## Описание используемого набора изображений.

Обучение и тестирование модели осуществляется на датасете COVID-19 Radiography Dataset, который собирается и обновляется группой исследователей из Катарского университета, Доха, Катар, и Университета Дакки, Бангладеш, вместе исследователями из Пакистана и Малайзии в сотрудничестве с врачами. На момент исследования база данных содержит 21165 рентгеновских снимков грудной клетки для случаев COVID-19, а также изображений нормальной и вирусной пневмонии. Этот набор данных по COVID-19, нормальным и другим легочным инфекциям.

Структура данных.

Датасет состоит из четырех частей.

1 Изображения, маркированные COVID:

3615 изображений собрано из различных доступных публично баз данных, онлайн ресурсов, и опубликованных исследований:

* 2473 изображений из PadChest датасета. Большого набора данных рентгеновских снимков грудной клетки с аннотированными отчетами с несколькими метками.
* 183 изображений из Немецкой медицинской школы
* 559 изображений из SIRM, Github, Kaggle & Tweeter
* 400 изображений из других ресурсов Github.

2. Обычные изображения:

10192 изображения из двух датасетов:

* 8851 Радиологического общества Северной Америки (RSNA)
* 1341 Kaggle.

3 Изображения непрозрачности легких:

6012 Изображения CXR непрозрачности легких собраны из набора данных CXR Радиологического общества Северной Америки (RSNA).

4 Изображения вирусной пневмонии:

1345 Данные о вирусной пневмонии собраны из базы данных рентгеновских снимков грудной клетки (пневмония).

Изображения сохранены в формате «png». Размер изображений 299х299 пикселей. Весь набор изображений размечен метками – «COVID», «Normal», «Viral Pneumonia», «Lung\_Opacity».

Датасет загружен на сетевой диск Google Drive, который подключен к проекту в среде Google Colaboratory. Управление подключением/отключением сетевого диска к проекту осуществляется с помощью модуля drive из google.colab.

## Техническая реализация.

Программная реализация ансамбля нейронных сетей выполнена на языке Python. Наличие широкого выбора готовых модулей для обработки данных и реализаций многих архитектур нейронных сетей.

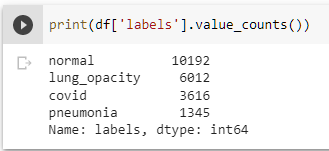
Вычисления производились на удаленном сервере Google Colaboratory.

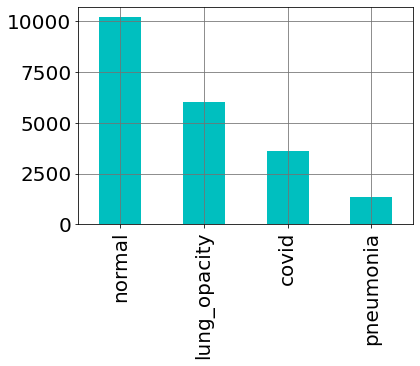
## Структура программного обеспечения

В модели используются три сверточные нейронные сети, описанные в Главе 1

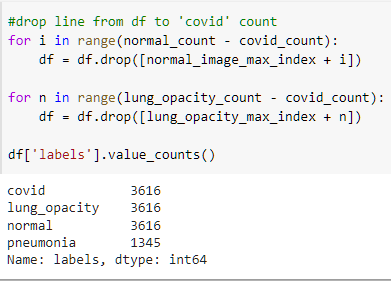
В первой части скрипта реализовано подключение библиотеки TensorFlow для экземпляров классов исследуемых нейронных сетей.

В подключенной библиотеке MatplotLib реализованы методы построения графиков, на основе передаваемых данных.



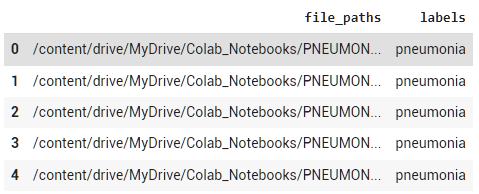


Выравниваем количество изображений по классам



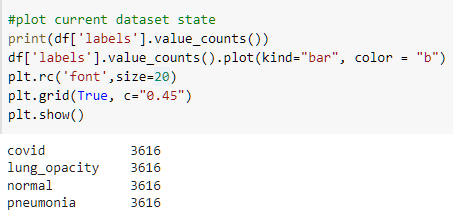
Добавляем второй датасет

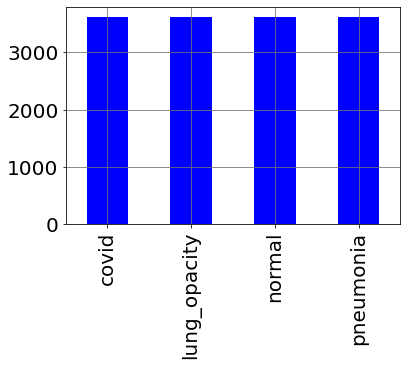
Chest X-Ray Images (Pneumonia) Dataset



Выравниваем датасет.

Результат



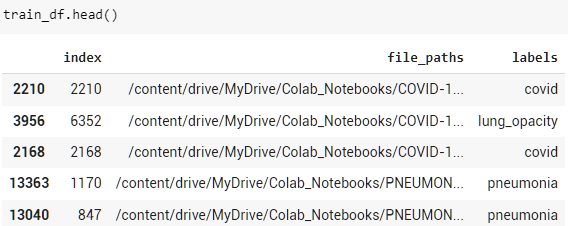


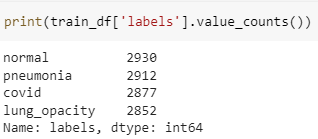
В итоге в датасете в каждом классе равное количество изображений

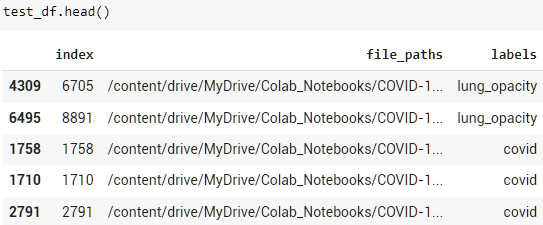
Разделение датасета на тренировочную часть и тестовую.

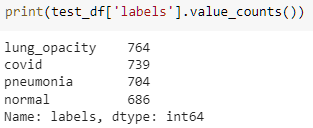
Для обучения сетей используется тренировочный датасет, это часть в 80% из общего объема датасета, Применено случайное перемешивание.

В результате разделения датасеты состоят из









– Train the models - DenseNet201, ResNet50V2, and Inceptionv3 using training set images and do the loss minimization based on the test set images. Calculate the weights of the 5-fold cross validation based on the test set.

– Run the trained models on the validation set images and select class label value 0 or 1 based on weighted average ensembling of the 3 models.

The overall proposed approach, as summarized in Fig. 6, includes:

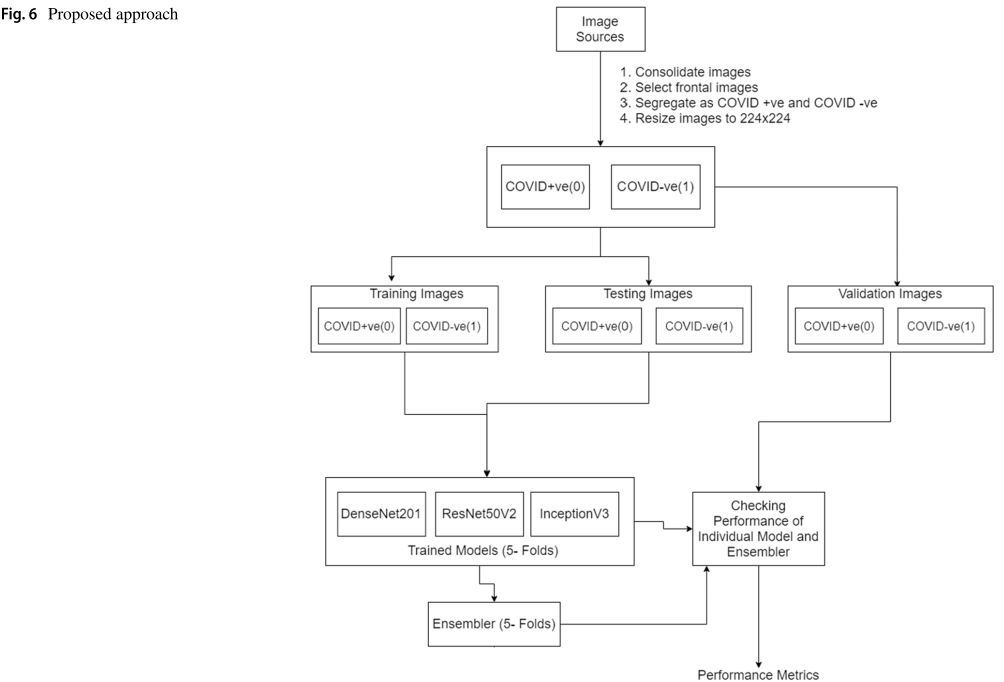
– Consolidate CXR images for healthy subjects, patients having pneumonia or other bacterial infection and COVID patients from different sources.

– Retain only frontal CXR images.

– Resize images to a uniform size.

– Divide the images into three portions—training, testing and validation datasets. One small portion is retained as validation set to test the efficacy of the trained model while the remaining portion is divided into 5 folds. Each time one separate fold is picked up as test data and the remaining folds as training data.

– While dividing the images into training, testing and validation sets, ensure that there is no patient overlap i.e., different images of the same patient is not present in multiple sets.



Выбор функций активаций

Выбор нейронных сетей

Предоработка

Метрики качества

Сравнение с бенчмарком

Для сравнения качества модели стоит сравнить с показателями отдельных сетей, составляющих ансамбль.

Сколько эпох обучения было выбрано.

Какой оптимизатор был выбран

Какие другие параметры обучения были выбраны.

Какое время в эпохах заняло обучение?

Матрицы смешения (конфьюза)