# Data

1. Coswara

* Data size: **4,465** audios (two types: shallow cough and heavy cough) - *.wav*
* Folder: [coswara\_cough\_data](https://drive.google.com/drive/folders/19q9fvUYGgttRT0esacdWMRWJ67VGuA2B?usp=sharing)
* Metadata: [coswara\_metadata.csv](https://drive.google.com/file/d/1rfiPzRgmkC5xQgXIhUXfutRuoREmVoWM/view?usp=sharing) (37 fields)
* Cough chunks (1 second long):
  + Data size: **22,878** - *.wav*
  + Folder: [coswara\_cough\_chunk](https://drive.google.com/drive/folders/1LzWWke3AcEHU21n84rlFEQDQTBQnQ5nh?usp=sharing)
  + Metadata: [coswara\_chunk\_metadata.csv](https://drive.google.com/file/d/1EllL4AqLbe7vOKCPOQjSyIfaK_ypSQWS/view?usp=sharing) (39 fields)

1. COUGHVID

* Data size: **20,072** audios - *.webm, .ogg*
* Folder: [coughvid\_cough\_data](https://drive.google.com/drive/folders/1QvNqdtMXKGZtmfjXlBvOOHehfy5hP9Io?usp=sharing)
* Metadata: [coughvid\_metadata.csv](https://drive.google.com/file/d/1JWARiYp4fGp-lAj-6pLxttO6v_LNbu6z/view?usp=sharing) (40 fields)
* Cough chunks (1 second long):
  + Data size: **84,981** - *.wav*
  + Folder: [coughvid\_cough\_chunk](https://drive.google.com/drive/folders/1GWX3nP-61bLuv0GLE86TdH905umaTWLH?usp=sharing)
  + Metadata: [coughvid\_chunk\_metadata.csv](https://drive.google.com/file/d/1E9j-vYRYlJ0FyZ1S6RAP1YYoygEzfosy/view?usp=sharing) (43 fields)

1. AICovidVN

* Data size: **5,247** audios -  *.wav*
* Folder: [aicovidvn\_cough\_data](https://drive.google.com/drive/folders/10q-_WPOaCoribs9jylO763RKR95jfdc5?usp=sharing)
* Metadata: [aicovidvn\_metadata.csv](https://drive.google.com/file/d/1DRKwxVh7MG5oRmdljIbUV3w0497sGnKp/view?usp=sharing) (5 fields)
* Cough chunks (1 second long):
  + Data size: **31,542** - *.wav*
  + Folder: [aicovidvn\_cough\_chunk](https://drive.google.com/drive/folders/1eO2cAc2JVoIErvwiwan8t-pdtkFVT8Ff?usp=sharing)
  + Metadata: [aicovidvn\_chunk\_metadata.csv](https://drive.google.com/file/d/1BlBnNOVghOozJVSRhYZDbabzd9B5gg8V/view?usp=sharing) (7 fields)

1. ESC-50

* Data size: **2,000** audios - .wav
* Folder: [ESC-50](https://drive.google.com/drive/folders/18XLkfTAbhlD_C1S_AABIcV_zid7ZIRZx?usp=sharing)
* Metadata: [esc50.csv](https://drive.google.com/file/d/1EVPrRRjoNYu3V9bqiGninFhnnnyjnm3i/view?usp=sharing)

# System components (For information on each component and relevant references: [reference](https://docs.google.com/document/d/1ipoKlPG8b_EZkLOTpCvofomkEzV2k8jZpfW1xUbAOa4/edit?usp=sharing))

1. Cough detection (detect whether the audio has a cough or not)

* **Metric**: **ACCURACY** (look into metrics used from previous research)

1. Covid19 recognition (recognize cough audio is whether covid or not)

* **Metric**: **ROC-AUC**, others (if needed) in order to compare previous results.

# Important folders

1. [Workspace/data](https://drive.google.com/drive/folders/1tzad41Nnt41NDKQsPiizV0J3Q05sAwLv?usp=sharing) - here lies all raw data
2. [Workspace/data preprocessing](https://drive.google.com/drive/folders/1MZQxZOBvzjWvXga4Q1Jcw0u5euH6H4nB?usp=sharing) - here lies all preprocessed data and its codes
3. [Workspace/detection](https://drive.google.com/drive/folders/1P8NCVIXo3optlvbOueYBQmZuGJGNrMXE?usp=sharing) - for cough detection model development
4. [Workspace/recognition](https://drive.google.com/drive/folders/1Tt9317JWJ6jne656f5fd6hcmHyftHIkP?usp=sharing) - for covid recognition model development

# What will we do?

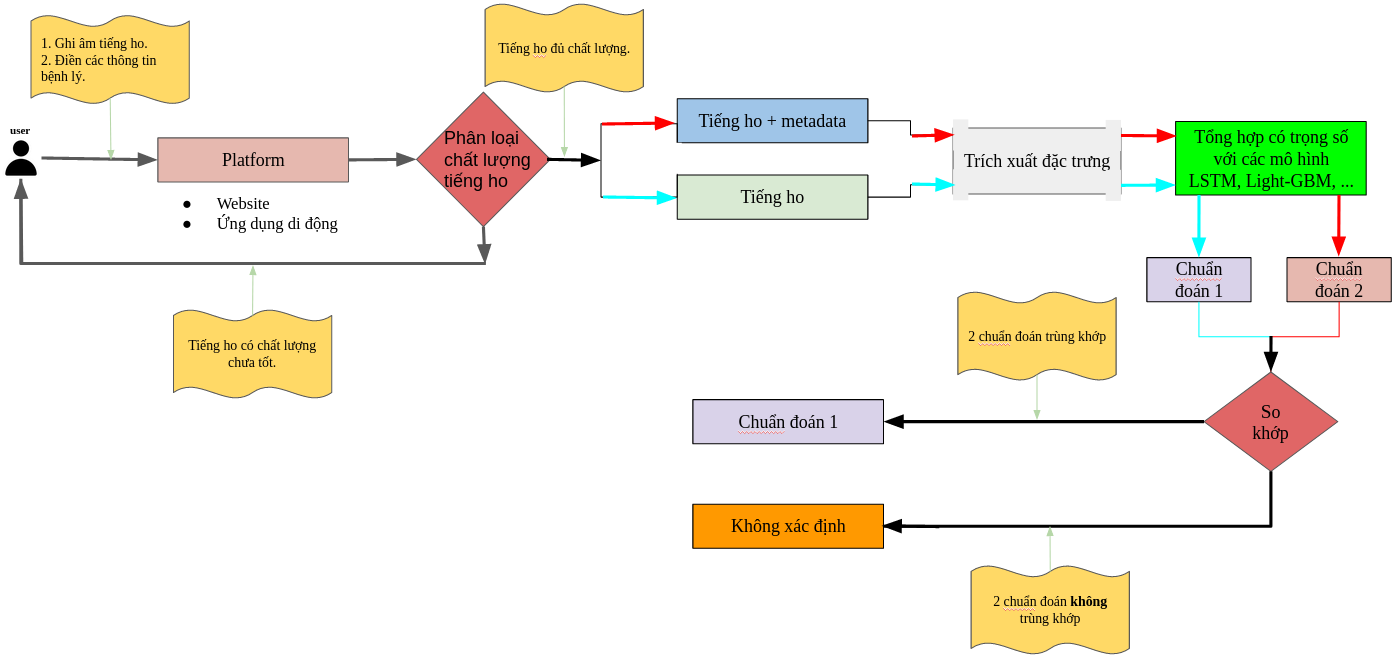
1. For cough detection:
   1. Develop a model for detecting cough:

* Audio sound from ESC-50: which kinds of sound should be used to train the model?
* Reference from the previous paper
* We train cough detector model for each Covid dataset (AICovidVN, Coswara, Coughvid) - **3 cough detector models:**

1. For covid recognition:
   1. Data reinforcement with GAN (1)
   2. Apply different audio segmentation **vs take raw audio**:

* slice by silence -> cough chunks (**Done**)
  + could concatenate chunks
  + apply different normalization or other methods to improve the quality of cough chunk
* …
  1. (if having time) Apply audio augmentation
  2. Rebuild model from previous research to do the comparison
  3. Normalize cough audio to the same sample rate (44k) (backlog) **vs remaining original (in progress)**
  4. Feature extraction: Mel-spectrogram, mfcc 13, 26, 39, zero-crossing rate, etc. (in progress): on chunks and raw audio
* 2-d features -> for cnn-based model
  + 2-d
  + 2-d + metadata
* 1-d features -> ml-based model
  1. **Ablation experiment:**
* (1) vs using the default number of audios
* Compare between different audio segmentation methods and segmentation vs raw audio
* Metadata + audio vs audio
* (if having time) Audio augmentation vs (1) -> to conclude: augmentation for cough sound is not good

# System architecture



# Feature extraction

## Before use, remove extracted feature whose labels is NULL

## Cough chunks

* 1-d features:
* Features are saved in .csv format with corresponding labels.
  + Feb 22, 2022 currently, we have features:
    - Melspectrogram: AICovidVN, Coswara, Coughvid
    - MFCC 13, 26, 39: AICovidVN, Coswara, Coughvid
    - Folder: [chunks](https://drive.google.com/drive/folders/1ITZdQ16kXEtgaj_K50hElsfQls7HUfLQ?usp=sharing)
* 2-d features:
* Features are saved in .npy format
* Due to a large number of chunks, a dataset is split into different batches. So, carefully look at the filenames to concatenate them in the right order.
* There are no labels in the *.npy file* so it is necessary to take the label from chunk metadata referenced in [Data](#_nuzo9yz1g5ib) when training.
  + Feb 22, 2022 currently, we have extracted features:
    - Melspectrogram: AICovidVN, Coswara, Coughvid
    - MFCC 13, 26, 39: Coswara, Coughvid
    - Folder: [chunks](https://drive.google.com/drive/folders/1ypYnT2NtXuszV75wtF8WDvbatC_pt4N2?usp=sharing)

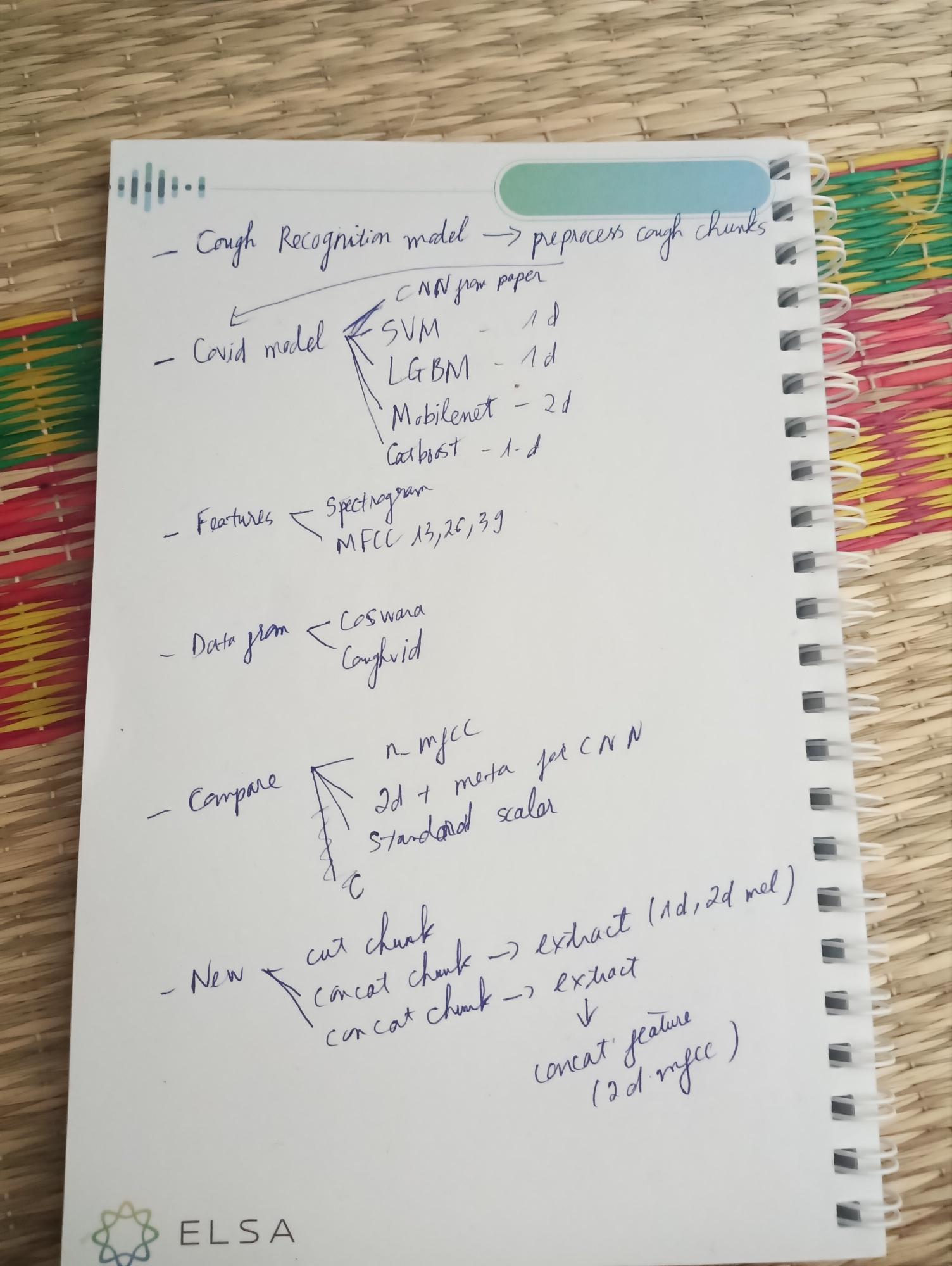
## Cough audio (in progress): feature extraction scenarios are alike chunks but apply for full raw audio.

* Some audio cannot be extracted so there is a metadata file like the original one but includes a field of *‘error’* (0: keep, 1: drop) used for filtering audio when matching labels for extracted features.
* 1-d features:
  + Mar 7, 2022: : currently, we have extracted features:
    - Melspectrogram: AICovidVN, Coswara, Coughvid
    - MFCC 13, 26, 39: AICovidVN, Coswara, Coughvid
    - Folder: [full\_audio](https://drive.google.com/drive/folders/1fi-A3L64Q3oWMk1WaP34LdYRCs00wKb8?usp=sharing)
* 2-d features:
  + Feb 26, 2022: currently, we have extracted features:
    - Melspectrogram: Coswara, Coughvid
    - MFCC 13, 26, 39: Coswara, Coughvid
    - Folder: [full\_audio](https://drive.google.com/drive/folders/1gA7KoR10WzGku4Oxv5LbcobCoZrZRbLg?usp=sharing)

# Roadmap

Feb 28, 2022: complete some models for cough detection and covid recognition and gan model

## Phase 1: paper for cough recognition and covid detection



* Data: Coswara, Coughvid
* Metrics:
  + Cough detection: Accuracy
  + Covid recognition: ROC-AUC
* Model:
  + CNN-based model: MobileNet, VGG, etc.
  + ML-based model: Light-GBM, Catboost, SVM, Random Forest, etc.
  + Implement models from previous research: [Covid](https://drive.google.com/drive/folders/1iqq0aT6v7DFlk5yw09XdbI0V6HNmt8rA?usp=sharing)
* New experiments in:
  + Use cough chunks
  + Feature extraction pipeline
* Comparison and Ablation:
  + cough chunk vs cough audio
  + Standard scaler, PCA, .etc.
  + # MFCC coefficients: 13, 26, 39
  + Audio vs audio + metadata

**Overleaf project:** [**https://www.overleaf.com/9319278816nskgdhhhbdjs**](https://www.overleaf.com/9319278816nskgdhhhbdjs)

## Phase 2: paper for cough recognition and covid detection + GAN (update later)

* Main goal:
  + use GAN to enhance the number of audios, as well as covid recognition
  + Demo system
* Data: AICovidVN, Coswara, Coughvid
* Model:
  + Models from AICovidVN challenge
  + etc.

# Notes

Working timeline: [timelines](https://docs.google.com/document/d/13zSdm8h8xHItqclRVSUOepmXrrMBvlZhTmN457xYMPs/edit?usp=sharing)