Classifying COVID-19 Severity using ML

Mariam Grigoryan
Yejin Cha
Gordon Kong

03/02/2021

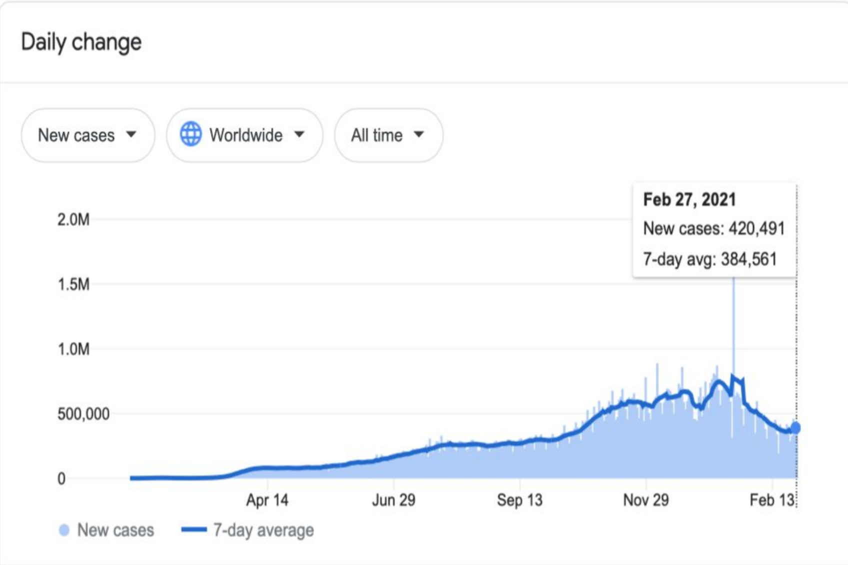
Research Question and Motivation

Aim to classify COVID-19 severity into 3 categories:

- Mild
- Moderate
- Severe

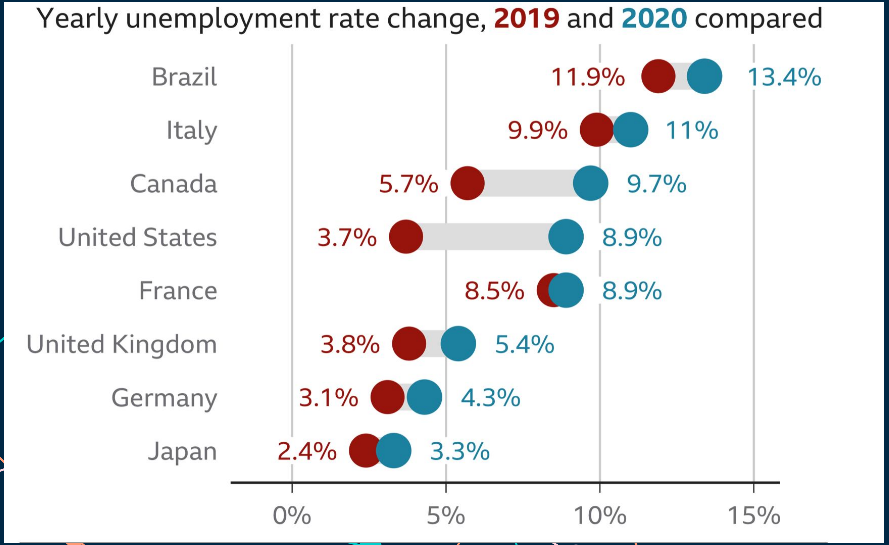
114,986,566 cases total

2.54M deaths



Daily change

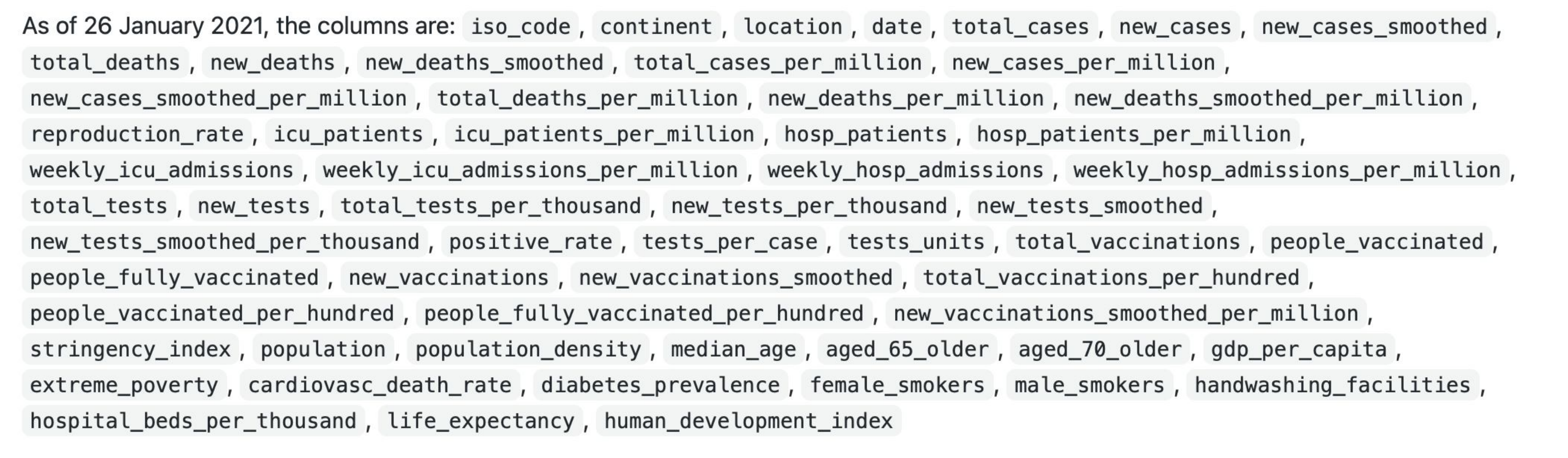
Feb 27, 2021
New cases: 420,491
7-day avg: 384,561



Yearly unemployment rate change, **2019** and **2020** compared

Data Sources

- Country-specific, frequently updated data
- https://ourworldindata.org

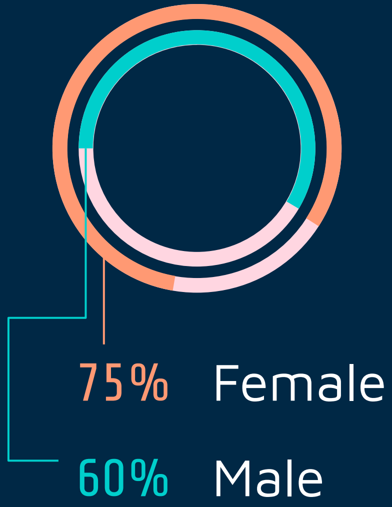


As of 26 January 2021, the columns are: `iso_code`, `continent`, `location`, `date`, `total_cases`, `new_cases`, `new_cases_smoothed`, `total_deaths`, `new_deaths`, `new_deaths_smoothed`, `total_cases_per_million`, `new_cases_per_million`, `new_cases_smoothed_per_million`, `total_deaths_per_million`, `new_deaths_per_million`, `new_deaths_smoothed_per_million`, `reproduction_rate`, `icu_patients`, `icu_patients_per_million`, `hosp_patients`, `hosp_patients_per_million`, `weekly_icu_admissions`, `weekly_icu_admissions_per_million`, `weekly_hosp_admissions`, `weekly_hosp_admissions_per_million`, `total_tests`, `new_tests`, `total_tests_per_thousand`, `new_tests_per_thousand`, `new_tests_smoothed`, `new_tests_smoothed_per_thousand`, `positive_rate`, `tests_per_case`, `tests_units`, `total_vaccinations`, `people_vaccinated`, `people_fully_vaccinated`, `new_vaccinations`, `new_vaccinations_smoothed`, `total_vaccinations_per_hundred`, `people_vaccinated_per_hundred`, `people_fully_vaccinated_per_hundred`, `new_vaccinations_smoothed_per_million`, `stringency_index`, `population`, `population_density`, `median_age`, `aged_65_older`, `aged_70_older`, `gdp_per_capita`, `extreme_poverty`, `cardiovasc_death_rate`, `diabetes_prevalence`, `female_smokers`, `male_smokers`, `handwashing_facilities`, `hospital_beds_per_thousand`, `life_expectancy`, `human_development_index`

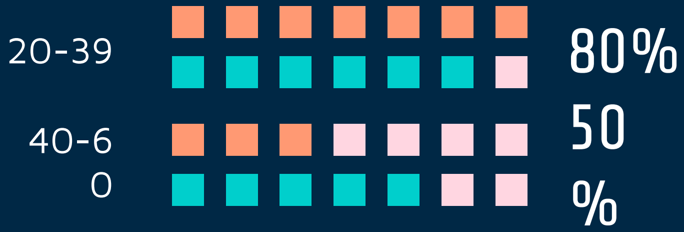
Features

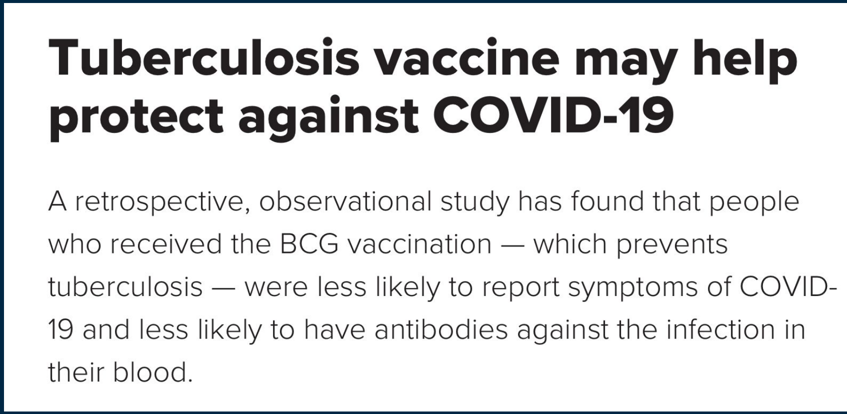
- Demographics
- Blood/urine data
- Smoking
- Other vaccinations
  - Ex: BCG

GENDER



AGE

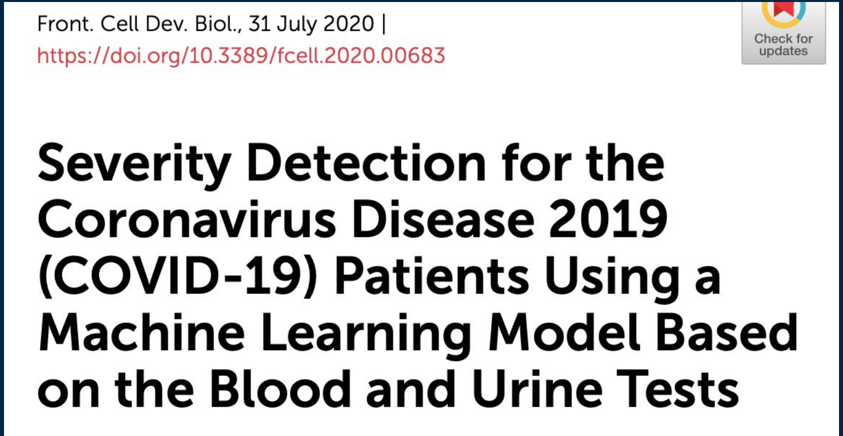




Tuberculosis vaccine may help protect against COVID-19

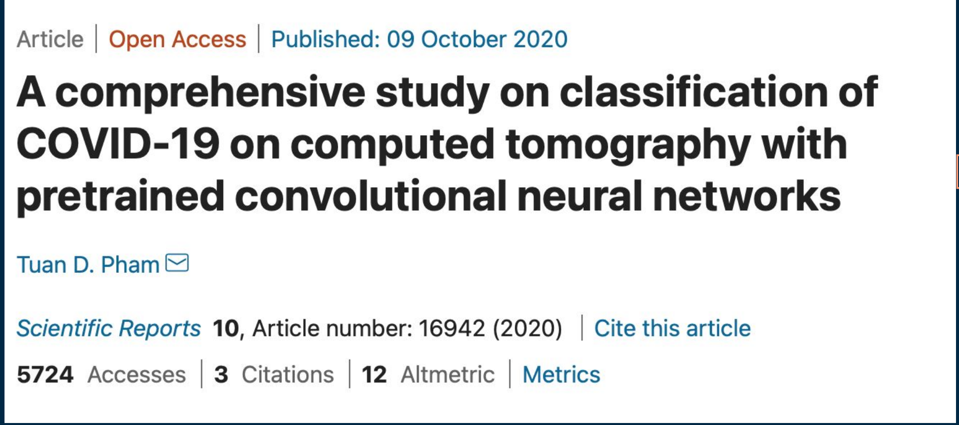
A retrospective, observational study has found that people who received the BCG vaccination — which prevents tuberculosis — were less likely to report symptoms of COVID-19 and less likely to have antibodies against the infection in their blood.

Previous Research


Front. Cell Dev. Biol., 31 July 2020 |
https://doi.org/10.3389/fcell.2020.00683

Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests

Binary Classifier:
Mild/moderate vs severe/death
Features: blood/urine data+ demographics
Data: 137 cases in China
Supervised
81% accuracy


Article | Open Access | Published: 09 October 2020

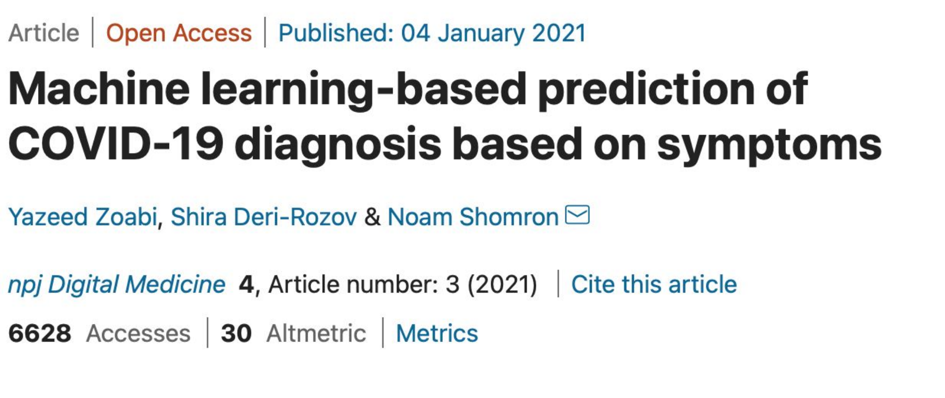
A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks

Tuan D. Pham

Scientific Reports **10**, Article number: 16942 (2020) | Cite this article

5724 Accesses | **3** Citations | **12** Altmetric | Metrics

COVID diagnosis based on CT images
Pre-trained CNN


Article | Open Access | Published: 04 January 2021

Machine learning-based prediction of COVID-19 diagnosis based on symptoms

Yazeed Zoabi, Shira Deri-Rozov & Noam Shomron

npj Digital Medicine **4**, Article number: 3 (2021) | Cite this article

6628 Accesses | **30** Altmetric | Metrics

https://www.nature.com/articles/s41746-020-00372-6

Planned ML Approach and Rationale

Data Pre-processing

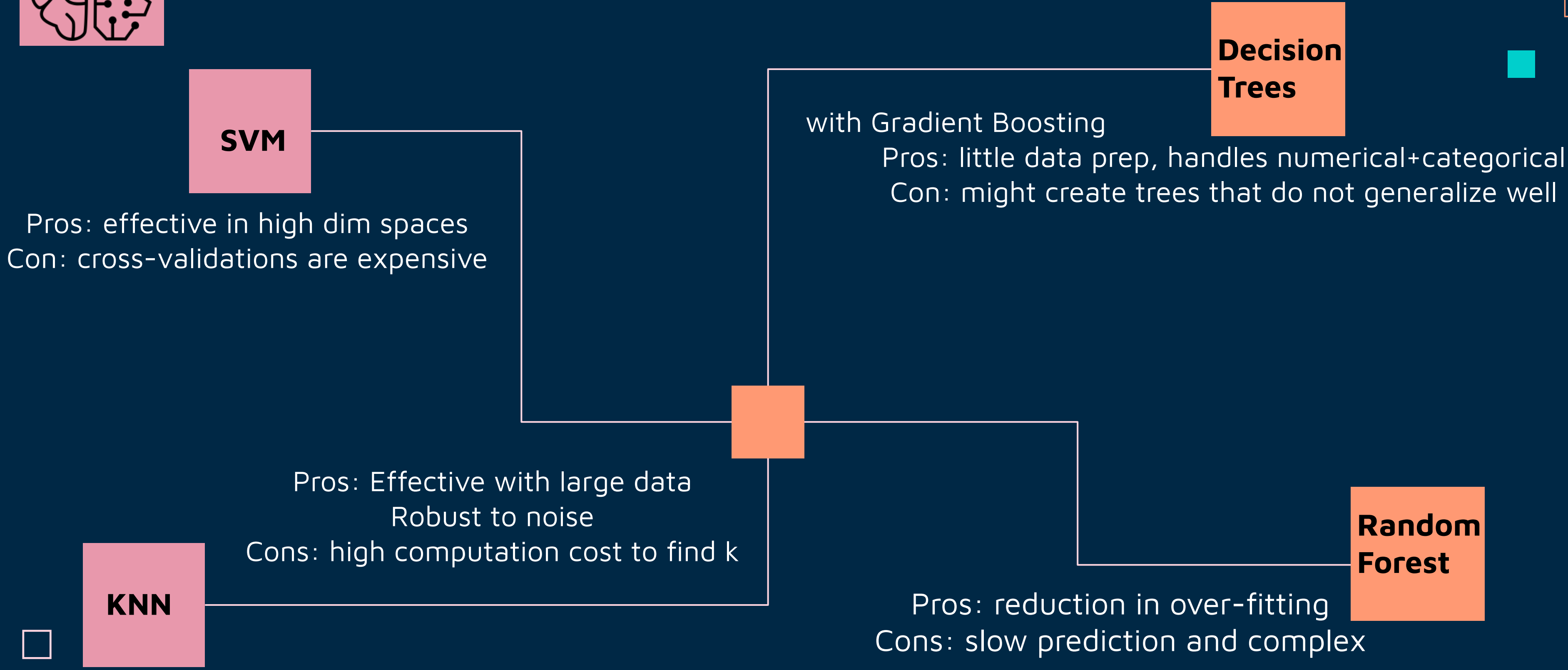
- Fill the missing entries with median of that normal range or 0
- Split samples at 80% training, 20% test datasets
- Normalize continuous values by the values in the training dataset
- Encode categorical features by one-hot strategy

Feature Selection

- Student t-test to evaluate the statistical association of each feature with the disease severity of the sample. Rank features based on their significance (p-values)

- Use Lasso (Least Absolute Shrinkage and Selection Operator) for identifying the most relevant features

Models



SVM

Pros: effective in high dim spaces
Con: cross-validations are expensive

Decision Trees

with Gradient Boosting
Pros: little data prep, handles numerical+categorical
Con: might create trees that do not generalize well

Pros: Effective with large data
Robust to noise
Cons: high computation cost to find k

KNN

Random Forest

Pros: reduction in over-fitting
Cons: slow prediction and complex

Model Evaluations

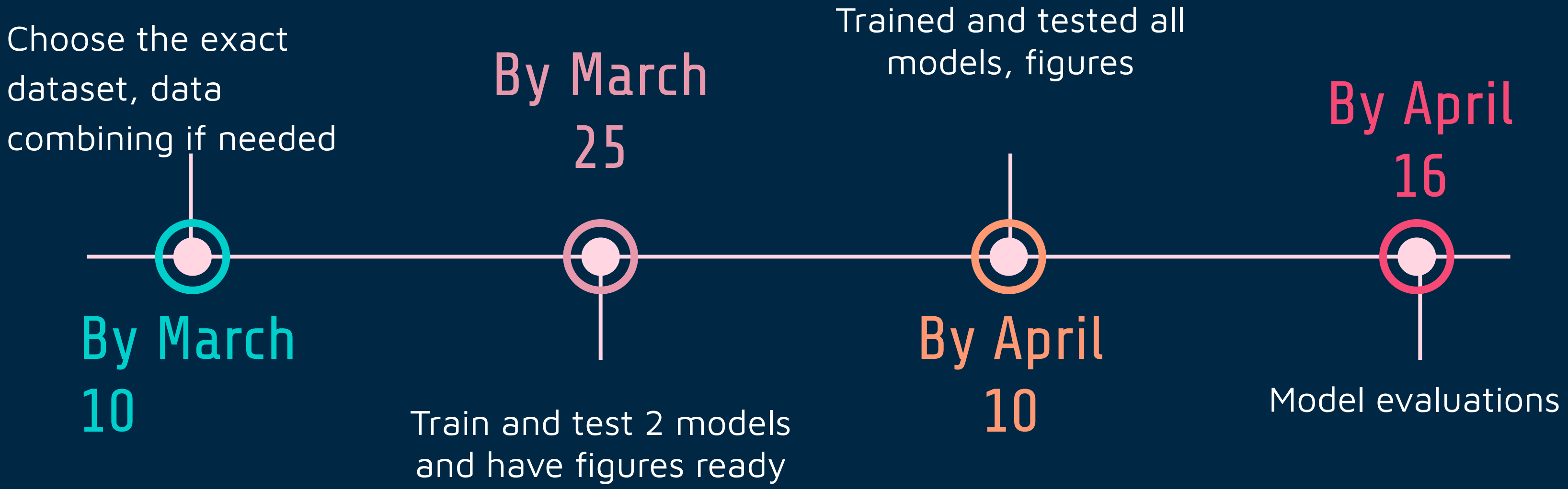
Compare the different models with the following metrics:

- **Accuracy**
- **F-1 score**
- **Specificity** (false positive, true negative rate)
- **Sensitivity** (recall, true positive, false negative rate)

Anticipated Challenges

- Keeping up with current research
- Large datasets
- Combining Multiple Datasets

Timeline



By March 10

Choose the exact dataset, data combining if needed

By March 25

Train and test 2 models and have figures ready

By April 10

Trained and tested all models, figures

By April 16

Model evaluations