MosMedData: результаты исследований компьютерной томографии органов грудной клетки с признаками COVID-19

Данный набор данных (датасет) содержит результаты компьютерной томографии органов грудной клетки с рентгенологическими признаками вирусной пневмонии (COVID-19), а также без признаков (норма). Для некоторых исследований представлена разметка областей интереса (зон уплотнений по типу «матового стекла» и консолидации) в виде бинарной пискельной маски. Данные исследований были собраны в отделениях лучевой диагностики лечебных учреждений города Москвы в период с 01.03.2020 по 25.04.2020.

Общая информация

Название набора данных

MosMedData: результаты исследований компьютерной томографии органов грудной клетки с признаками COVID-19

Внутренний код

COVID19_1110

Классы разметки

2–C, 2–A

Ключевые слова

компьютерная томография, КТ, органы дыхательной системы, вирусная, инфекция, легкие, грудная клетка, COVID-19

Язык

Английский, русский

Финансирование

Внутреннее финансирование

Версия набора данных

1.0

Постоянная ссылка

https://mosmed.ai/datasets/covid19_1110

Дата публикации

28.04.2020

Аффилиация и авторы

Авторы

- Морозов Сергей Павлович[1]
- Андрейченко Анна Евгеньевна[1]
- Блохин Иван Андреевич[1]
- Владзимирский Антон Вячеславович[1]
- Гележе Павел Борисович[1]
- Гомболевский Виктор Александрович[1]
- Гончар Анна Павловна[1]
- Ледихова Наталья Владимировна[1]
- Павлов Николай Александрович[1] (n.pavlov@npcmr.ru)
- Чернина Валерия Юрьевна[1]

Аффилиация

1. Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»

Структура набора данных

```
.
|-- dataset_registry.xlsx
|-- LICENSE
|-- README_EN.md
|-- README_RU.md
|-- README_EN.pdf
|-- README_RU.pdf
|-- masks
|   |-- study_BBBB_mask.nii.gz
|   |-- ...
|   `-- study_BBBB_mask.nii.gz
`-- studies
    |-- CT-0
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    |-- CT-1
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    |-- CT-2
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    |-- CT-3
    |   |-- study_BBBB.nii.gz
    |   |-- ...
    |   `-- study_BBBB.nii.gz
    `-- CT-4
        |-- study_BBBB.nii.gz
        |-- ...
        `-- study_BBBB.nii.gz
```

- `README_EN.md` и `README_RU.md` содержат общую информацию о наборе данных в формате `Markdown` на английском и русском языках соответственно; та же информация в формате `PDF` представлена в `README_EN.pdf` и `README_RU.pdf`.
- `LICENSE` содержит описание лицензии Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported (CC BY-NC-ND 3.0) License.
- `dataset_registry.xlsx` содержит перечень исследований, включенных в набор данных, путь к соответствующему файлу и путь к маске (при наличии).
- В директории `studies` находятся директории `CT-0`, `CT-1`, `CT-2`, `CT-3` и `CT-4`, в каждой из которых содержатся исследования в формате `NIfTI`, заархивированные в `Gzip`. Названия исследований построены по шаблону `study_BBBB.nii.gz`, где `BBBB` – уникальный порядковый номер исследования во всем наборе данных (сквозная нумерация).
- В директории `masks` находятся бинарные маски разметки в формате `NIfTI`, заархивированные в `Gzip`. Названия масок построены по шаблону `study_BBBB_mask.nii.gz`, где `BBBB` – порядковый номер соответствующего исследования.

Обзор данных

Параметр	**Значение**
Количество исследований, ед.	1110
Количество пациентов, чел.	1110
Распределение по полу, % (М/ Ж/ др.)	42/ 56/ 2
Распределение по возрасту, лет (мин./ медиана/ макс.)	18/ 47/ 97

Параметр	Значение
Количество бинарных пиксельных масок разметки класса A, ед.	50
Распределение по классам (разметка класса C), ед. (KT-0/ KT-1/ KT-2/ KT-3/ KT-4)	254/ 684/ 125/ 45/ 2

Особенности подготовки исследований

- Одно исследование относится к одному пациенту.
- Каждое исследование включает одну трехмерную реконструкцию в мягкотканном режиме.

```
SeriesDescription LIKE '%BODY%'
```

- При преобразовании формата `DICOM` в `NIfTI` в серии сохранено каждое 10-е изображение.

```
InstanceNumber % 10 = 0
```

Принцип разметки класса C

Исследования разделены на 5 категорий[1]:

- **KT-0** (директория `/studies/CT-0`): норма и отсутствие КТ-признаков вирусной пневмонии.
- **KT-1** (директория `/studies/CT-1`): зоны уплотнения по типу «матового стекла». Вовлечение паренхимы легкого =< 25%.
- **KT-2** (директория `/studies/CT-2`): зоны уплотнения по типу «матового стекла». Вовлечение паренхимы легкого = 25–50%.
- **KT-3** (директория `/studies/CT-3`): зоны уплотнения по типу «матового стекла» и консолидации. Вовлечение паренхимы легкого = 50–75%.
- **KT-4** (директория `/studies/CT-4`): Диффузное уплотнение легочной ткани по типу «матового стекла» и консолидации в сочетании с ретикулярными изменениями. Вовлечение паренхимы легкого >= 75%.

Обратите внимание: разделение исследований было проведено *перед* преобразованием `DICOM` в `NIfTI` и *перед* сохранением каждого 10-го изображения.

Обратите внимание: разделение исследований было проведено *на основании рентгенологических (КТ) признаков*, а не на результатах лабораторного исследования (например, с помощью полимеразной цепной реакции) или клинической верификации.

> 1. Лучевая диагностика коронавирусной болезни (COVID-19): организация, методология, интерпретация результатов : препринт № ЦДТ – 2020 – II. Версия 2 от 17.04.2020 / сост. С. П. Морозов, Д. Н. Проценко, С. В. Сметанина [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». – Вып. 65. – М. : ГБУЗ «НПКЦ ДиТ ДЗМ», 2020. – 78 с.

Принцип разметки класса A

Для ограниченного количества исследований в наборе данных (50 исследований) экспертами НПКЦ ДиТ ДЗМ созданы бинарные пиксельные маски областей интереса (зон уплотнений по типу «матового стекла» и консолидации) (директория `/masks/`). Маски сохранены в формате `NIfTI` и заархивированы в `Gzip`. Маски имеют те же координаты, что и соответствующее исследование.

При разметке использовалось программное обеспечение MedSeg (© 2020 Artificial Intelligence AS).

Правила использования и распространения

Лицензия

Цитирование

Рекомендованная форма для цитирования:

> Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzymyrskyy, A.V., Ledikhova, N.V., Gombolevskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A.V. and Chernina, V.Y., 2020. MosMedData: Chest CT Scans With COVID-19 Related Findings Dataset. *arXiv preprint arXiv:2005.06465*.

Распространение

Данный датасет не должен распространяться без указания:

- авторов;
- аффилиаций;
- правообладателя (копирайта);
- постоянной ссылки на оригинальный набор данных;
- ссылки на лицензию.