# 基于影像和检验数据的联合智能新型冠状病毒诊断

**Abstrat**

自2019年11月武汉发现第一个新型冠状病毒肺炎病例来，中国乃至全球经过一系列的传播已有超10万的被传染者。新型冠状病毒既给病患带来极大的痛苦乃至死亡，也对正常社会秩序和发展造成极大的破坏。在前线医护以大勇气对病患实施救治的同时，本团队基于科研工具的支持研发了一种针对新型冠状病毒的人工智能诊断方法。此方法结合了患者的影像表现数据和检验数据，影像数据使用卷积神经网络对其进行分类，检验数据则使用XGBoost进行分类。两种方法最后再进行结合，得到非常好的结果。

**Introduction**

新型冠状病毒肺炎是人类历史中第一个引起全球性的传染性疾病，其病毒具有传播能力强，隐蔽性强等特点[17]。病毒在人类的呼吸系统中生存，造成了肺部炎症感染[18]，甚至攻击人的脑部和肾脏[16][21]。期间肺部感染的CT呈双肺感染有毛玻璃症状[18]，在血常规、肌酶、凝血功能、炎性指标等人体功能上也呈现异常情况[16]。基于这两种明显的数据差异，一些特殊的机器学习和人工智能方法能够通过学习得到模型来判断新型冠状病毒病人和正常非新型冠状病毒病人。

根据权威调查显示，新型冠状病毒除了攻击肺部，还会对肌肉、心脏、血管、肝胆、肾脏甚至对脑组织进行攻击，因此实验室检查中患病患者会有多达14种人体功能出现变化[16]。它们具体表现为：发病早期外周血白细胞总数正常或减少，可见淋巴细胞计数减少；肌肉相关酶，即肝酶、乳酸脱氢酶、肌酶和肌红蛋白会增高，部分危重者可见肌钙蛋白增高；多数患者c反应蛋白(CRP)和血沉升高，降钙素原正常；严重者D-二聚体升高、外周血淋吧细胞进行性减少。然而普通肺炎降钙素原一般升高，且肌肉相关酶也表现不明显，D-二聚体和外周血淋吧细胞同样也表现不明显。因此基于新型冠状病毒的特殊实验室检查结果，通过一些机器学习的分类方法能很明显的将新型冠状病毒的病人筛选出来。

机器学习方法如SVM[6]、随机森林[2]、决策树[1]、卷积神经网络[4]、循环神经网络[7]等在当前社会应用广泛。它们在互联网产业中推动了很多成功产品，也在医疗行业开始显示它们的惊人本领，有U-NET[8]等卷积神经网络帮助影像医生看CT、MRI图像，有SVM，决策树等帮助基因测序和分类[9][10][11]，有LSTM等循环神经网络帮助医疗诊断[12]，有GCN[5]等图神经网络和GAN网络[14]帮助寻找化合物[13][15]等等。这些机器学习方法大力助推当前医学的发展。

本团队是基于医院的团队，所在市共有672名新型冠状病毒肺炎患者，有重症患者xx名。本医院为新型冠状病毒肺炎收治定点医院之一，共收治了xx名患者，其中重症有xx名。基于本团队能够拿到的实验室数据不是很多，本团队采用今年来在小数据大数据上都表现很好的工具XGBoost分类器[3]来构建新型冠状病毒肺炎的分类模型。XGBoost是一个基于决策树CART（或者说GBDT）的Gradient Boosting方法。XGboost方法这种使用boosting的方法能提高模型的准确率，同时XGBoost的线形器能够使用L1和L2正则化来防治模型过拟合，而且XGBoost能很好的处理稀疏数据。这些特点非常适合本团队的数据规模。因此本团队使用XGBoost工具来拟合我们最后的分类模型。

我们将依数据介绍，XGboost方法使用，实验结果，软件生成这4个部分来介绍我们的方法。

**Method**

**数据介绍**

新型冠状病毒引起的病毒型感染肺炎主要表征在人体的肺部大面积炎症感染[18]，但是也会攻击脾脏、心脏血管、肝胆、肾脏乃至脑组织和胃[16]。新型冠状病毒对人体多器官的攻击造成人类的多个实验室指标发生变化，根据中国广大医务工作者的总结，很多指标会明显发生变化。它们具体是：

表格 1 新冠病毒患者实验室检查信息表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 一般或多数患者表现 | | | | | | | | |
| 外周血白细胞总数 | 可见淋巴细胞计数 | 肝酶 | 乳酸脱氢酶 | 肌酶 | 肌红蛋白 | C反应蛋白（CRP） | 血沉 |
| 减少 | 减少 | 增高 | 增高 | 增高 | 增高 | 升高 | 升高 |
| 严重危重部分表现 | | | | 特殊情况 |  | | |
| 肌钙蛋白 | D-二聚体 | 外周血淋巴细胞 | 炎症因子 | 降钙素原 |
| 增高 | 升高 | 进行型减少 | 升高 | 正常 |

上表中，和免疫功能有关的外周血白细胞总数，淋巴细胞计数减少。和炎症有关的c反应蛋白和血沉升高。肝酶即丙氨酸氨基转移酶和天冬氨基酸转移酶增高，这些酶和肝脏系统有关。乳酸脱氢酶、肌酶和肌红蛋白这些和肌肉包括心肌相关的指标也升高。严重者一些其他相关指标也表现异常，肌钙蛋白和心肌有关，D-二聚体和纤维蛋白溶解有关，外周血淋巴细胞即T细胞和B细胞和免疫功能有关，炎症细胞因子升高。期间特殊的是降钙素原相比普通肺炎反而保持正常。

该病毒攻击范围广，和普通肺炎表现区别在基本都会攻击肌肉，炎症反应大也广，所以在实验室检查指标上有其非常特殊的地方。我们完全可以利用这些显著特点做出一个区别新型冠状病毒肺炎和非新型冠状病毒肺炎的正常人或者其他病人的两类分类器，即新型冠状病毒肺炎作为我们的正例，非新型冠状病毒肺炎的正常或其他感冒类或者普通肺炎病人都统作为负类。而我们选取新型冠状病毒肺炎的上述实验室检查指标为特征，因为男女年龄的差异性，性别年龄这里也作为特征。

本团队统计了院内自1月22日钟南山院士认定人传人以来来本院发热门诊检查的所有病人或者正常人的检查数据。共有xx名发热病人，其中是新型冠状病毒肺炎患者有xx名，其他病患或者正常人为了类别平衡随机选取了xx名。其中因为必要和不必要的原因，有占了xx%的病人都做了上述的共13种检查或者其中的多种。数据还是具有一点的稀疏性，数量少的特点。下表是具体的检测数目表（？？？？）

**方法描述**

根据特征的数据特点：数据规模不大且具有一定的稀疏性，本团队决定使用一种现在被广泛使用的分类工具XGBoost来作为分类器。XGBoost是一个基于决策树的Boosting方法。特别的是，XGboost的基本部分是决策树CART或者说回归树，然后使用Boosting方法来组合多颗CART树来提升准确度。它的基本过程为先基本确定前面的树，然后通过结果残差不断提升后续的树，这是一个不断Boosting（提升）的过程。每棵树的构造又是一个回归树，通过拟合残差来构造回归树。

根据上述过程，整个过程的损失函数为下列公式：

(1)

公式（1）中的为惩罚项，可以使用L1，L2正则。正则化可以防止过拟合。而且在每棵树的构造过程中还会使用学习率和列采样。学习率即权重更新快慢，防止变化过快。列采样即每棵树构造都会随机选取一定数目的特征来构造。这三个方式都会有效防止过拟合[3]。

XGBoost针对缺失值使用Sparsity-aware Split Finding方法，即在缺失数值那里分裂时会给一个默认的方向让它继续分裂下去生成树。树的节点的默认左右两个方向会通过学习得到[3]。这个学习过程等于说根据数据情况处理了默认值，所以也是合理利用了数据集，有利于整个模型。

因此，XGBoost的多项防止过拟合以及处理缺失值的方式非常有利于本团队的数据集。

**Experiment**

通过数据库的查询，本团队所在医院共检查了20个项目，包括年龄和性别。它们分别是白细胞数、淋巴细胞绝对值、淋巴细胞百分比、丙氨基酸基转移酶、天冬氨基酸基转移酶、乳酸脱氢酶、肌酸肌酶、肌红蛋白、心肌钙蛋白、高敏C反应蛋白、血沉、降钙素原、D-二聚体、T细胞和B细胞、年龄和性别。共检查了xx名病人，每项检查数目为下表（或者列出来？）

通过数据特征，XGBoost的参数拟使用树的个数，学习率，列采样，树深度，L1正则化和L2正则化。因为前面数据分析了差异性，所以有些特征可能有决定性作用，因此本团队会使用特征选择加强结果。在实验中，本团队为了防止过拟合还会对数据进行归一化处理，。。。。

实验结果balabala。。。

**Conclusion**

新型冠状病毒肆虐全球，疫情紧迫，攻克认识它是医生的职责，也是每一个科研者应该要做的。本团队基于本地医院的数据使用XGBoost对新型冠状病毒的实验室检查做了个二类分类器，结果为XX，表现优异。实验结果中发现，xx和xx以及xx对分类影响较大，说明新型冠状病毒区别在普通肺炎或者感冒以及正常人上对xx器官的确攻击较大。

Reference

1. Denison David G. T, Mallick Bani K, Smith Adrian F. M. A Bayesian CART algorithm[J]. Biometrika, 1998(2):2.
2. Breiman L . Random Forests[J]. Machine Learning, 2001, 45(1):5-32.
3. Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.
4. Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).
5. Kipf T N , Welling M . Semi-Supervised Classification with Graph Convolutional Networks[J]. 2016.
6. Chang C C, Lin C J. LIBSVM: a library for support vector machines," 2001. Software available at http://www. csie. ntu. edu. tw/~ cjlin/libsvm[J]. 2001.
7. Hochreiter, S, Schmidhuber, J. Long Short-Term Memory[J]. Neural Computation, 9(8):1735-1780.
8. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
9. Brown M P S , Grundy W N , Lin D , et al. Knowledge-Based Analysis of Microarray Gene-Expression Data by Using Support Vector Machines[J]. Proceedings of the National Academy of Sciences, 2000, 97(1):262-267.
10. Bartsch Jr G, Mitra A P, Mitra S A, et al. Use of artificial intelligence and machine learning algorithms with gene expression profiling to predict recurrent nonmuscle invasive urothelial carcinoma of the bladder[J]. The Journal of urology, 2016, 195(2): 493-498.
11. Wang C W. New ensemble machine learning method for classification and prediction on gene expression data[M]. Encyclopedia of Healthcare Information Systems. IGI Global, 2008: 982-989.
12. Lipton Z C, Kale D C, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks[J]. arXiv preprint arXiv:1511.03677, 2015.
13. Benhenda M. ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity?[J]. arXiv preprint arXiv:1708.08227, 2017.
14. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]. Advances in neural information processing systems. 2014: 2672-2680.
15. Xu N, Wang P, Chen L, et al. MR-GNN: multi-resolution and dual graph neural network for predicting structured entity interactions[J]. arXiv preprint arXiv:1905.09558, 2019.

1. [Dahai Zhao](javascript:;), [Feifei Yao](javascript:;), [Lijie Wang](javascript:;), [Ling Zheng](javascript:;), [Yongjun Gao](javascript:;), [Jun Ye](javascript:;), [Feng Guo](javascript:;), [Hui Zhao](javascript:;), [Rongbao Gao](javascript:;). A comparative study on the clinical features of COVID-19 pneumonia to other pneumonias. Clinical infectious diseases, ciaa247. 2020
2. Zhang L, Shen F, Chen F, et al. Origin and evolution of the 2019 novel coronavirus[J]. Clinical Infectious Diseases, 2020.
3. Lei J, Li J, Li X, et al. CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia[J]. Radiology, 2020: 200236.
4. Xu, X., Yu, C., Zhang, L. *et al.* Imaging features of 2019 novel coronavirus pneumonia. Eur J Nucl Med Mol Imaging (2020)
5. Han Y, Yang H. The transmission and diagnosis of 2019 novel coronavirus infection disease (COVID‐19): A Chinese perspective[J]. Journal of Medical Virology, 2020.
6. Zhang C, Shi L, Wang F S. Liver injury in COVID-19: management and challenges[J]. The Lancet Gastroenterology & Hepatology, 2020.