Here is what we are going to do:

目录

[Goal 1](#_Toc9513984)

[A brief introduction to tuberculosis: 1](#_Toc9513985)

[Pulmonary nodule, solitary - CT scan 1](#_Toc9513986)

[肺结核的分类以及诊断 1](#_Toc9513987)

[对应的胸部影像学检查标准如下： 2](#_Toc9513988)

[Current Paper 2](#_Toc9513989)

[Architect 2](#_Toc9513990)

[Summary 3](#_Toc9513991)

## Goal

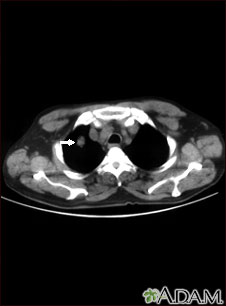
Try to classify different types of Pulmonary Tuberculosis

## A brief introduction to tuberculosis:

Pulmonary TB is caused by the bacterium Mycobacterium tuberculosis (M tuberculosis), which is contagious, meaning the bacteria is easily spread from an infected person to someone else. You can get TB by breathing in air droplets from a cough or sneeze of an infected person. **The infection may stay inactive (dormant) for years.**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Please Be Cautious About This Malady\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

### Pulmonary nodule, solitary - CT scan

This CT scan shows a single lesion (pulmonary nodule) in the right lung. This nodule is seen as the light circle in the upper portion of the dark area on the left side of the picture. A normal lung would look **completely** black in a CT scan.

## 肺结核的分类以及诊断

根据最新《中 华 人 民 共 和 国 卫 生 行 业 标 准》（2018年5月1日实施），肺结核指结核病变发生在肺、气管、支气管和胸膜等部位。分为以下 5 种类型：

a) 原发性肺结核：包括原发综合征和胸内淋巴结结核（儿童尚包括干酪性肺炎和气管、支气管结核）；

b) 血行播散性肺结核：包括急性、亚急性和慢性血行播散性肺结核；

c) 继发性肺结核：包括浸润性肺结核、结核球、干酪性肺炎、慢性纤维空洞性肺结核和毁损肺等；

d) 气管、支气管结核：包括气管、支气管粘膜及粘膜下层的结核病；

e) 结核性胸膜炎：包括干性、渗出性胸膜炎和结核性脓胸。

### 对应的胸部影像学检查标准如下：

原发性肺结核

原发性肺结核主要表现为肺内原发病灶及胸内淋巴结肿大，或单纯胸内淋巴结肿大。儿童原发性肺结核也可表现为空洞、干酪性肺炎以及由支气管淋巴瘘导致的支气管结核。

血行播散性肺结核

急性血行播散性肺结核表现为两肺均匀分布的大小、密度一致的粟粒阴影；亚急性或慢性血行播散性肺结核的弥漫病灶，多分布于两肺的上中部，大小不一，密度不等，可有融合。儿童急性血行播散性肺结核有时仅表现为磨玻璃样影，婴幼儿粟粒病灶周围渗出明显，边缘模糊，易于融合。

继发性肺结核

继发性肺结核胸部影像表现多样。轻者主要表现为斑片、结节及索条影，或表现为结核瘤或孤立空洞；重者可表现为大叶性浸润、干酪性肺炎、多发空洞形成和支气管播散等；反复迁延进展者可出现肺损毁，损毁肺组织体积缩小，其内多发纤维厚壁空洞、继发性支气管扩张，或伴有多发钙化等，邻近肺门和纵隔结构牵拉移位，胸廓塌陷，胸膜增厚粘连，其他肺组织出现代偿性肺气肿和新旧不一的支气管

播散病灶等。

d) 气管、支气管结核

气管及支气管结核主要表现为气管或支气管壁不规则增厚、管腔狭窄或阻塞，狭窄支气管远端肺组织可出现继发性不张或实变、支气管扩张及其他部位支气管播散病灶等。

e) 结核性胸膜炎

结核性胸膜炎分为干性胸膜炎和渗出性胸膜炎。干性胸膜炎为胸膜的早期炎性反应，通常无明显的影像表现；渗出性胸膜炎主要表现为胸腔积液，且胸腔积液可表现为少量或中大量的游离积液，或存在于胸腔任何部位的局限积液，吸收缓慢者常合并胸膜增厚粘连，也可演变为胸膜结核瘤及脓胸等。

## Current Paper

In Google, we can easily get access to researches in pulmonary tuberculosis with deep learning and machine learning, mainly concerning DCNN. However, it seems that most of the current researches are involved in the aspect of X-ray, rather than CT scan. You can search in Google Scholar.

## Architect

First, we have to figure out how many samples we could possess. Actually, different models suit different scales of data. Here is a post discussing this problem. (<https://machinelearningmastery.com/much-training-data-required-machine-learning/>). One paper based on the X-ray images suggested that they had manipulated more than 100,000 anonymized CXR images and their associated data which compiled from more than 32,000 patients.

Second, we have to evaluate the accuracy of our predictions. It seems that a rate around 0.85 on the training set or a number 0.75 on the test set is acceptable. We should consult current papers to make it sure.

Third, coding is necessary. We could ask help from Keras, a high-level neural networks API, written in Python and capable of running on top of TensorFlow, CNTK, or Theano. Personally, I prefer TensorFlow since Google has published the 2.0 edition, much easier to use. However, we should bear in mind that python 3.7 is not compatible on this API now, and 3.6 is all right.

Forth, it’s time for parameter tuning and diagnose. These procedures are extremely tiresome and annoying. But fortunately, that’s the last step.

## Summary

我们必须考虑数据的问题。数据是有个性的，任何算法都必须符合数据的禀性。所以，在数据到来之前，只需要一点初步的讨论。任何过于深入的探讨都可能使徒劳无功的。

如果我们只有1000个病人样本，每个样本由50张图像构成。那么可以估计，计算的性能是没有任何问题的。任何拥有GTM 950及以上显卡的笔记本电脑都可以瞬间完成工作。但是由于数据量过小，恐怕DCNN或者CNN效果都保证不了。一般来说，对于分类问题，每一类的样本量都至少在1000以上才适合深度学习。因此可以先考虑机器学习算法，主要有softmax regression/randomForest/xgboost/lightBGM等。一般模型融合后效果会不错。但是我没有处理过医疗影像，不太确定。

对于样本，我们应该分成三部分。建议按照65%:25%:10%的比例相应划分数据，对应分为训练集、交叉验证集、测试集。除了汇报结果时，任何时候都不应该动用测试集。我们在训练集上训练模型，用交叉验证集选择超参数，用测试集去汇报结果。

模型应该由简到繁，不断根据模型诊断的意见添枝加叶。不应本末倒置，浪费时间和算力。

后期可以尝试添加其他的特征提高预测准确率与稳健性。同时注意不可以只关注预测准确率，F1等几个指标也很要紧。