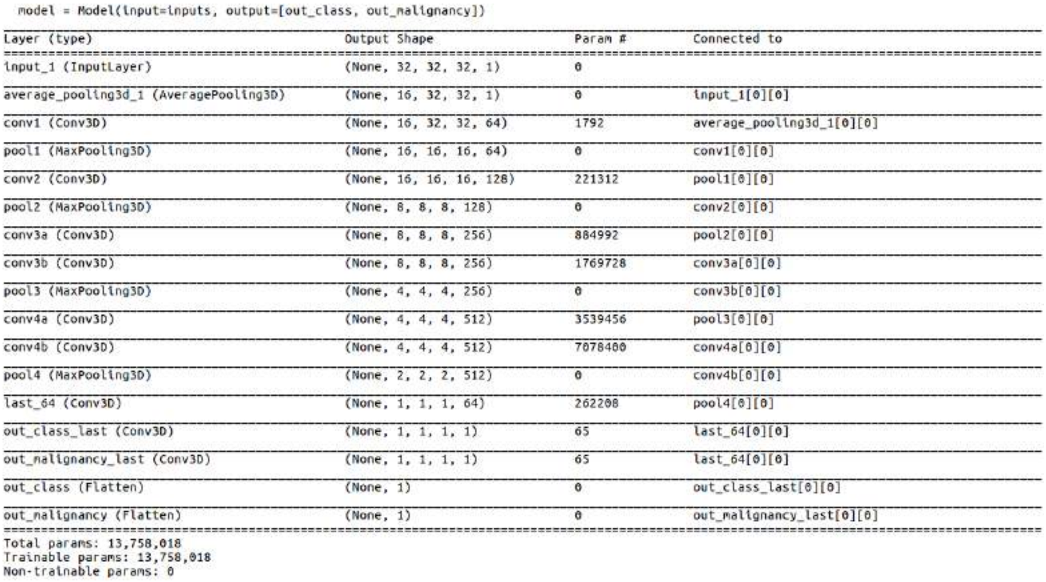


Figure 10: Hierarchy of the U-net (Ronneberger 2015).

image, reducing overfitting and effectively increasing the dataset size. The first portion of the model was a 3D Fully Convolutional Neural Network (FCNN) of a skip architecture. This CNN was tasked with nodule candidate region selection; rather than having a radiologist feed into the neural network possible malignant lung nodules, this hyper-efficient network broke down the CT scan into possible malignant nodule regions. This preliminary step is advantageous due to the nature of the FCNN network: it operates completely within the spatial realm without a need for the expensive fully-connected translational conclusion layers, resulting in the elimination of pointless analysis of all regions by the ensuing discriminators. This FCNN would supply the four following discriminator CNNs with candidate regions. The first of the constructed discriminators was a



```
model = Model(input=inputs, output=[out_class, out_malignancy])
____________________________________________________________________________________________________
Layer (type)                     Output Shape          Param #     Connected to
====================================================================================================
input_1 (InputLayer)             (None, 32, 32, 32, 1) 0
____________________________________________________________________________________________________
average_pooling3d_1 (AveragePooling3D) (None, 16, 32, 32, 1) 0     input_1[0][0]
____________________________________________________________________________________________________
conv1 (Conv3D)                   (None, 16, 32, 32, 64) 1792       average_pooling3d_1[0][0]
____________________________________________________________________________________________________
pool1 (MaxPooling3D)             (None, 16, 16, 16, 64) 0          conv1[0][0]
____________________________________________________________________________________________________
conv2 (Conv3D)                   (None, 16, 16, 16, 128) 221312    pool1[0][0]
____________________________________________________________________________________________________
pool2 (MaxPooling3D)             (None, 8, 8, 8, 128)  0           conv2[0][0]
____________________________________________________________________________________________________
conv3a (Conv3D)                  (None, 8, 8, 8, 256)  884992      pool2[0][0]
____________________________________________________________________________________________________
conv3b (Conv3D)                  (None, 8, 8, 8, 256)  1769728     conv3a[0][0]
____________________________________________________________________________________________________
pool3 (MaxPooling3D)             (None, 4, 4, 4, 256)  0           conv3b[0][0]
____________________________________________________________________________________________________
conv4a (Conv3D)                  (None, 4, 4, 4, 512)  3539456     pool3[0][0]
____________________________________________________________________________________________________
conv4b (Conv3D)                  (None, 4, 4, 4, 512)  7078400     conv4a[0][0]
____________________________________________________________________________________________________
pool4 (MaxPooling3D)             (None, 2, 2, 2, 512)  0           conv4b[0][0]
____________________________________________________________________________________________________
last_64 (Conv3D)                 (None, 1, 1, 1, 64)   262208      pool4[0][0]
____________________________________________________________________________________________________
out_class_last (Conv3D)          (None, 1, 1, 1, 1)    65          last_64[0][0]
____________________________________________________________________________________________________
out_malignancy_last (Conv3D)     (None, 1, 1, 1, 1)    65          last_64[0][0]
____________________________________________________________________________________________________
out_class (Flatten)              (None, 1)             0           out_class_last[0][0]
____________________________________________________________________________________________________
out_malignancy (Flatten)         (None, 1)             0           out_malignancy_last[0][0]
====================================================================================================
Total params: 13,758,018
Trainable params: 13,758,018
Non-trainable params: 0
```

Figure 11: 3D VGG-like C3D architecture used in ensemble model. The receptive field is 32mm x 32mm x 32mm and Z-axis is down sampled first to lighten the load on the net (does not affect performance because Z-axis much courser than X and Y axes on most scans). In addition, three fully-connected layers bottleneck features, and predict nodules and malignancy.

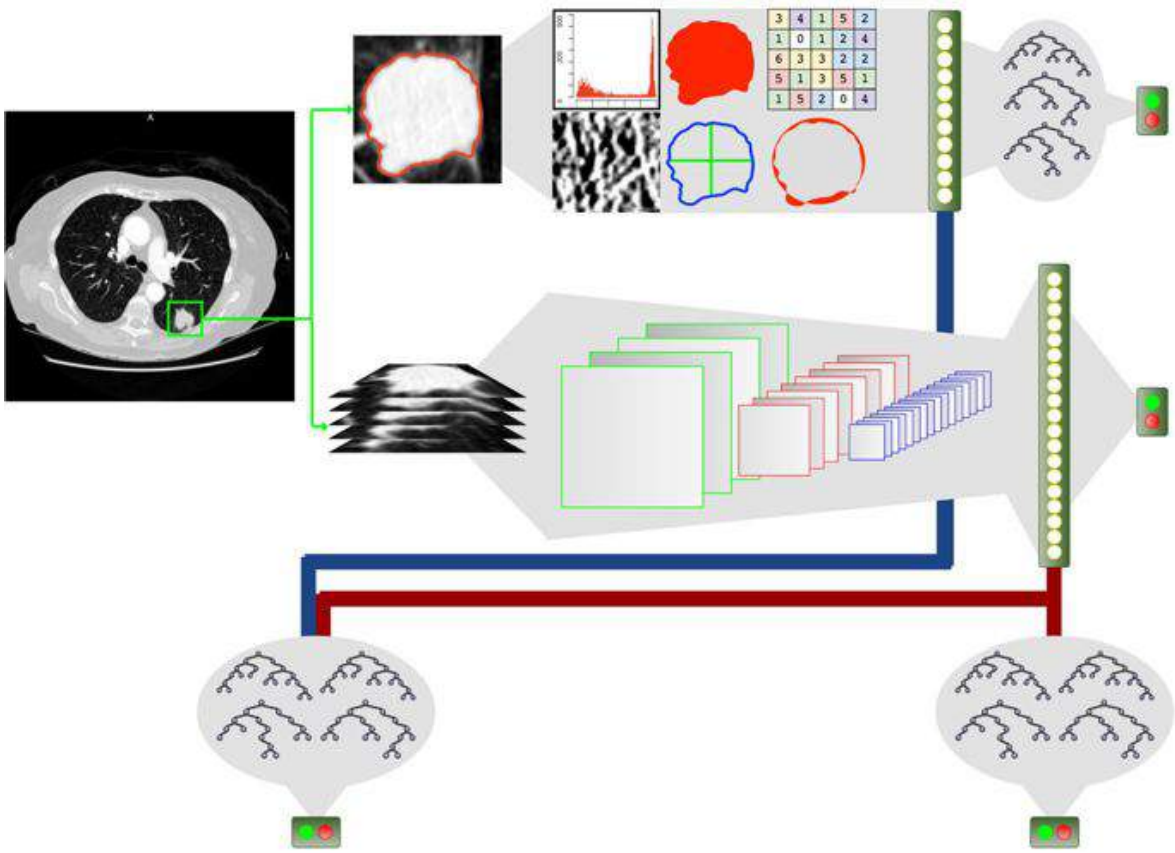


Figure 12: A high-level pipeline for prediction of NoduleX (Causey et. al). The proposed model additionally uses UNET deep CNNs, an anomalous tissue classifier, and a linear classifier to predict a final cancer prediction score.

3D U-net (hierarchy shown in Figure 12), implemented to inexpensively and accurately segment candidate regions for malignant pulmonary nodules. Appended to the tail of this network were fully-connected layer to predict the malignancy and size of each nodule. The U-net type of CNN is specialized for Biomedical Image Segmentation, consisting of "a contracting path to capture context and a symmetric expanding path that enables precise localization"; this architecture has outperformed the traditional sliding-window network on several 2D light-microscopy cell tracking competitions (Ronneberger 2015). However, the U-net acts as a course detector rather than a "fine-grained probability map" which may limit its functionality (Hammock 2017). For this reason, two 3D discriminator-scanning convolutional neural networks were engineered, both variations of the VGG-like C3D architecture. Again, fully-connected layers were appended to the end of each of these networks to predict nodule position, size, and malignancy. This architecture of the sliding-window CNN has been known to effectually detect malignant nodules in 3D space (Hammock 2017). Finally, a second 3D U-net was trained on hand-labeled data to predict the amount of anomalous tissue in the CT scan as opposed to the malignancy and location of nodules (cancerous tissue that did not originate in the lungs or otherwise unidentifiable tissue). The radiologist-annotated samples used malignancy labels from 1 (very likely not malignant) to 5 (very likely malignant); these labels were squared to emphasize malignant nodules in the dataset. The nodule location loss function employed was the dice coefficient due to its evaluation of spatial overlap between true and predicted nodule segmentations. The malignancy loss function employed was log loss to measure classification performance.

The malignant nodule detector 3D U-net, one of the VGG-like C3D scanning CNNs, and the anomalous tissue U-net were fit to a random sample of 80% of the lung cancer nodule data and augmented data (to up sample positives), and scored on the remaining 20% of the data. The second VGG-like C3D scanning CNN was fit to a hand-selected portion of the data possessing high false negative and positive predictions by the other networks. Finally, a linear classifier model was constructed to forecast the risk of cancer development within the next one year—this model intelligently weights the malignancy and size results of the two VGG-like C3D networks (predicted at scale of 1x and 0.5x) and U-net as well as the anomalous tissue percentage to produce a final numerical indicator of the probability of lung cancer development within the next year.

Results

This section contains analysis of the Gradient Boosted Machine (GBM) intelligent assessment survey construction and an Ensemble Convolutional Neural Network (CNN) CT scan malignant nodule detection obtained results.

The GBM patient feature survey performed exceedingly well on the validation set, with an area under the receiver optimizer characteristic (ROC) curve of 0.92 (applicable models > 0.80) as seen in Figure 14. The ROC curve plots the false positive rate versus the true positive rate at different decision thresholds, describing how well the model can differentiate between lung cancer cases and control patients. The metric of the area under the ROC curve lies between 0.5 (random guessing) and 1. Values that are closer to 1 indicate a better predictive value. In addition, the inversely correlated precision-recall curve exhibits a high curve, presenting high precision and recall in middle decision boundary ranges. Table 1 shows the predictions vs labels of the GBM. The accuracy of the model constructed was 85.13%, which is notable considering that the dataset was well-balanced (~12,000 positive and ~11,000 control samples). With a threshold of 0.513, the f1 score ($F1 = 2 * \frac{precision*recall}{precision+recall}$) was calculated to be 86.04% (Table 2). This evaluation metric accounts for false negatives and false positives, ensuring that the model has not overfit. Although age and pack-years possessed the highest predictive value, race and ethnic group, BMI, and Chronic Pulmonary Obstructive Disease or COPD (hcc 111) also had significant predictive value as seen in Figure 15. When age was limited to people 40 and greater, the accuracy of the model was 82.11%. Surprisingly, the blood pressure measurement did not influence the model. This model performs significantly better than the current age and pack-year high-risk detection method. Sample statistics calculated on the EMR dataset using the current, inflexible algorithm produced a sensitivity ($Sensitivity/Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives}$) of 23% as compared to the GBM sensitivity of 88%. After GBM intelligent survey analysis, the predicted high-risk patients would conduct a low-dosage CT scan for malignant pulmonary nodule assessment.

The Low-Dosage CT Scan Cancer Prediction CNN also performed very well. For individual scans, the dice-coefficient loss algorithm was employed to compute the intersection over union of predicted nodules. The leading CNN achieved a dice similarity coefficient of 0.8553 (between 0 and 1 where 1 signifies more similar). For individual nodule detection, the precision was 91.30% and the recall (sensitivity) was 89.72% (Table 2). This varied based on Free Response Operating The hand-labeled large pulmonary masses significantly improved the loss. In predicting cancer probability, the cross entropy loss of the validation set was 0.3943 ($\mathrm{H}(\mathrm{p},\mathrm{q}) = -\sum_{x \in X} p(x) \log{[q(x)]}$). This effectively computes the log error between the true and predicted labels in a binary distribution. Additionally, the area under the ROC curve was about 0.94, displayed in Figure 14. These results are outstanding considering the small size of the dataset. Moreover, the classification accuracy of the Ensemble CNN classifier was 87.16%. However, 69.88% of validation cases were healthy, showing an imbalance between classes. Augmentation greatly aided the classifier in reducing overfitting, diminishing the cross entropy loss by a factor of 3.23. In addition, the anomalous tissue classifier helped to a small extent, gaining the classifier a ~ 0.010 loss advantage. Morphology prediction (luminance, shape, spiculation), gradient clipping, and alternate training further aided the model. The best performing individual CNN of the Ensemble possessed a cross-entropy loss of 0.4110, with alternate training and augmentation. This attests to the power of ensembling, which allowed a somewhat substantial improvement in predictive power.

Overall, the results show that pipeline is highly effectual, with a GBM survey accuracy of 85.13% and an Ensemble CNN accuracy of 87.16% on validation sets, producing a fully-automated model for high-risk lung cancer candidate selection and malignant pulmonary nodule detection/assessment.

Discussion

This research proposes a Gradient Boosted Machine (GBM) for initial lung cancer risk assessment and a 3D Convolutional Neural Network (CNN) Ensemble for malignant pulmonary nodule detection.

The GBM employs personalized patient statistics including age, prescriptions, ethnicity, body mass index, blood pressure, and diagnoses to better assess the true risk of patients. This algorithm may have significant applications in the medical realm. This system possesses a sensitivity a factor of four times higher than the current technique, displaying a remarkable decrease in false negatives, or patients that may be misclassified by previous algorithms as not high-risk when they are truly high-risk. Exploratory data analysis (EDA) on the acquired EMR Database displayed serious flaws with the current predictive risk factor system—more than one third of the cancerous patients had less than 30 pack-years and over five percent of cancerous patients were less than 50 years old. This algorithm succeeds in capturing these previously-omitted risk groups by use of external personalized variables; Cardiovascular prescriptions and Chronic Kidney Disease Stage 3 and Chronic Pulmonary Obstructive Disease diagnoses were labeled by the algorithm as likely indicators of increased lung cancer risk. In addition, Caucasians exhibited higher lung cancer risk than Hispanic and Asian ethnic groups.

This system of affordable and fast analysis grants patients an important intelligent preliminary risk score, thereby helping increase early detection. This is essential in a disease where symptoms show solely in advanced stages, with a terminal prognosis. Expeditious detection catches nodules while in premature stages, engendering possible cure via lobectomy or chemotherapeutic agents.

Moreover, the engineered 3D CNN Ensemble for nodule detection obtains outstanding prediction results. This hierarchy of a Fully CNN for nodule candidate regions, multiple 3D CNNs for course and probabilistic nodule discrimination, and a linear regression classifier boasts great ability to determine nodules that will become malignant over time.

The ensemble algorithm yielded best results when combined with a radiomics approach; integrating lobulation, spiculation and luminance predictions into the model aided, especially in borderline malignancy predictions (difficult, indeterminate nodules). Interestingly, the Z-location of the nodule contained a somewhat high feature importance. In addition, the manual labeling of large nodules in the Luna dataset and class adjustments allowed for better one-year cancer predictions by accounting for discrepancies between datasets. Large nodules are most indicative of malignancy; therefore, correct labeling and balancing of scans possessing them greatly aided in reducing false negatives. Scan augmentations (flip, rotate, swap) alter training scans to provide the network with seemingly novel scans—this effectively increased the dataset size exponentially, increasing accuracy and reducing overfitting.

The high accuracy and AUC score display that this Ensemble Model is applicable to aid in medical diagnoses. The model's sensitivity score of 88.54% is comparable to radiologists, who vary with sensitivity levels of 30 to 97% (most experienced radiologists show levels of 80%+) depending on the nature of the input (Rubin 2016). Additionally, the false positive rate per scan is significantly less than current CAD models, allowing the model to help in the primary patient screening, rather than its current secondary screening to the radiologist standing. This decreases radiologist involvement in screening, thereby greatly improving the timeliness and affordability of detection.

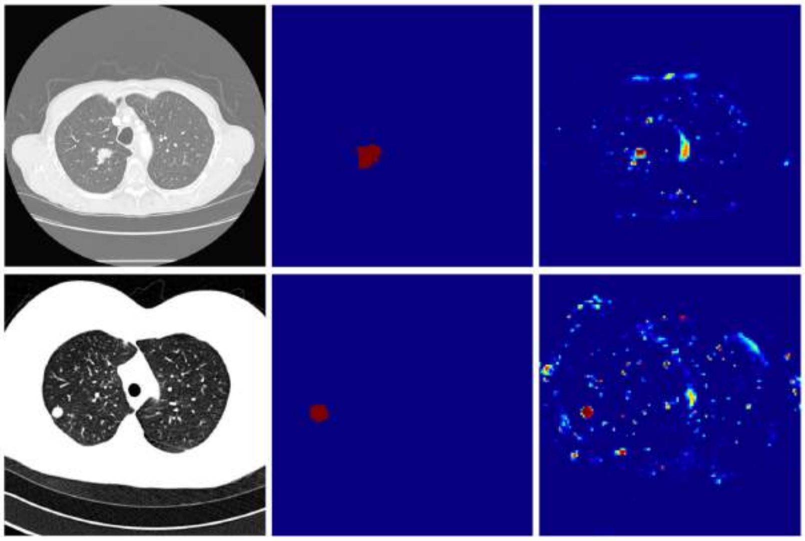


Figure 13: Visualization of FCN candidate region predictions in comparable study (Hamidian, et al 2017).

A similar study conducted by Hamidian, et al attempted the same task as the discussed research, using a Fully Convolutional Network and sole Discriminator Network (rather than Ensemble) for CT Pulmonary Nodule Detection. The model was trained on the LIDC data without preprocessing. The study reached a sensitivity of 80% with a mean 15.28 false positives per scan and showed that the Fully Convolutional Network (FCN) paired with Discriminator yields an 800-fold improvement in processing time per CT scan compared to a sole Discriminator CNN (5 seconds as opposed to 4200 seconds). This research confirms the choice of the FCN as an extremely viable solution for candidate assessment. Additionally, this study displays the outstanding accuracy of the proposed ensemble network, dataset modification, and multitudinous included training techniques, which attained almost 10% better sensitivity and many fewer false positives than the compared study.

In summary, this complete lung cancer diagnostic pipeline provides extensive benefits compared to current techniques. The algorithm utilizes personalized patient features to generate an intelligent initial risk score, with a remarkable decrease in false negatives, or patients that may be misclassified by previous algorithms as not high-risk when they are truly high-risk, compared to the in-use risk assessment; this technique captures omitted patient groups (sensitivity increased from 0.23 to 0.88), thereby helping increase early detection. The CNN Ensemble obtains statistically comparable predictions to experienced radiologist readings of scans, enabling use as a primary screening tool. The proposed algorithm also predicts a great fewer number of false positive and false negative nodules than current CAD models. The combined system increases early-detection rates and decreases radiologist involvement in screening, thereby greatly improving the timeliness, accuracy, and affordability of lung cancer detection.

Still, there are potential limitations to this model. The 3D hierarchy occupies a large amount of memory when the model grows, so the running speed is limited. In addition, the great number of trainable parameters and small dataset size may result in overfitting, and not generalize to a diverse population. Techniques such as augmentation, alternate training, and regularization were used to mitigate this problem.

Further research to improve this algorithm is possible. Primarily, the training dataset size must be increased: only ~1600 patients can not include all variations in nodules. Other untested architectures, optimizers, loss functions, and hyperparameters may produce improved results. Finally, the development of the algorithm into an app or website would enable easy access to physicians in the field.

Acknowledgements

Firstly, I would like to acknowledge Ms. Klose, for aiding in the editing process of the report and board, as well as guidance throughout the year. In addition, I would like to thank my mentor, Dr. Drew Clausen, for providing me with the resources and guidance to pursue this significant undertaking throughout the course of action. Finally, I would like to acknowledge all the people who are diagnosed with lung cancer for granting me inspiration to conduct this research and possessing with unmeasurable optimism.

References

Adetiba, Emmanuel, and Oludayo O. Olugbara. “Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features.” *The Scientific World Journal*, vol. 2015, 15 Feb. 2015, pp. 1–17., doi:10.1155/2015/786013.

Causey, J. L., Zhang, J., Ma, S., Jiang, B., Qualls, J. A., Politte, D. G., . . . Huang, X. (2018). Highly accurate model for prediction of lung nodule malignancy with CT scans. *Scientific Reports,8*(1). doi:10.1038/s41598-018-27569-w

Chon, A., Balachandar, N., & Lu, P. (2017). Deep Convolutional Neural Networks for Lung Cancer Detection. *Deep Convolutional Neural Networks for Lung Cancer Detection*. Retrieved from http://cs231n.stanford.edu/reports/2017/pdfs/518.pdf

Cruz, Joseph A., and David S. Wishart. “Applications of Machine Learning in Cancer Prediction and Prognosis.” *Cancer Informatics*, vol. 2, 2006, p. 117693510600200., doi:10.1177/117693510600200030.

Giles, Robert, Kinahan Paul, and Hrick, Hedvid. “Radiomics: Images Are More than Pictures, They Are Data.” *Radiology,* doi: 10.1148/radiol.2015151169 2015. Retrieved from https://pubs.rsna.org/doi/full/10.1148/radiol.2015151169

Girvin, Francis, and Jane P. Ko. “Pulmonary Nodules: Detection, Assessment, and CAD.” *American Journal of Roentgenology*, vol. 191, no. 4, 2008, pp. 1057–1069., doi:10.2214/ajr.07.3472.

Hamidian, S., Sahiner, B., Petrick, N., & Pezeshk, A. (2017). 3D Convolutional Neural Network for Automatic Detection of Lung Nodules in Chest CT. *Proceedings of SPIE--the International Society for Optical Engineering*, *10134*, 1013409.

Hammack, Daniel, and Julian De-Wit. “Forecasting Lung Cancer Prediction with Deep Learning.” *Kaggle Technical Reports*, Kaggle, 2017, github.com/dhammack/DSB2017/blob/master/dsb_2017_daniel_hammack.pdf.

Liao, F., Liang, M., Li, Z., & Hu, ,. (2017). Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network. Retrieved from https://arxiv.org/pdf/1711.08324.pdf.

Lung Cancer Risk Factors. (2012). Retrieved 2018, from https://www.nationaljewish.org/health-insights/health-infographics/lung-cancer-risk-factors

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics,7*. doi:10.3389/fnbot.2013.00021

Peto, R. “Smoking, Smoking Cessation, and Lung Cancer in the UK since 1950: Combination of National Statistics with Two Case-Control Studies.” *Bmj*, vol. 321, no. 7257, May 2000, pp. 323–329., doi:10.1136/bmj.321.7257.323.

Rubin G. D. (2015). Lung nodule and cancer detection in computed tomography screening. *Journal of thoracic imaging*, *30*(2), 130-8.

Shelhamer, E., J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640-651, 1 April 2017.
doi: 10.1109/TPAMI.2016.2572683

Tanoue, L.t. “Evaluation of Patients with Pulmonary Nodules: When Is It Lung Cancer?: ACCP Evidence-Based Clinical Practice Guidelines (2nd Edition).” *Yearbook of Pulmonary Disease*, vol. 2009, 2009, pp. 156–157., doi:10.1016/s8756-3452(08)79216-8.

Yu, Kun-Hsing, et al. “Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features.” *Nature Communications*, vol. 7, 2016, p. 12474., doi:10.1038/ncomms12474.

Appendix

Table 1: Confusion matrix of the Gradient Boosted Machine Results. This chart shows true positives, false positives, true negatives, and false negatives.

		True Cancer Cases		
		Negative	Positive	Error
Predicted Cancer Cases	Negative	8980	2180	0.1953
	Positive	1229	10536	0.1045
	Total	10209	12716	0.1487

Table 2: Evaluation Metrics for Gradient Boosted Machine (GBM) and Convolutional Neural Network (CNN) Ensemble Lung Cancer Development Predictions

Evaluation Metric	**GBM (%)**	**CNN Ensemble (%)**
Accuracy	85.13	87.16
F1 Score	86.04	90.50
Precision/Positive Predictive Value	82.86	91.30
Recall (Sensitivity)/ True Positive Rate	89.55	89.72
Specificity/True Negative Rate	80.47	83.23

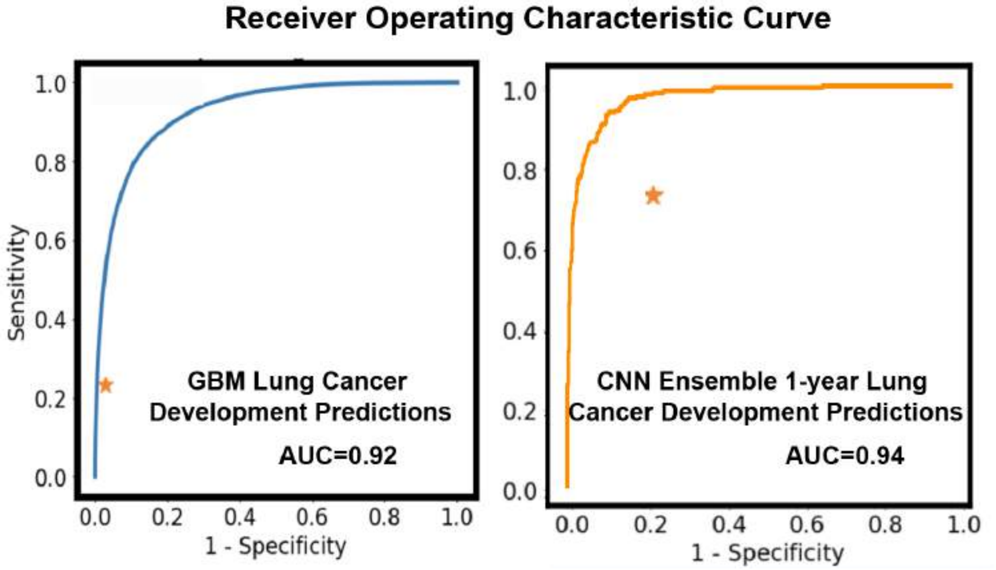


Figure 14: ROC Curves for Stage 1 and Stage 2 algorithm predictions. Orange star represents prevalent predictive algorithm specificity and sensitivity—Lung Screening Trial Recommendation (left) and Chon et al Deep 3D CNN (rights)

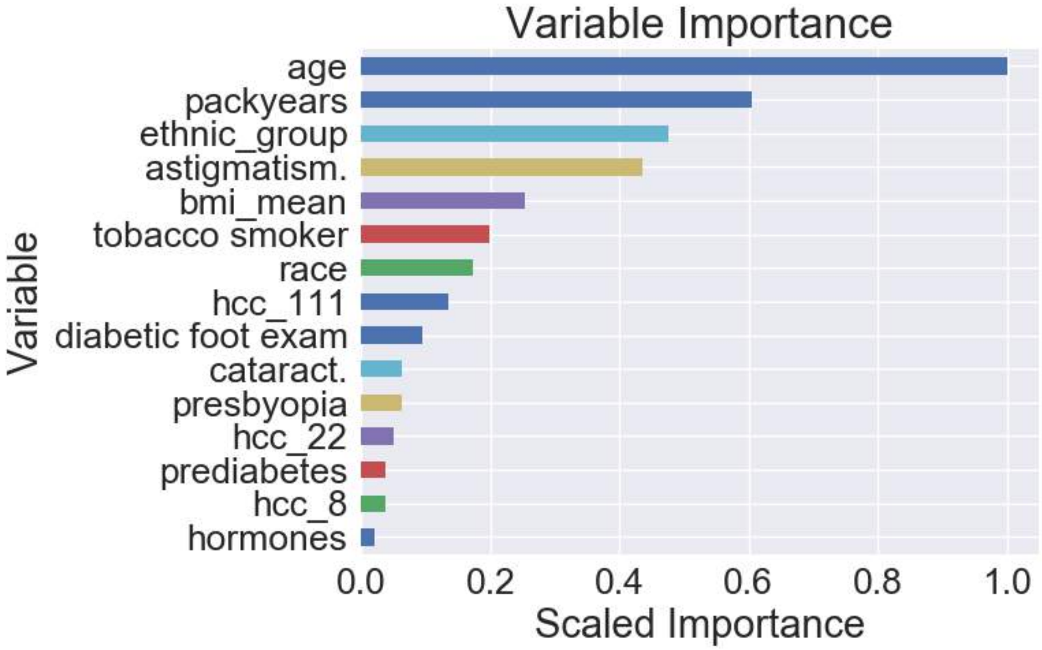


Figure 15: Gradient Boosted Machine Variable Importance Graph.

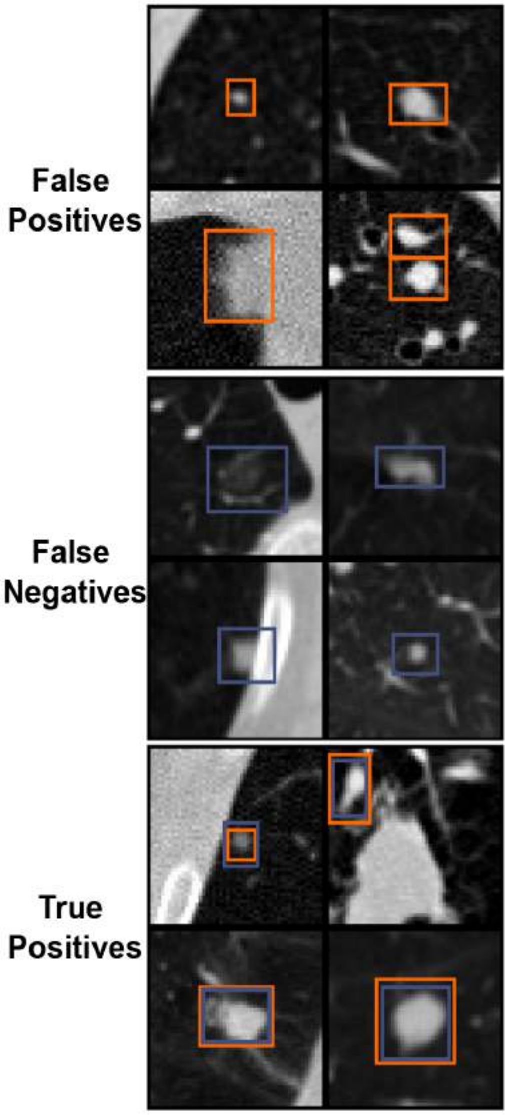


Figure 16: Analysis of CNN Ensemble False Positive, False Negative, and True Positive Predictions (orange=predicted, blue=true label)

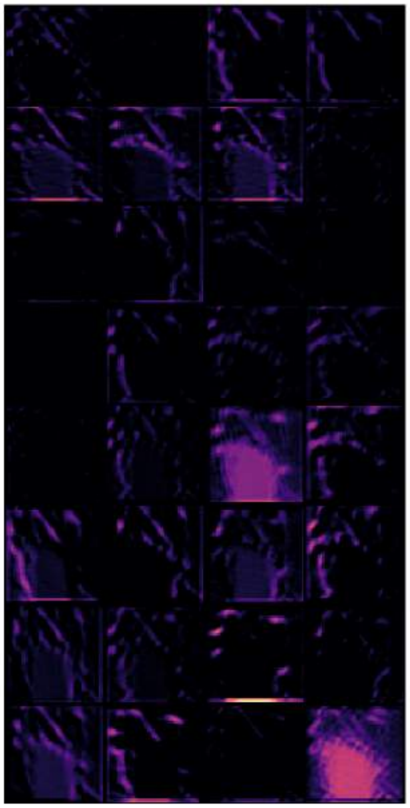


Figure 17: Visualization of Convolutional Layers