

## 数据集

- Dataset的收集方式：
  - The expert annotators reviewed 50 second long EEG samples plus matched spectrograms covering 10 a minute window centered at the same time and labeled the central 10 seconds.
- Files
  - data里一共两种文件csv和parquet
  - data里面的csv文件是metadata.
  - raw data是spectrograms和eegs文件夹里面的parquet文件. metadata决定了怎么构造samples
- visualize raw data的可以参考笔记本
  - <https://www.kaggle.com/code/mikhailzemskov/understanding-stft-and-spectrograms>
  - <https://www.kaggle.com/code/cdeotte/how-to-make-spectrogram-from-eeeg/notebook>
- dataset 构造可以用
- HMS-HBAC: KerasCV Starter Notebook (这个notebook主要是extract samples的代码)
  - <https://www.kaggle.com/code/awsaf49/hms-hbac-kerascv-starter-notebook/notebook>

### spectrograms data:

每个parquet文件都以唯一的spectrogram id命名。

dataframe 的columns 一共有401个，第一个col是time，其余column代表探头？每个探头分别有10个columns。见图。

dataframe (df\_spec) 长度为 320 (注意time column 全是奇数 最大的time stamps是639)

```
df_spectrogram = pd.read_parquet(SPECTROGRAM_SAMPLE_FILE)
df_spectrogram.head(10)
```

:

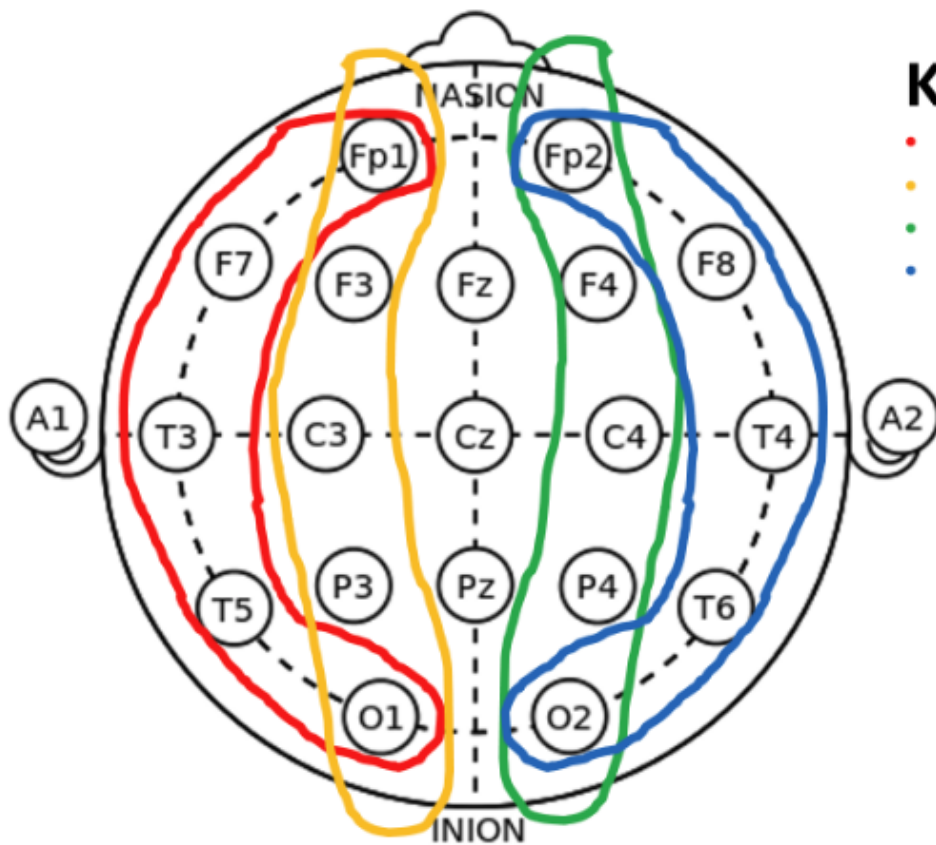
	time	LL_0.59	LL_0.78	LL_0.98	LL_1.17	LL_1.37	LL_1.56	LL_1.76	LL_1.95	LL_2.15	...	RP_1
0	1	4.26	10.98	9.05	13.65	11.49	8.930000	18.840000	19.26	19.240000	...	0.31
1	3	2.65	3.97	12.18	13.26	14.21	13.230000	9.650000	8.11	11.280000	...	0.15
2	5	4.18	4.53	8.77	14.26	13.36	16.559999	19.219999	17.51	22.650000	...	0.29
3	7	2.41	3.21	4.92	8.07	5.97	12.420000	10.820000	14.96	21.809999	...	0.33
4	9	2.29	2.44	2.77	4.62	5.39	7.080000	9.840000	12.27	14.410000	...	0.44

### eeg data:

每个parquet文件都以唯一的eeg id命名。

dataframe 的columns 除了没有time column，其余和spectrograms类似，见表格。这里我只能大胆猜测electro

node是一一对应的。。。



## KEY

- LL - Left Temporal Chain
- LP - Left Parasagittal Chain
- RP - Right Parasagittal Chain
- RR - Right Temporal Chain

dataframe (df\_egg) 长度为18000

Q: 因为eeg和spectrogram是matched的, 但是time stamps数目不一样。time stamps是不同的单位吗? 分别是什么单位?

```
df_eeg = pd.read_parquet(EEG_SAMPLE_FILE)
df_eeg.head(10)
```

:

	Fp1	F3	C3	P3	F7	T3	T5	O1
0	-80.519997	-70.540001	-80.110001	-108.750000	-120.330002	-88.620003	-101.750000	-104.489999
1	-80.449997	-70.330002	-81.760002	-107.669998	-120.769997	-90.820000	-104.260002	-99.730000
2	-80.209999	-75.870003	-82.050003	-106.010002	-117.500000	-87.489998	-99.589996	-96.820000
3	-84.709999	-75.339996	-87.480003	-108.970001	-121.410004	-94.750000	-105.370003	-100.279999
4	-90.570000	-80.790001	-93.000000	-113.870003	-129.960007	-102.860001	-118.599998	-101.099999

metadata: 来自train.csv 基本上规定了sample的构造 (见下一节) <input, label> pair

```
df_metadata = pd.read_csv(METADATA_PATH)
df_metadata.head(10)
```

	eeg_id	eeg_sub_id	eeg_label_offset_seconds	spectrogram_id	spectrogram_sub_id	spectrogram_label_c
0	1628180742	0	0.0	353733	0	0.0
1	1628180742	1	6.0	353733	1	6.0
2	1628180742	2	8.0	353733	2	8.0
3	1628180742	3	18.0	353733	3	18.0
4	1628180742	4	24.0	353733	4	24.0

# Dataset的构造 <input,label> pair

(个人理解)

- 在train.csv文件里面记录了每一个sample和对应的label。每个sample对应着一个spectrogram id, eeg id, eeg\_sub\_id(spectrogram\_sub\_id), eeg\_label\_offset\_seconds(spectrogram\_label\_offset\_seconds), label\_id, votes...等等
- 假设df\_spec\_id1是对应某个spectrogram\_id的数据, 那么对应每一个offset, 我们能从这个dataframe中提取一个sample `df_spec_id1[:, offset : offset+320]`。
- 一共能提取的samples是 `df_spec_id1[:, offset : offset+320] for offset in metadata[metadata.spectrogram_id == id1]['spectrogram_label_offset_seconds']`
- eeg的sample input构造同理。且与每一个spectrogram sample input对应。这个对应是有csv里面的对应确定的。
- Question: metadata里面, 对应同一个sample, eeg\_label\_offset\_seconds 和 spectrogram\_label\_offset\_seconds 的值有时候是相同的 有时候是不同的。。。我不是太理解。。。而且看eeg和spectrogram的长度是不同的, 鉴于它们都是同一个时间段matched的pair 所以我推测对应的时间单位是不一样的。如果时间单位不一样, 我的理解是eeg和spectrogram的offset应该是成比例才对。。。不是吗。。。?

## notes

大致看了一下各种notebooks。一部分在折腾features, 譬如eeg那几列数据的交互作用; 一部分在做各种信号处理 (signal processing), 包括去噪 各种傅立叶变换 小波变换。一部分在搞CV model。。。

-----03-13

<https://zhuanlan.zhihu.com/p/147561580>

<https://zhuanlan.zhihu.com/p/549254941>

[https://blog.csdn.net/youzi12345678/article/details/89321276?](https://blog.csdn.net/youzi12345678/article/details/89321276?spm=1001.2101.3001.6650.2&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=5)

[spm=1001.2101.3001.6650.2&utm\\_medium=distribute.pc\\_relevant.none-task-blog-](https://blog.csdn.net/youzi12345678/article/details/89321276?spm=1001.2101.3001.6650.2&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=5)

[2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&depth\\_1-](https://blog.csdn.net/youzi12345678/article/details/89321276?spm=1001.2101.3001.6650.2&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=5)

[utm\\_source=distribute.pc\\_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-](https://blog.csdn.net/youzi12345678/article/details/89321276?spm=1001.2101.3001.6650.2&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=5)

[85716900.235%5Ev43%5Econtrol&utm\\_relevant\\_index=5](https://blog.csdn.net/youzi12345678/article/details/89321276?spm=1001.2101.3001.6650.2&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7ECTRLIST%7ERate-2-89321276-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=5)

[https://blog.csdn.net/weixin\\_44086494/article/details/105881174?](https://blog.csdn.net/weixin_44086494/article/details/105881174?spm=1001.2101.3001.6650.6&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=10)

[spm=1001.2101.3001.6650.6&utm\\_medium=distribute.pc\\_relevant.none-task-blog-](https://blog.csdn.net/weixin_44086494/article/details/105881174?spm=1001.2101.3001.6650.6&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=10)

[2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&depth\\_1-](https://blog.csdn.net/weixin_44086494/article/details/105881174?spm=1001.2101.3001.6650.6&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=10)

[utm\\_source=distribute.pc\\_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-](https://blog.csdn.net/weixin_44086494/article/details/105881174?spm=1001.2101.3001.6650.6&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=10)

[105881174-blog-85716900.235%5Ev43%5Econtrol&utm\\_relevant\\_index=10](https://blog.csdn.net/weixin_44086494/article/details/105881174?spm=1001.2101.3001.6650.6&utm_medium=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&depth_1-utm_source=distribute.pc_relevant.none-task-blog-2%7Edefault%7EBlogCommendFromBaidu%7ERate-6-105881174-blog-85716900.235%5Ev43%5Econtrol&utm_relevant_index=10)