选题一　癫痫间期颅内脑电异常活动分析

郝千越 2018011153 无 85

目录

0 说明

本人所学专业为电子信息工程而非医学、生物，课程学习中、作业完成中听老师、助教讲解，查阅资料、文献，对相关专业背景略有一二了解，其受益匪浅。然而所欠缺与所待补习之处仍数不胜数，本作业中结合所给参考文档、例程及相关资料，尽力了解与应用相关知识。但受时间、精力所限，多有疏漏与不足，如本文内容中涉及相关专业知识有所不当与谬误，恳请原谅与指正。

1 数据预处理与观察

本课程设计的背景为癫痫间期颅内脑电异常活动分析，癫痫患者在发作间期的颅内脑电信号存在一定的异常表现，通过分析这些异常脑电信号的时间、位置分布，可以为患者的治疗提供重要的指导。课程设计提供的数据包括两位患者颅内脑电信号，其数据形式描述如下：

表 1 数据形式

患者编号	记录时长（h）	电极数目	采样率（Hz）	异常记录数目
S1	24	86	1000	1330070
S2	24	126	1000	813920
总计	——	——	——	2143990

原始数据文件中，两位患者的记录各自以 2h 为划分组织为 12 个文件，为方便后续处理，将每位患者的所有记录合并为一个.npz 文件，相关代码为 merge.py。分别可视化两位患者异常记录中的 15 条如下：

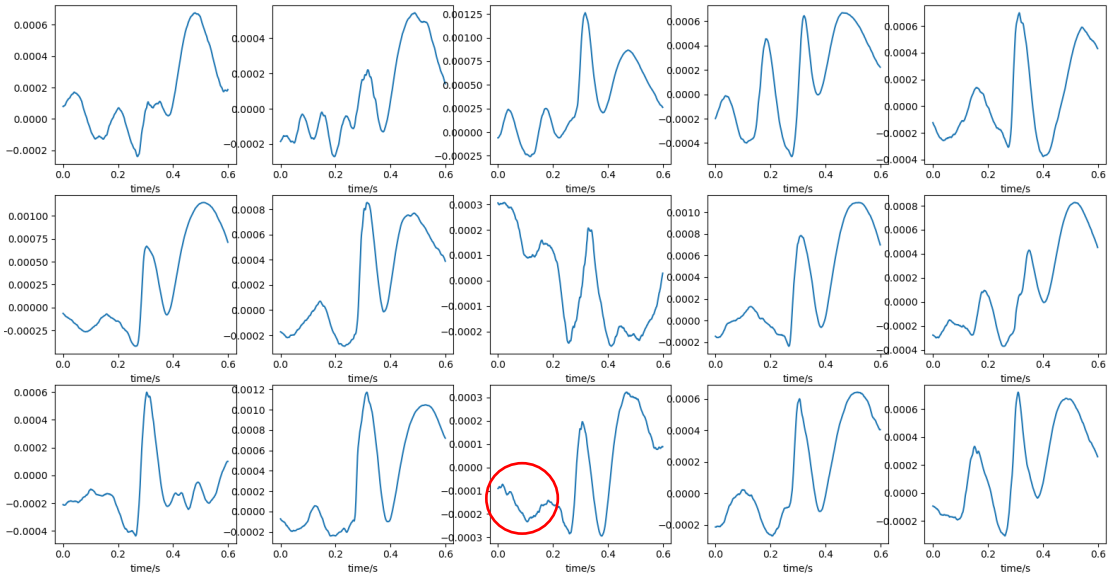

图 1 患者 1 部分异常记录波形

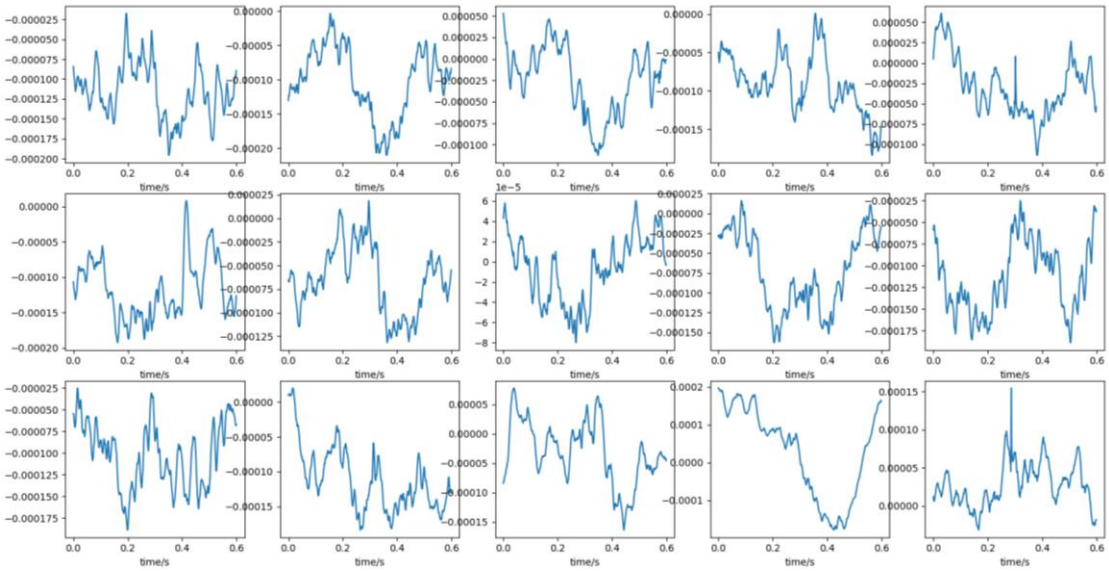

图 2 患者 2 部分异常记录波形

直接观察时域波形发现，两位患者的颅内脑电波形差异较大，S1 患者呈现较为平滑的曲线，而 S2 患者呈现明显叠加了高频信号的曲线。但经过仔细观察，S1 患者的平滑曲线上亦叠加了较为明显的高频率抖动，例如图 1 中红圈标注位置所示，同时观察到 S1 患者的波形幅度普遍比 S2 大，由此推测两者呈现不同形态的原因如下：脑电波形由频率较低、幅度较大的低频信号和频率较高、幅度较小的高频信号叠加而成。两位患者中，由于某些原因 S1 探测到的低频信号幅度大于 S2，而两患者叠加的高频信号幅度相近。因此，S2 患者的高频信号被凸显出来，导致波形的直观形态不同。结合滤波和频域分析可以更清晰的证明该解释：

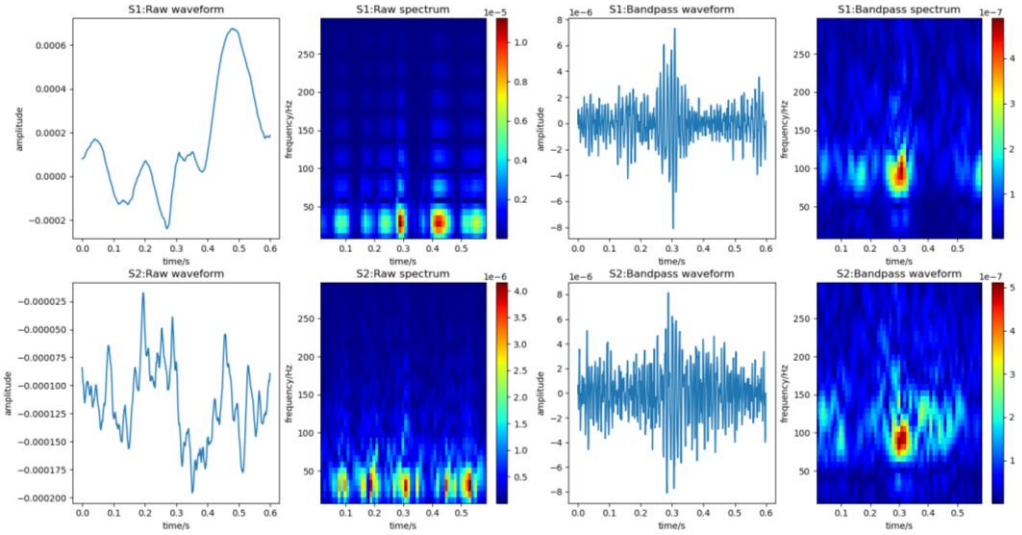

图 3 患者脑电频谱分析

上图分别从 S1、S2 中取出一条典型记录进行短时傅里叶变换（STFT）频谱分析。左侧四图是原始波形和直接进行频谱分析结果[1]。可见两患者信号的能量[2]主要集中在 50Hz 以下，即前述的“幅度较大”的低频信号，同时注意到两者的强度存在 2-3 倍的差距。

[1] 此处频谱分析使用长度 30pixels 的 Tukey-Hanning 时间窗，窗移为 10pixels，傅里叶系数取 128 个。由于所取时频域精度均较高，因此不使用 demo 中的高斯滤波器，可以直接得到呈现岛状分布的频谱图。

[2] 实际上图中给出的为短时傅里叶变换系数，真正的功率谱密度应为信号自相关函数而非信号本身的短时傅里叶变换系数。但两者均有能量、幅度含义，本文中不做严格区分。

右侧四图是用通带 3dB 点为 80Hz 的 10 阶数字 Butterworth 高通滤波器（该 IIR 滤波器传递函数可以由冲激不变法给出）对波形进行滤波后再进行频谱分析。可以看到滤去低频信号后，两者剩余部分为 80Hz 以上的高频信号。两者高频信号幅度基本一致，由此证实了上述对于时域波形形态差异原因的分析。

同时注意到，这一小幅度的高频信号相比于低频信号强度低 1-2 个数量级，其能量分布在 80Hz-250Hz 之间，正是我们要寻找的“颅内脑电异常活动”。将幅度做对数压扩后再画出未经滤波的频谱图，可以清晰地看到高频信号与低频信号的强度关系。

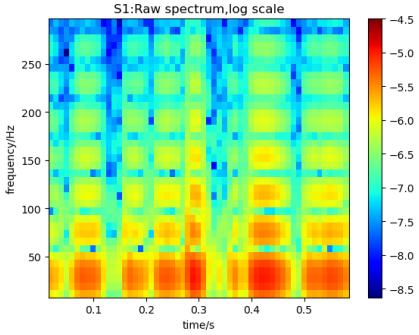

图 4 患者脑电频谱图（对数压扩）

2 任务一：无监督分类

2.1 问题分析

下面尝试利用无监督分类的办法对 Spike 和 HFO 进行区分。结合上面的讨论，经过带通滤波、短时傅里叶变换变换后可以得到位于 80Hz 以上的高频颅内脑电异常活动信号的特征。下面取 S2 患者的前 12 个样本，用与上述相同的滤波器和 STFT 设置，其频谱图如下。此后为除去频谱绝对幅度的影响从而便于将频谱归类，将频谱做最大值为 1 的线性归一化。同时为便于观察更广的频谱特性，此处频谱展示范围扩大到 500Hz。

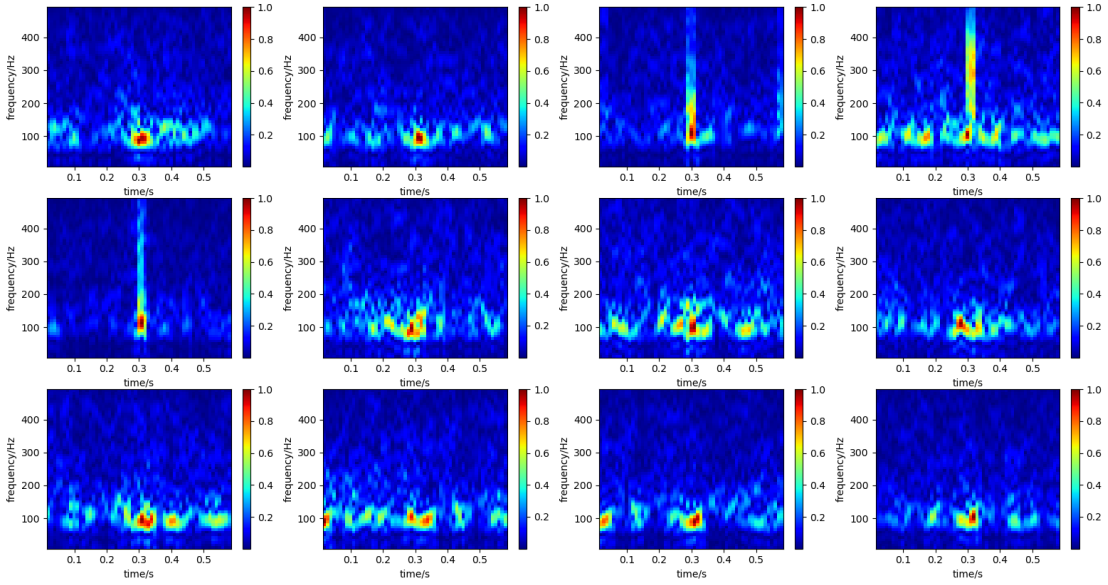

图 5 S2 患者的前 12 个样本归一化频谱图

可以明显看出高频活动可以分成两类。一类活动能量集中在 150Hz-200Hz 以下，另一类活动除 150Hz-200Hz 以下的能量外还有延伸至 200Hz 以上，甚至可以达到近 500Hz。上图中第 2、3、4 幅图就是第二类的典型代表。根据任务说明中对于 Spike 和 HFO 的说明，

第一类可能对应的是 Spike，而第二类对应的则是 HFO。下面使用无监督聚类的方法区分这两类高频活动。

2.2 手工降维，K-means 聚类

上面得到的频谱图尽管能够看出明显的类别差异，但由于频谱图为规模$T \times F$的二维张量S_{tf}，其特征过多，因此需要进行降维操作。此处首先使用观察样本特点并进行手工降维操作的方法，观察发现，沿时间轴进行特征降维能够包络频谱图的特征差异，即应用函数$f(\vec{x}), \mathbb{R}^T \to \mathbb{R}^1$，得到降维后的$F$维向量特征$S_f^*$如下：

$$S_f^* = f\left([S_{1f}, S_{2f} \dots S_{Tf}]\right) \triangleq f(S_f)$$

尝试应用两种降维操作，$f_1(\vec{x}) = mean(\vec{x})$即取某频点所有时刻平均强度；$f_2(\vec{x}) = max(\vec{x})$即取某频点所有时刻最大强度。

使用$f_1(\vec{x}) = mean(\vec{x})$进行降维后聚类，观察聚类结果。下图分别为选取 S1 患者和 S2 患者少量样本，变换到频域后展示的分类结果。

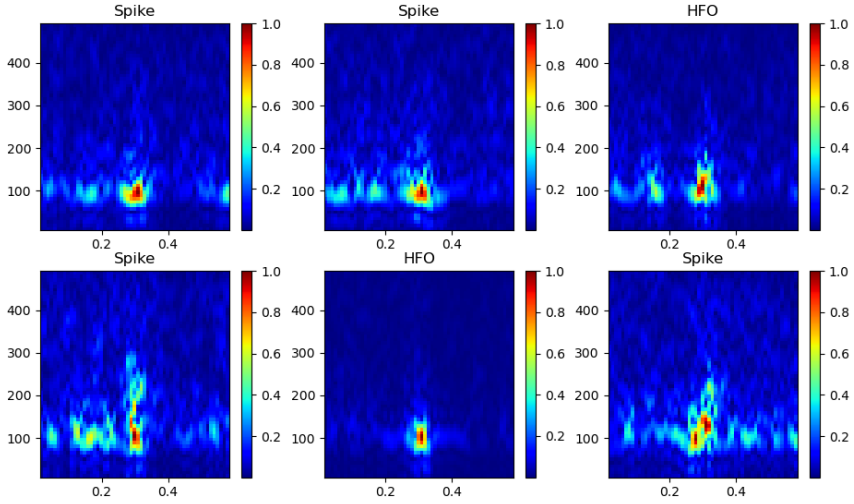

图 6 平均值降维 S1 患者部分样本聚类结果

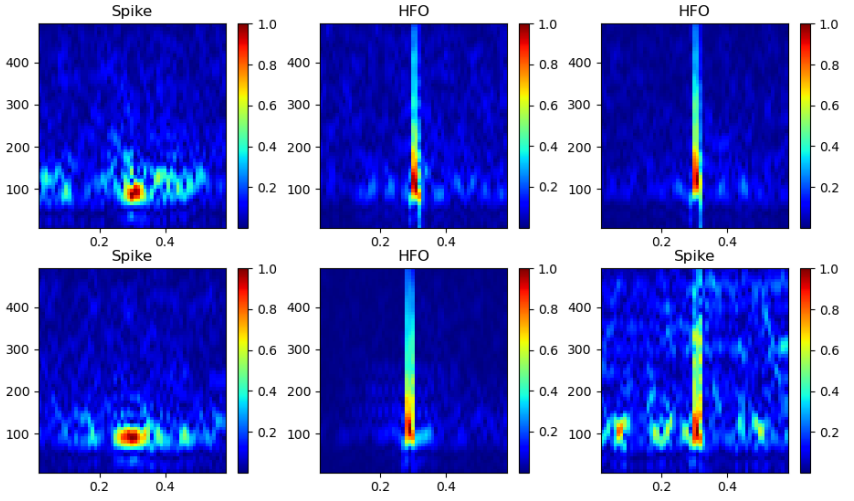

图 7 平均值 S2 患者部分样本聚类结果

可以看到聚类结果并不理想。许多具有明显特征的 HFO 活动被聚类为 Spike，同时也有一些目测应当属于 Spike 的样本被分为 HFO。再使用$f_2(\vec{x}) = max(\vec{x})$进行降维后聚类，观察聚类结果。下图选取 S1 患者和 S2 患者相同样本，变换到频域后展示的分类结果。

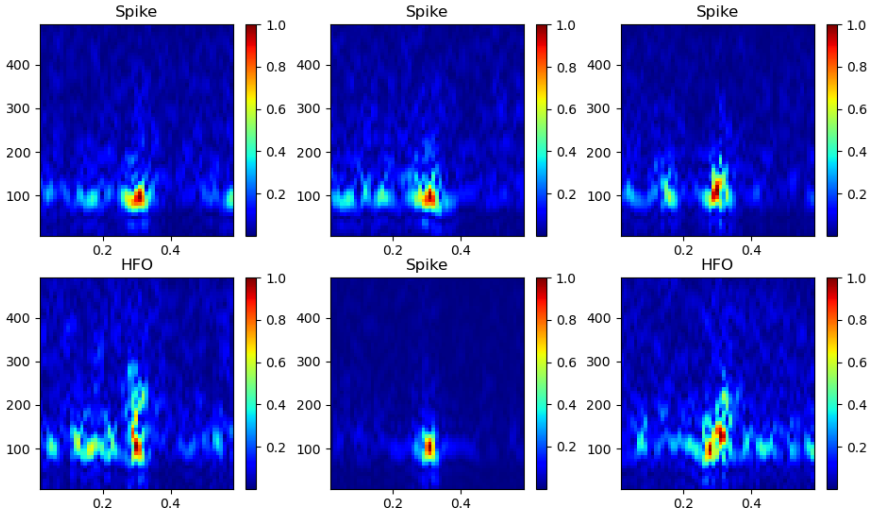

图 8 最大值降维 S1 患者部分样本聚类结果

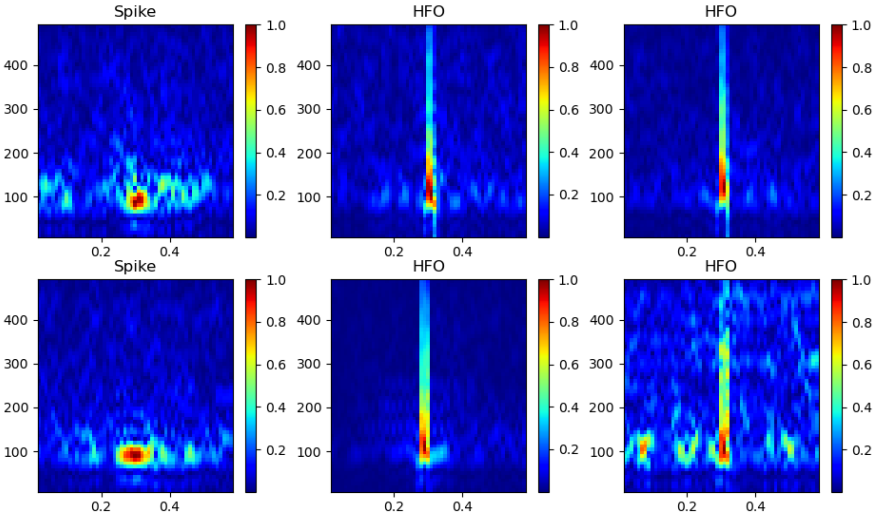

图 9 最大值降维 S2 患者部分样本聚类结果

可以看到使用最大值进行特征降维后，聚类效果良好，图示样本中具有高频活动特性的被正确分为 HFO，而其他被正确分为 Spike。随机选取一些此处未画出的样本进行验证，绝大部分被正确分类。

对于两种降维方式效果差异大致分析如下：由于样本采集、裁剪等原因，不可避免地存在一定的噪声。如果使用平均值，则全部频谱都会对最后结果产生影响，导致分类效果不佳。由于决定异常活动类型的主要是峰值处的频谱，因此使用最大值降维可以忽略噪声影响而抓住本质因素，获得较好的分类效果。目前深度学习常用模型中，多使用最大池化而非平均池化，一定意义上也考虑了这种原因，即最大池化可以忽略小扰动而抓住主要特征。

2.3 手工降维，PCA，K-means 聚类

下面在 2.1 中表现较好的最大值降维方法得到的 63 维频域特征向量S_f^*做进一步降维处理。主成分分析法（PCA）是一种有效的特征降维方法，能够将高维样本投影在方差尽可能大的低维空间上，即降低了样本的维度而便于后续处理，又保留了便于区分的特征。

同时比较不同的数据降维力度对聚类性能的影响。用无监督聚类性能评价指标 Calinski and Harabasz score 来评价聚类效果，下文简称为 CH score。CH score 定义如下：

$$CH = \frac{tr(B_k)/(k-1)}{tr(W_k)/(N-k)}$$

其中k为类别数目，N为样本总数，B_k表示类间散度矩阵，其计算为

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

其中c_q是第q类样本的质心，n_q是第q类样本的数目，c_E是所有样本的质心，W_k表示类内散度矩阵，其计算为

$$W_k = \sum_{q=1}^{k} \sum_{x_q} (x_q - c_q)(x_q - c_q)^T$$

容易看出，CH score 表示类间差异程度和类内相似程度的比值，其值越大代表聚类效果越好。下面以计算代价、CH score 两个指标分析降维至不同维度时的分类性能，结果如下[3]：

表 2 不同 PCA 降维力度的比较

降维后特征维度	PCA 计算耗时（s）	Kmeans 计算耗时（s）	CH score
60	39.26	23.12	1197971
50	39.01	21.05	1197971
40	33.31	19.69	1197971
30	23.39	18.70	1197970
20	19.85	16.22	1197969
10	12.38	15.96	1197967

可以看到，随着降维后样本维度的降低，计算样本投影过程中矩阵规模减小，因此 PCA 降维的计算耗时减小；同时随着降维后样本维度下降，Kmeans 迭代计算过程中消耗的时间也减小。观察 CH score 发现，随着样本维度的下降，聚类效果几乎不变。这正是由于 PCA 的特性保留了样本方差最大的维度，同时原样本也具有区分性较为明显的特点。因此降维后剩余的维度仍然能有效地区分不同样本，故类间散度和类内散度的比值基本不变。

下图展示了选取 S2 患者的少量样本，提取频域特征并降维至 10 维后进行聚类的分类效果。可以看出，将数据维度压缩至原数据的 1/6 以下时，聚类结果仍是准确的，具有岛状高频活动特性的 HFO 被区分开来。

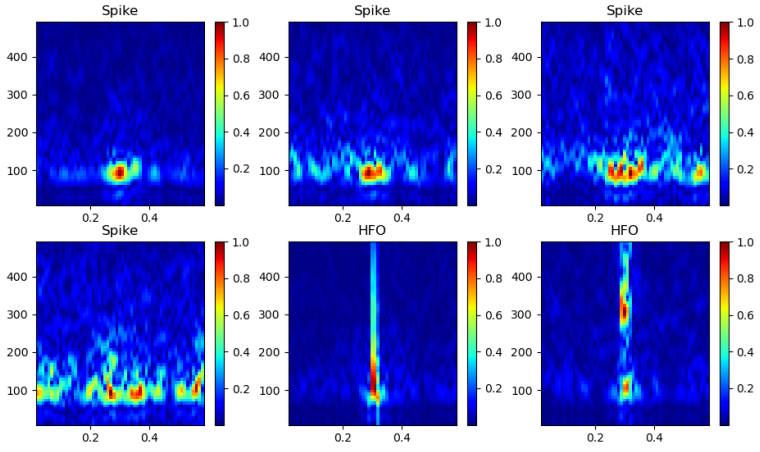

图 10 PCA 降至 10 维后 S2 患者部分样本聚类结果

更极端地，将样本压缩至 3 维，使之可以在几何空间中被可视化。此时 PCA 计算消耗 11.95s，Kmeans 计算消耗 11.23s，均小于较高维度时的情形。同时，CH score 为 1197936，相比于较高维度有相对明显的下降，但仍为较大值，代表有效的聚类。在三维空间中可视化

[3] 测试平台硬件配置为 Intel Core i7-8750H @2.4GHz

10000 个样本点来示意降维后的聚类效果。可以看出，聚类后的点基本在空间中分布为两簇，呈现聚类特性。

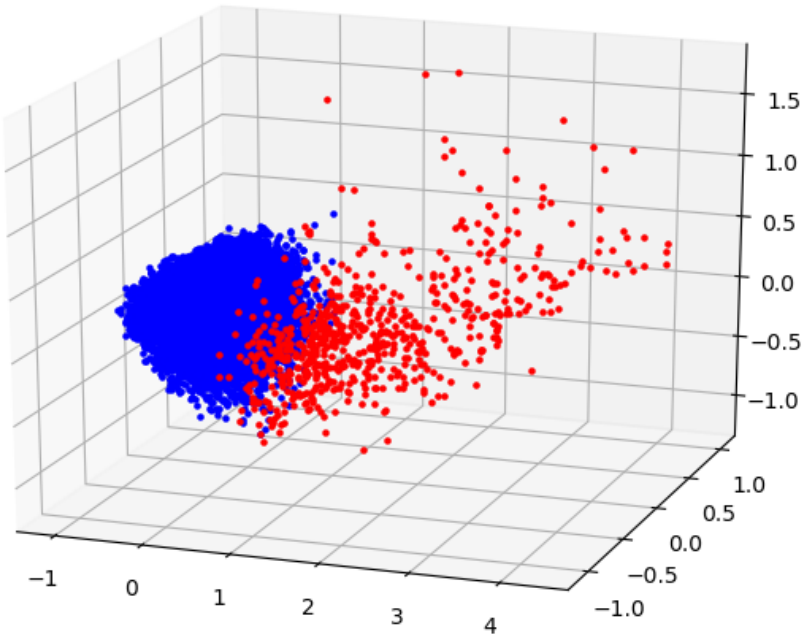

图 11 PCA 降维至 3 维后点的空间分布

2.4 典型的 Spike 和 HFO

结合上述探究过程，选取最大值降维后使用 PCA 降维至 30 维的数据继续进行后续探究。观察一些样本后，从 S1、S2 患者的样本中各找出一个较为典型的 Spike 和一个较为典型的 HFO，展示如下。其中第一行为原始时域波形，第二行为原始波形的频谱图，第三行为经过 80Hz 以上的高通滤波器后的时域波形，第四行为经过 80Hz 以上的高通滤波器后的频谱图。

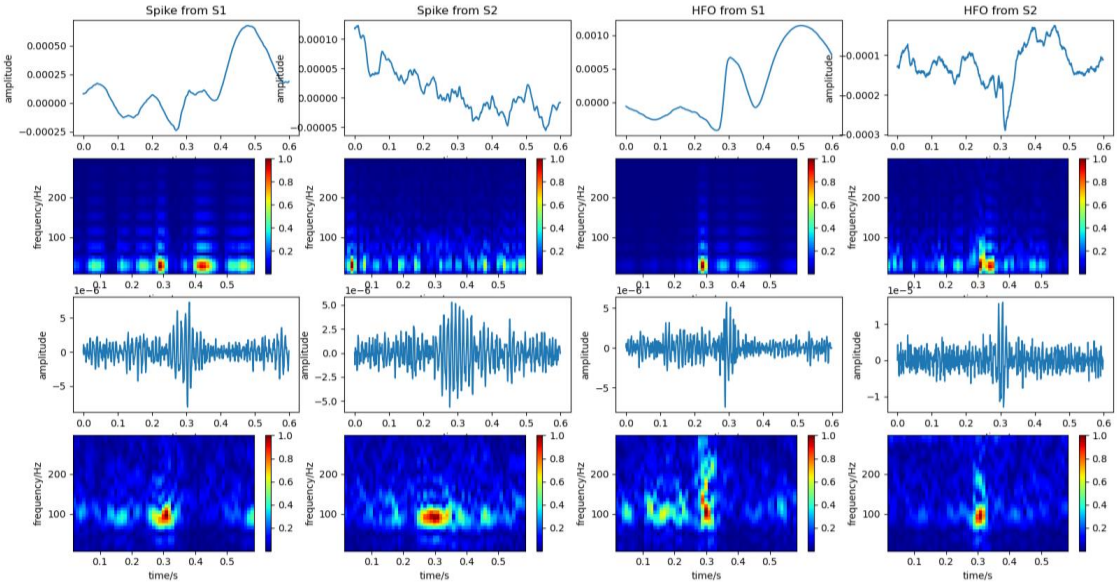

图 12 两位患者样本中的典型 Spike 和 HFO

可以发现，两种脑电活动在原始时域波形上并没有较为规律性的特征，不易于直接区分。同时，直接观察频谱时，绝大部分能量集中在 80Hz 以下，高频部分的区别被掩盖，同样不易区分两类活动。经过 80Hz 以上的高通滤波滤去低频活动后，可以看出剩余的高频活动存

在一定的直观差异：Spike 的高频振动相对更加规律而 HFO 的高频振动则更加杂乱。这是由于 HFO 中包含更多的高频分量，故时域波形呈现更加复杂的叠加。这一点在经过高通滤波后剩余的频谱上可以明显看出：Spike 的频谱集中在 150Hz 左右以下，而 HFO 的频谱存在明显的 200Hz 附近的岛状形态。这一岛状形态正对应着时域波形中杂乱的高频振动。

2.5 其他方法探究

2.5.1 DBSCAN 算法

上面的“高通滤波——频域变换——最大值降维——PCA 降维——Kmeans 聚类”的方法能够比较有效的无监督区分两类脑电活动，下面再尝试一些其他的方法进行分类。

首先使用另一种聚类方法——DBSCAN 聚类方法。在传统的 Kmeans 方法中，聚类数目需要人为给出，而这种方法能够根据类内、类间的散度自动分裂或合并聚类，从而可以自动决定聚类数目。由于 DBSCAN 算法的计算复杂度远高于 Kmeans 算法，这里仅取出 20000 个样本在变换到频域并降维后进行聚类，其聚类计算过程耗时 17.99s。聚类后，展示同图 7、图 9 中 S2 患者的部分样本上的分类效果，可见并不能准确地区分两类样本。

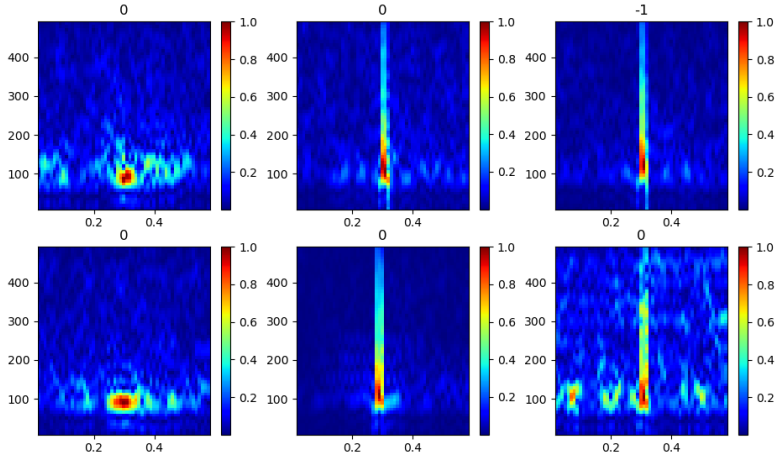

图 13 DBSCAN 算法 S2 患者部分样本聚类结果

由上述过程分析，DBSCAN 算法计算复杂度高，计算耗时长，因而难以应用与大量样本的场景，从而在较小样本上获得的效果较差。在明确已知聚类数目时，Kmeans 计算速度快，是首选的聚类算法。

2.5.2 时域波形直接聚类

上述过程均为经过滤波、变换至频域后进行特征降维与聚类。下面尝试直接从时域区分两种信号。首先将 600 个采样点的信号降采样至 100 个采样点，然后使用 Kmeans 算法直接聚类，展示同图 7、图 9 中 S2 患者的部分样本上的分类效果，可见并不能准确地区分两类样本。

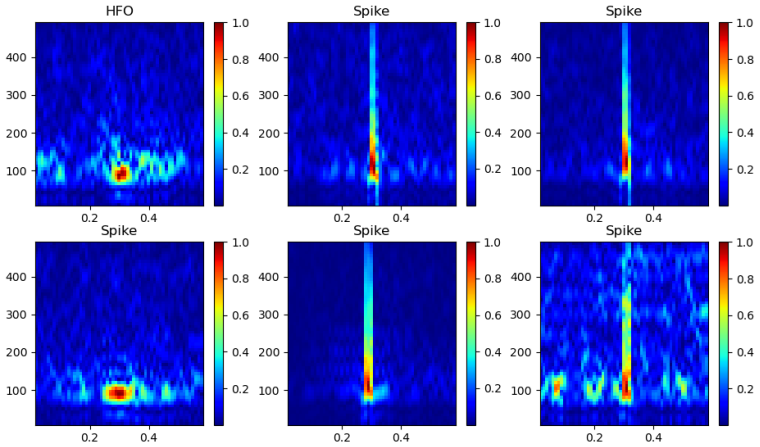

图 14 时域直接聚类后 S2 患者部分样本聚类结果

这一结果首先印证了 2.4 中观察典型的两种信号波形后得出“原始时域波形不具有较为明显的区分特征”这一结论。同时，更加数字信号处理的相关理论，采样率为 1000Hz 的信号具有最大的频率为 500Hz，时域聚类前的降采样操作极大损失了信号的高频分量，而高频分量正是区分两种活动的关键。如果不进行降采样而对 600 维的时域信号直接聚类，又存在计算复杂度可接受性上的挑战。

2.5.3 小结

上述两种探究表明，进行高通滤波、频域变换这一特征提取操作对于本问题的解决是十分关键的，直接进行时域聚类结果较差。同时选择 Kmeans 算法比较适合本问题场景，其他算法不仅计算复杂度极大，结果表现也很有限。

3 任务二：规律探索

结合上述探究过程，使用“高通滤波——频域变换——最大值降维——PCA 降维至 30 维——Kmeans 聚类”的方法能够比较有效的无监督区分两类脑电活动，从而得到样本的分类标签。该计算过程中时间开销如下：

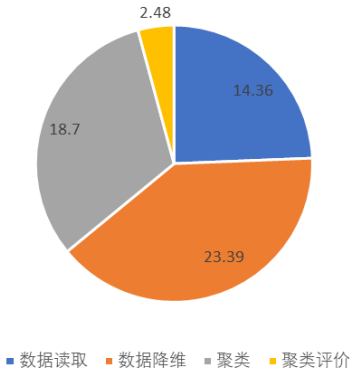

图 15 计算时间开销（s）

下面使用得到的标签进行两种脑电活动的规律探索。

3.1 异常活动的时间特点

首先研究异常脑电活动的时间分布特征。统计一天内（0-24 时）中两位患者的全部 2143990 条记录，以 1min 为统计粒度，得到一天中 Spike 和 HFO 发生次数的时间分布如下。

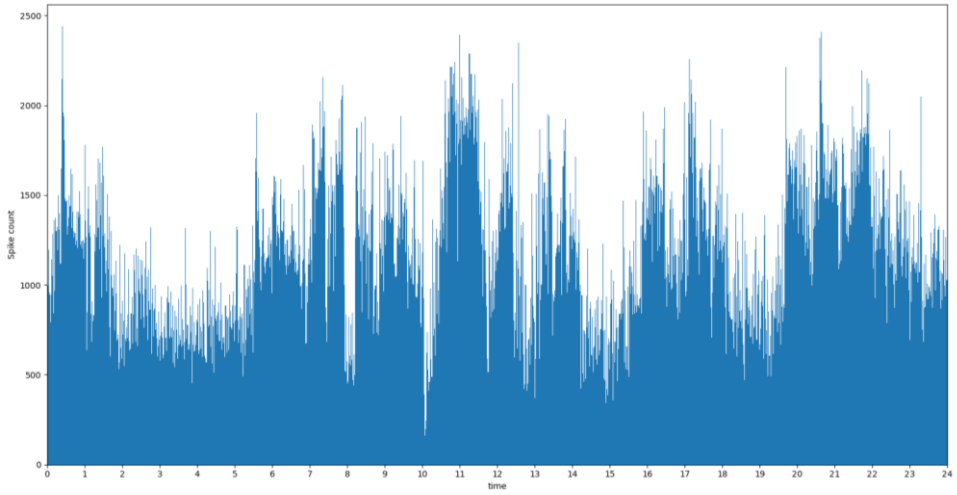

图 16 24 小时中 Spike 的数量分布

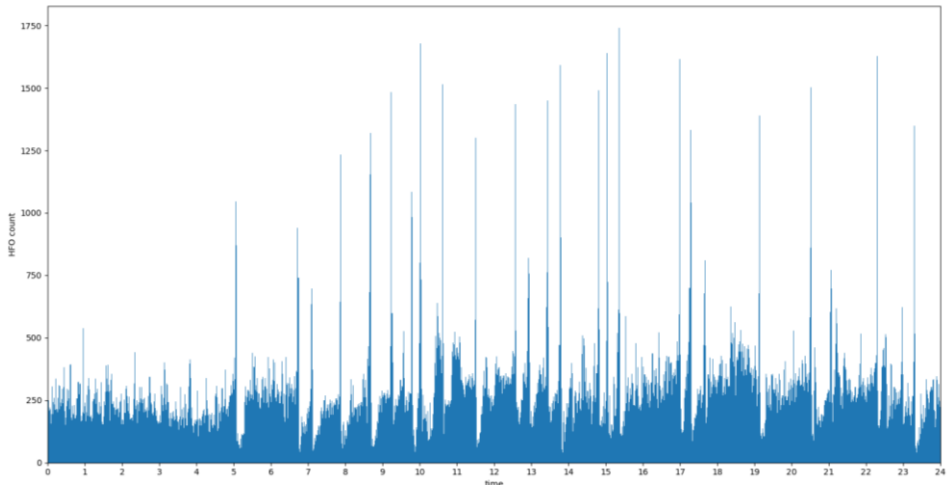

图 17 24 小时中 HFO 的数量分布

从总体数量上来看，Spike 发生的次数多于 HFO（两图纵轴单位长度不同）。观察时间分布特性可以发现，两者的发生都具有一定的周期特性。其中 Spike 的发生密度变化的周期较长，而 HFO 的变化周期较短，同时 HFO 时常出现突然增多的陡峭峰线，Spike 的变化相对比较平缓。同时观察到，在夜间（0-5 时），HFO 出现的密度相对较低且比较平稳，在日间则变化剧烈。

下面从频谱上再次观察上述结论。按照 2.2 中方法得到每一条异常脑电活动的 63 维频域特征向量。按照 1min 为粒度将发生在同一个时间片段内的频谱特征向量取平均值，分类统计所有 2143990 条记录，将结果归一化后排列在时间轴上画出，结果如下。

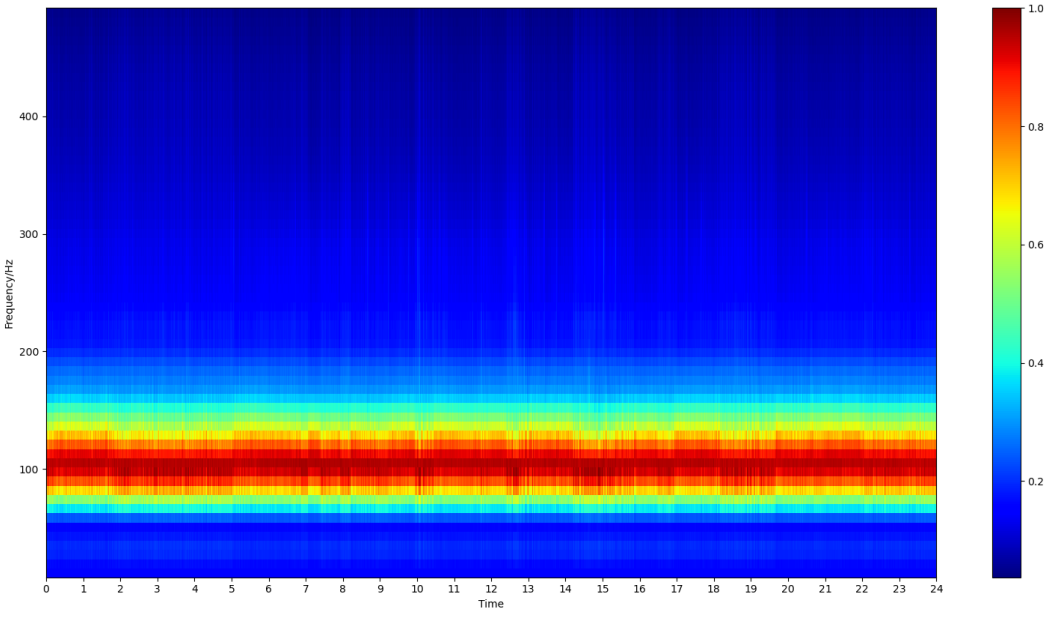

图 18 24 小时中 Spike 的平均频谱分布

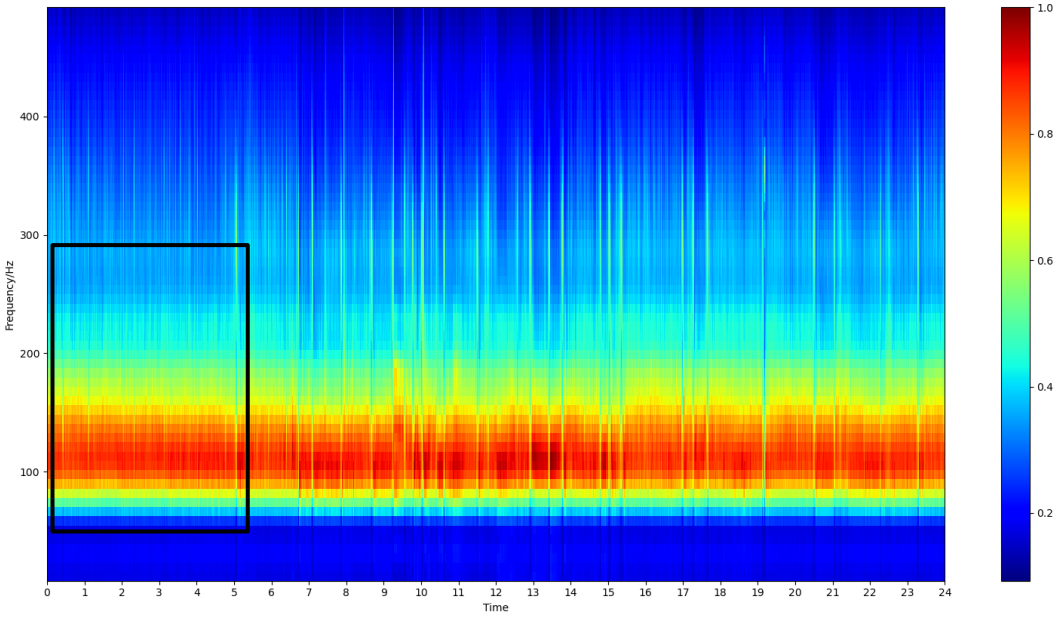

图 19 24 小时中 HFO 的平均频谱分布

由于 Spike 的频谱多分布在 150Hz 以下，故图 17 中基本没有高频区域的能量分布。同时已经从图 15 中观察到 Spike 的发生频次变化较为平缓，可在图 17 中对应观察到平均频谱随时间没有剧烈变化。

HFO 的频谱分布有 200Hz 以上的高频段，故图 18 中存在一条条插入高频区的“能量线”，这些线的分布也对应着图 16 中 HFO 时常出现突然增多的陡峭峰线。同时图 18 中黑框部分颜色相对较浅（能量较低），同时尖锐的高频峰线不突出，对应着图 16 中夜间（0-5 时）HFO 出现的密度相对较低且比较平稳。

3.2 异常活动的空间特点

所给出的颅内脑电数据来自多个不同位置的电极，下面分析这些电极采集的数据之间是否存在空间分布的差异性。以 S2 患者为例，统计全部 126 个电极上 813920 数据的分布，结果如下。

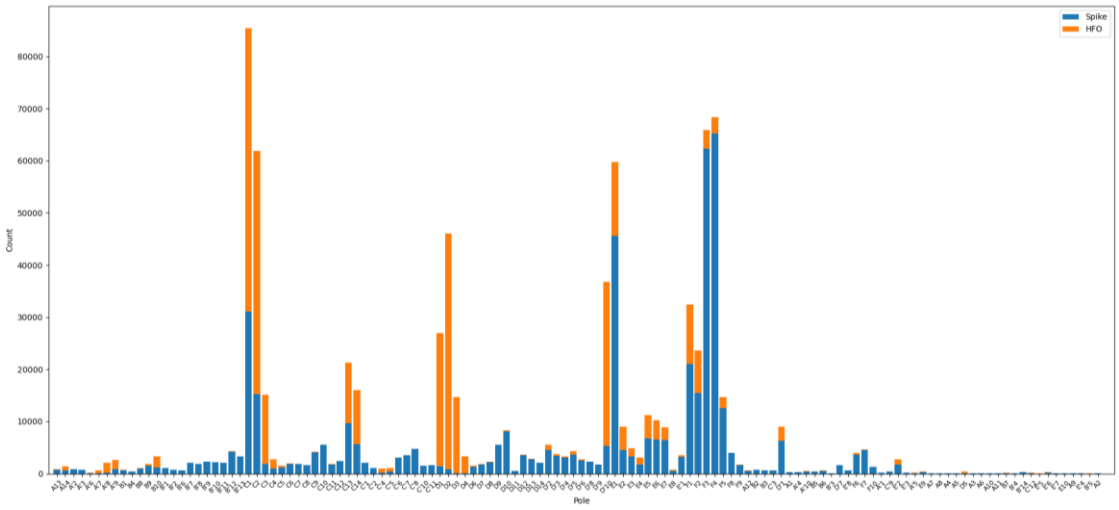

图 20 S2 患者脑电活动电极分布

可以看出，异常活动的空间分布很不均匀。大部分电极上采集到的异常活动数量很少，而少数几个电极上采集到的异常活动数量远远高于其他。同时，不同电极上异常活动的种类

比例极不相同。如 F1-F8 电极上采集到的几乎全部为 Spike，而 D1-D4 电极上采集到的几乎全部为 HFO，而 C1、C2 等电极上采集到的两种活动比例接近。

下面从频谱上观察不同电极上的异常活动分布情况。画出几个电极上所有活动的平均频谱、所有 Spike 的平均频谱和所有 HFO 的平均频谱如下。

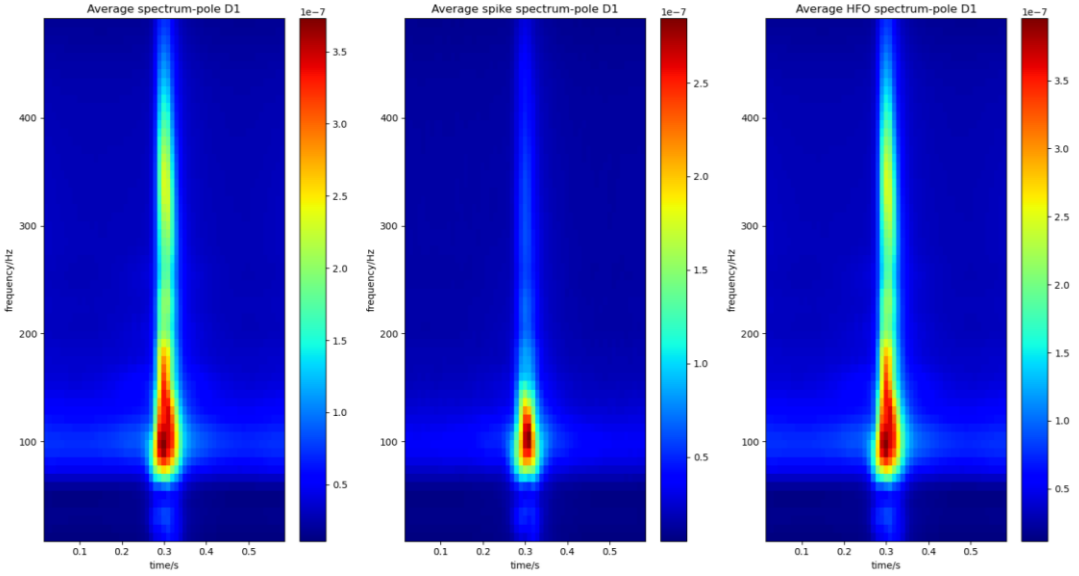

图 21 S2 患者 D1 电极平均频谱

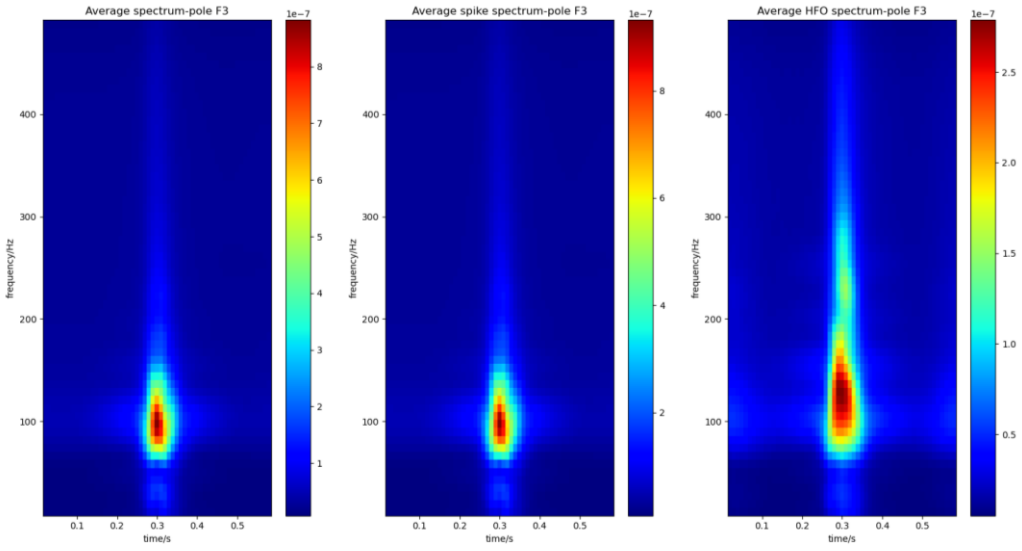

图 22 S2 患者 F3 电极平均频谱

从图 19 中可以看到 D1 电极活动以 HFO 为主，因此其总平均频谱与 HFO 平均频谱十分相近；而 F3 电极活动以 Spike 为主，因此其总平均频谱与 Spike 平均频谱相近。两者的 Spike 平均频谱比较接近，而 HFO 平均频谱差异明显，D1 的 HFO 活动高频频率高于 F3 电极。由此可以得出，同样是 HFO 活动，不同电极处也存在差异。

由上述观察可以发现，异常活动并非均有分布在全脑之中，而是集中在某些位点，同时不同位点有不同特性。结合这些位点的位置和其活动分布特性，可能为针对性地治疗某些位置提供指导。

4 总结

本课程作业中，以真实患者的颅内脑电活动数据为基础，进行了脑电活动分类、脑电活动规律探究。在完成相关任务的过程中，我认识到了如下几方面：

首先，对于大量的数据进行分析时，首先要结合问题的特性和数据的特点适当地处理数据，使得数据变换成最容易指向希望得到的结论的形式。如本任务中，直接在时域分析数据可能效果不佳，而变换到频域则能够获得良好的效果。

其次，不同于用于教学或练习的小数据集，真实的数据往往规模较大，因此处理起来也有一定的难度。在编写相关程序时，要注意内存管理、并行计算等编程技巧，以减小计算的开销和代价。同时，在无法直接对原始数据进行处理和分析时，我们需要在分析的准确度和计算的可行性之间权衡，数据降维就是这样的一种操作。恰当的数据降维能够以较小的分析准确度牺牲换取极大的计算可行性。

尽管前期课程学习和科研参与的过程中接触过较多的数据分析、机器学习等内容，但分析的对象多是通信信号、移动网络数据。本学期的课程是我第一次接触医学相关背景的数据和分析，从中我感受到其分析方式方法既有相通之处，又有独特的特点，也由此感受到了自然科学美妙的相通性于广阔性。

5 文件清单

表 3 文件清单

文件名	说明
报告.pdf	本文档
labels.npy	分类标签*
/code	代码文件夹
ep_utils.py	读取数据等工具函数
data_merge.py	合并数据
data_spectrum_process.py	将数据变换到 63 维频域特征
figure_3.py	2.1 中绘制图 3
figure_5.py	2.1 中绘制图 5
kmeans.py	2.2 中 Kmeans 聚类
show_result.py	绘制部分样本聚类结果
PCA_kmeans.py	2.3 中 PCA 降维后 Kmeans 聚类
show_3d_result.py	2.3 中绘制降维至 3 维的空间点
typical_waveform.py	2.4 中绘制典型波形
PCA_DBSCAN.py	2.5 中 DBSCAN 算法探究
temporal_kmeans.py	2.5 中时域直接聚类探究
time_merge.py	合并时间标签
time_statistic.py	3.1 中时间分布统计
time_spec_statistics.py	3.1 中时间频谱统计
pole_statistic.py	3.2 中电极空间分布统计
pole_spec_statistics.py	3.2 中电极平均频谱统计

*分类标签为 np.array 形式，其中 0 表示 Spike，1 表示 HFO，顺序为全部样本按 data.merge.py 产生的合并数据文件中顺序排列