生物信息学实验大作业：癌组织转录组差异分析

----------生信实验第 13 组成员与分工----------

罗潇琪 2000012264：实验背景调研、转录组数据分析

刘苏瑶 2000012172：实验背景调研、转录组数据分析

李浩田 2200012155：实验背景调研、转录组数据分析、神经网络搭建

项目 github 网址：https://github.com/lhtPeking/CancerAnalysis.git

一、实验准备

[实验目的]

利用 R 工具对胶质母细胞瘤（Glioblastoma，GBM）和嗜铬细胞瘤和副神经节瘤（Pheochromocytoma & Paraganglioma，PCPG）进行差异表达分析功能比较，比较其正常细胞与肿瘤细胞、肿瘤细胞与肿瘤细胞、正常细胞与正常细胞之间的基因表达差异，并绘制相应的火山图、PCA 图等进行可视化分析；

进一步地，通过卷积神经网络的方法完成对不同癌变种类与正常组织的区分（三分类问题），助力临床上的癌症诊断分析；

[背景介绍]

胶质母细胞瘤（Glioblastoma，GBM）是一种高度恶性的中枢神经系统肿瘤，在大脑和脊髓中生长形成，是最常见和最具侵袭性的原发性脑肿瘤之一；它属于 IV 级星形细胞瘤，生长迅速，预后极差。胶质母细胞瘤的临床表现取决于肿瘤的位置、大小和生长速度，常见症状包括：头痛、癫痫发作、神经功能缺损（包括肢体无力、感觉异常、视力障碍等）、认知和行为改变（包括记忆力减退、注意力不集中、性格改变等）、其他症状（如恶心、呕吐、步态不稳等）；

嗜铬细胞瘤（Pheochromocytoma）和副神经节瘤（Paraganglioma）都是起源于神经嵴的神经内分泌肿瘤。嗜铬细胞瘤起源于肾上腺髓质的嗜铬细胞，而副神经节瘤起源于肾上腺外的自主副神经节的神经内分泌细胞，二者的组织病理学、流行病学乃至分子病理生物学特征有一定重叠，因此，在收集到的数据集中将嗜铬细胞瘤（Pheochromocytoma）和副神经节瘤（Paraganglioma）进行合并（PCPG）；这两种肿瘤的临床表现主要由过量分泌儿茶酚胺（如肾上腺素和去甲肾上腺素）引起，常见症状包括：高血压、心血管系统（心动过速、心律失常、心肌病等）、代谢异常（包括体重减轻、高血糖等）、神经系统（焦虑、恐慌发作、震颤等）、其他症状（恶心、呕吐、腹痛等）；

胶质母细胞瘤（GBM）、嗜铬细胞瘤和副神经节瘤（PCPG）都与神经系统发育密切相

关。神经系统包括中枢神经系统（大脑和脊髓）和外周神经系统（神经和神经节），其主要功能是接收、处理和传递信息，控制和协调身体的各项活动。神经系统的发育对于个体的生长、发育和功能实现至关重要；神经系统肿瘤的后果严重，可能会引起患者神经功能障碍、癫痫发作、颅内压增高、神经心理问题等。对于肿瘤，早期诊断和综合治疗是改善预后的关键，因此，通过恰当的方式发现肿瘤细胞与正常细胞的差异，并及时进行诊断，对于肿瘤的治疗是至关重要的；

差异基因分析（Differential Expression Analysis，DEG）是指通过比较不同样本（如健康组与疾病组，正常细胞与肿瘤细胞等）中基因表达水平的差异，识别出在不同状态下有显著表达差异的基因，称为差异基因集。DEG 的原理是判断组间差异（不同样本之间的差异）是否显著大于组内差异（误差）。R 语言中的 DESeq2 工具基于负二项分布模型，通过标准化、统计建模和假设检验等步骤，可以识别出不同条件（如不同实验组）下显著差异表达的基因，再利用使用 EnhancedVolcano 包绘制火山图，就可以直观展示基因的显著性和表达变化。

[实验数据]

基因表达矩阵数据来源于 The Cancer Genome Atlas Program（TCGA），通过 TCGA 官网数据库 https://portal.gdc.cancer.gov/进行下载。下载的文件包括：

TCGA-GBM.htseq_counts.tsv.gz（60,489 identifiers X 173 samples）和

TCGA-PCPG.htseq_counts.tsv.gz（60,489 identifiers X 186 samples）；

[实验工具]

实验使用 R studio 代码平台，实验中使用的程辑包包括：readr，ggplot2，ggsignif，ggpubr，ggplot2，BiocManager：GenomicFeatures, AnnotationDbi，Enhancedvolcano；

神经网络部分在 vscode 上通过 python 完成，基于 linux-64 platform（WSL），更细致环境配置（通过 conda 管理）要求详见 requirements.txt 文档；

二、差异分析实验结果

1. GBM 中正常细胞与肿瘤细胞的差异表达分析

利用 R 语言中的 DESeq2 工具，得到 GBM 中正常组与肿瘤组的差异表达分析结果（其中

log2FoldChange 反映了表达量的相对上调或下调，padj 则反映出该处差异判断的显著性，下同）：

		baseMean	log2FoldCha	lfcSE	stat	pvalue	padj
2	ENSG000002	1763.03165	5.9435797	0.33224177	17.889321	1.428382129	5.82894179536356e-67
3	ENSG000001	2409.96632	7.18524692	0.42114379	17.0612675	2.882326489	5.88109897012058e-61
4	ENSG000002	987.101594	4.89280041	0.29831971	16.4011972	1.875060820	2.5505827315865e-56
5	ENSG000001	5043.72615	7.16091376	0.47455846	15.0896347	1.894883373	1.93316001752766e-47
6	ENSG000002	256.379856	4.8194838	0.32159196	14.9863316	9.019952444	7.36172438680215e-47
7	ENSG000002	316.885077	4.86829862	0.32836166	14.8260262	9.944833932	6.76381305212586e-46
8	ENSG000001	1605.09875	7.23834295	0.497525	14.5487019	5.952182489	3.46995232919006e-44
9	ENSG000002	429.333424	4.74035114	0.32887778	14.4137168	4.242512264	2.16410550590465e-43
10	ENSG000001	807.709644	6.59466319	0.46103045	14.3041815	2.060477465	9.34266271201733e-43
11	ENSG000001	1989.99178	7.56965991	0.52974885	14.2891484	2.557174061	1.04353159089321e-42
12	ENSG000001	1226.70504	7.17013477	0.50255083	14.2674817	3.489536299	1.29455452095493e-42
13	ENSG000002	559.720624	4.54982458	0.32232253	14.115751	3.037623886	9.57216575357825e-42
14	ENSG000001	581.023186	3.65091828	0.25864643	14.1154793	3.049356861	9.57216575357825e-42
15	ENSG000000	2327.97242	-2.2880569	0.16450626	-13.908631	5.614375458	1.63651024067625e-40

Figure.1 GBM 中正常细胞与肿瘤细胞差异表达分析结果

对 GBM 中正常细胞与肿瘤细胞差异表达分析结果进行火山图可视化处理，可以更为清晰地看到二者在基因表达中的差异和变化：

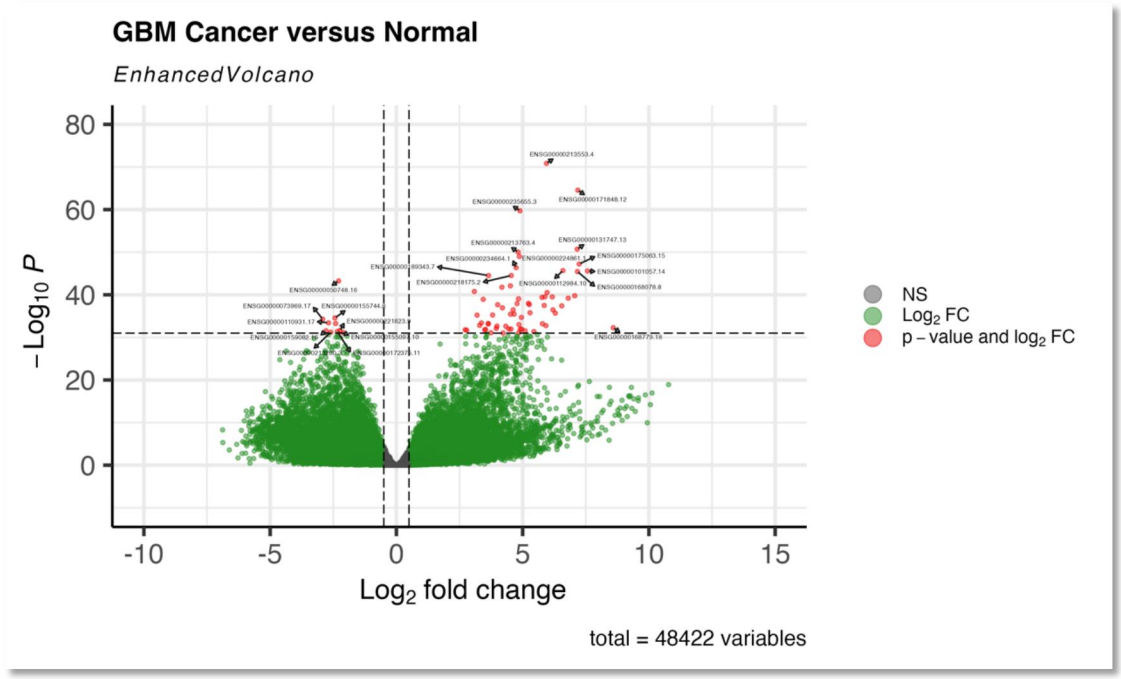

Figure2. GBM 中正常细胞与肿瘤细胞差异表达分析火山图

在火山图中，X 轴表示 GBM 肿瘤细胞与正常细胞之间基因表达变化的幅度，正值表示在肿瘤样本中表达较高，而负值表示在正常样本中表达较高；Y 轴表示观察到的变化的统计显著性，值越高，变化越具有统计显著性，使用 p 值的负对数帮助更清晰地可视化小 p 值。不同颜色的点表示不同的含义：绿色点所代表的基因具有显著的倍数变化（在 log2 尺度上

大于±0.5），但不符合严格的 p 值阈值（小于 10e-32）；红色点所代表的基因同时满足显著的倍数变化和 p 值阈值，这些基因被认为是高度显著的，值得进一步研究；灰色点所代表的基因既不满足倍数变化标准也不满足 p 值显著性标准。可以看到绝大多数的基因都保持正常（变化幅度不大），是少数的原癌基因或抑癌基因发生突变导致的细胞状态的显著变化；通过火山图可视化我们可以很直观的找出驱动癌变的“main gene”。

2. PCPG 中正常组与肿瘤组的差异表达分析

利用 R 语言中的 DESeq2 工具，得到 PCPG 中正常组与肿瘤组的差异表达分析结果。

		baseMean	log2FoldCha	lfcSE	stat	pvalue	padj
2	ENSG00000(	11115.5551	4.89516551	0.33612961	14.5633275	4.806035706	1.23870550404588e-43
3	ENSG00000(	11998.5689	5.34436606	0.36796891	14.5239608	8.542796579	1.23870550404588e-43
4	ENSG00000(	3161.90122	5.51398559	0.40165	13.7283347	6.869436294	6.64045508492513e-39
5	ENSG00000(	7647.68715	5.73581748	0.42082412	13.6299638	2.657122269	1.92641364566541e-38
6	ENSG00000(	9196.39425	5.53145994	0.41062936	13.4706878	2.326881472	1.34959125421313e-37
7	ENSG00000(	5919.06724	5.32303305	0.40325397	13.2001999	8.750257102	4.22929093286302e-36
8	ENSG00000(	1517.67163	5.0640847	0.38832641	13.040794	7.170637190	2.9706925504612e-35
9	ENSG00000(	8435.32896	4.41194125	0.33933139	13.0018661	1.193939196	4.32802958650129e-35
10	ENSG00000(	4755.8701	5.17276333	0.3986685	12.9750992	1.693768649	5.11736148983929e-35
11	ENSG00000(	2605.23655	5.76979523	0.44501349	12.9654388	1.921276738	5.11736148983929e-35
12	ENSG00000(	16655.3751	4.3564608	0.33602602	12.964653	1.941068151	5.11736148983929e-35
13	ENSG00000(	10218.9073	4.61952734	0.35830718	12.8926452	4.951577808	1.19663130383272e-34
14	ENSG00000(	3396.19524	3.89455385	0.30344308	12.834545	1.050181035	2.34271154000917e-34
15	ENSG00000(	2976.04076	4.9060779	0.38598372	12.7105825	5.164825684	1.06985674903154e-33

Figure3. PCPG 中正常细胞与肿瘤细胞差异表达分析结果

对 PCPG 中正常细胞与肿瘤细胞差异表达分析结果进行火山图可视化处理，可以更为清晰地看到二者在基因表达中的差异和变化：

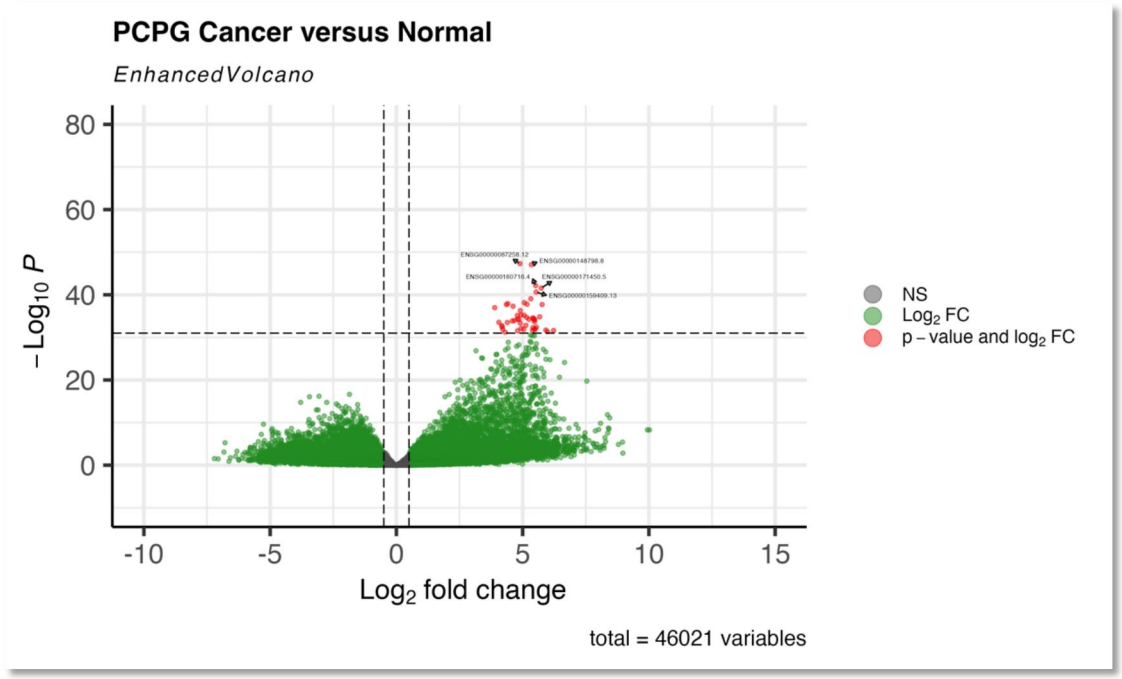

Figure4. PCPG 中正常细胞与肿瘤细胞差异表达分析火山图

与前面的分析一样，X 轴表示 PCPG 肿瘤细胞与正常细胞之间基因表达变化的幅度，正值表示在肿瘤样本中表达较高，而负值表示在正常样本中表达较高；Y 轴表示观察到的变化的统计显著性，值越高，变化越具有统计显著性，使用 p 值的负对数帮助更清晰地可视化小 p 值。绿色点所代表的基因具有显著的倍数变化（在 log2 尺度上大于±0.5），但不符合严格的 p 值阈值（小于 10e-32）；红色点所代表的基因同时满足显著的倍数变化和 p 值阈值，这些基因被认为是高度显著的，值得进一步研究；灰色点所代表的基因既不满足倍数变化标准也不满足 p 值显著性标准。由图可知，与 GBM 一样，PCPG 中绝大多数的基因都保持正常（变化幅度不大），是少数的原癌基因或抑癌基因发生突变导致的细胞状态的显著变化。

3. GBM 与 PCPG 中肿瘤细胞的差异表达分析

通过对 GBM 和 PCPG 中肿瘤细胞进行基因差异表达分析，得到不同肿瘤类型中肿瘤细胞的基因表达的差异情况：

```
> head(GBMvsPCPG_DEG)
log2 fold change (MLE): condition GBM vs PCPG
Wald test p-value: condition GBM vs PCPG
DataFrame with 6 rows and 6 columns
                    baseMean log2FoldChange     lfcSE      stat    pvalue      padj
                   <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
ENSG00000008283.14 65530.485       -5.60190  0.138576  -40.4248         0         0
ENSG00000012223.11 13935.624        8.43456  0.208895   40.3770         0         0
ENSG00000016391.9    600.803        4.98491  0.113574   43.8912         0         0
ENSG00000017483.13  1426.583        5.42257  0.136128   39.8342         0         0
ENSG00000043355.9    265.099       10.14576  0.207658   48.8581         0         0
ENSG00000046653.13 21235.893        5.70724  0.148000   38.5625         0         0
```

Figure5. GBM 与 PCPG 中肿瘤细胞的差异表达分析结果

对 GMB 与 PCPG 中肿瘤细胞的差异表达分析结果进行火山图可视化处理，可以更为清晰直观地看到其在基因表达中的差异:

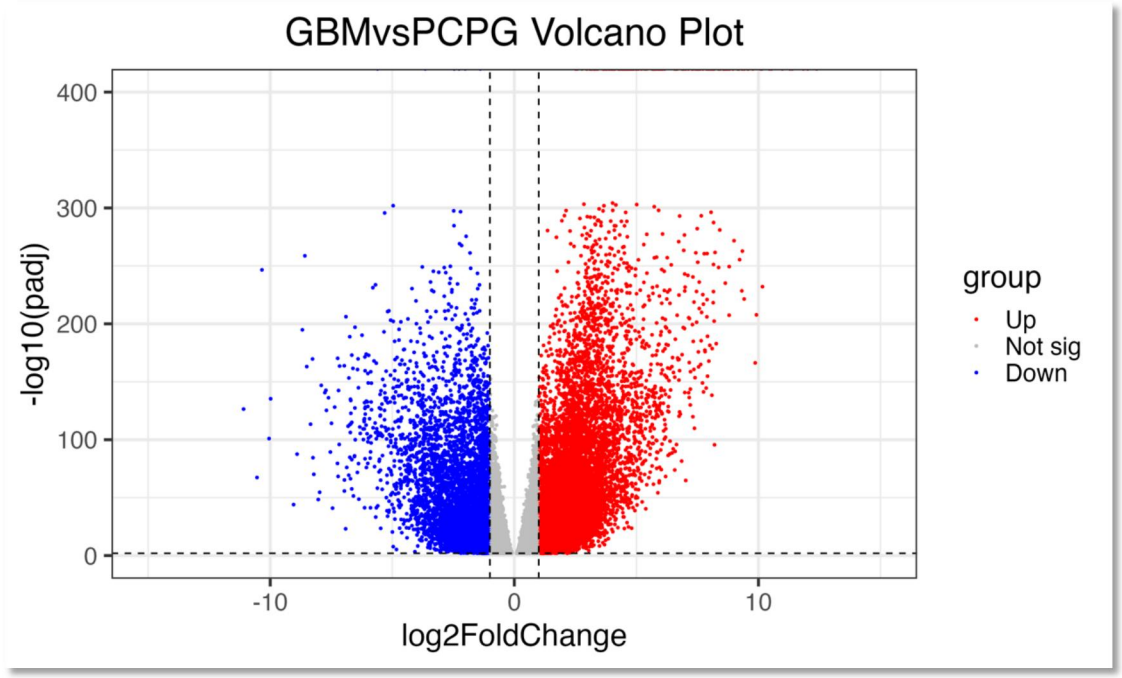

Figure6. GBM 与 PCPG 中肿瘤细胞的差异表达分析火山图

在火山图中，X 轴表示 GBM 和 PCPG 肿瘤细胞之间基因表达变化的幅度。正值表示在 PCPG 中表达较高，负值表示在 GBM 中表达较高；Y 轴表示观察到的基因表达变化的统计显著性，值越高，变化越具有统计显著性，使用调整后的 p 值来考虑多个比较的影响。不同颜色的点表示不同的含义：红色点所代表的基因为上调基因；蓝色点所代表的基因为下调基因；灰色点所代表的基因为无显著变化的基因。火山图中添加了虚线，用于指示显著性阈值。明显可以看出，GBM 和 PCPG 之间的基因表达差异非常显著（p 值非常小），表明两个肿瘤的性质存在本质区别。

4. GBM 与 PCPG 中肿瘤细胞的转录组主成分分析(PCA)

从 Figure6 中的火山图差异分析中我们可以看出两种癌症组织间的基因表达差异是很大的（参与的差异基因数量庞大），为了能够更好地表示出癌症组织间的差异，我们对转录组数据进行了 PCA 分析，对协方差矩阵进行奇异值分解后将高维数据降维至二维，可视化到下图所示的二维平面上：

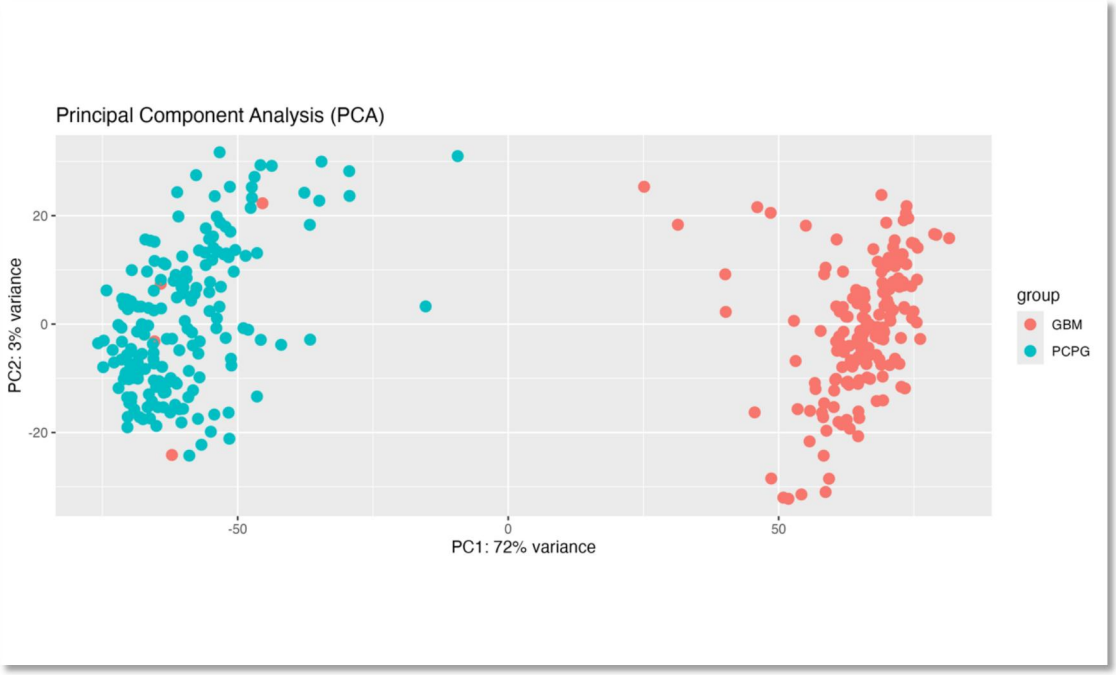

Figure7. GBM 和 PCPG 肿瘤细胞表达情况 PCA 图

Figure8. GBM 和 PCPG 肿瘤细胞表达情况 PCA K-means clustering 聚类图

在 Figure7&8 中，红色的点表示 GBM，蓝色的点表示 PCPG。在 Figure7 中，X 轴（PC1：72%方差）表示 GBM 和 PCPG 肿瘤细胞之间基因表达变化的主要差异，PC1 值为正表示 PCPG 中表达较高，PC1 值为负表示 GBM 中表达较高；Y 轴（PC2：3%方差）表示次要的基因表达变化。在 PC1 上，红色点（GBM 样本）和蓝色点（PCPG 样本）之间有明显的分离，表明这两组样本在基因表达上的主要差异由 PC1 解释。在 PC2 上，样本没有明显分离，表明次要的基因表达变化较小。在 Figure8 中，X 轴（PC1）代表第一主成分，解释了数据中主要的方差。Y 轴（PC2）代表第二主成分，解释了数据中的次要方差。在 PC1 上，红色点和蓝色点之间有明显的分离，表明这两组样本在基因表达上的主要差异由 PC1 解释。由 Figure7&8 可知，在 PCA 降维表出时 GBM 和 PCPG 能较为明显的区分成两类，与基于简单平面欧氏距离的 2-means clustering 的聚类对比，只有少数点的相互掺入，表明从转录组的层面的确能够非常有效的区分出这两种临床上可能不那么容易区分的癌症组织。

5. GBM 与 PCPG 中正常细胞的差异表达分析

通过对 GBM 和 PCPG 中正常细胞进行基因差异表达分析，来判断不同的肿瘤类型是否会对个体癌组织之外的正常细胞基因表达产生影响：

```
> head(res12)
log2 fold change (MLE): Group 2 vs 1
Wald test p-value: Group 2 vs 1
DataFrame with 6 rows and 6 columns
    baseMean log2FoldChange     lfcSE      stat      pvalue        padj
   <numeric>      <numeric> <numeric> <numeric>   <numeric>   <numeric>
1 1070.08885       2.051667  0.175937 11.661398 2.00720e-31 8.09358e-30
2    4.88666       0.158966  1.428790  0.111259 9.11411e-01 9.35753e-01
3 1535.40351       1.612843  0.225225  7.161023 8.00773e-13 8.08840e-12
4  597.15485       0.222203  0.175421  1.266681 2.05269e-01 2.89495e-01
5  108.49475       0.359266  0.349305  1.028515 3.03708e-01 3.98374e-01
6  311.88208       0.615829  0.416696  1.477886 1.39438e-01 2.09918e-01
```

Figure9. GBM 与 PCPG 中正常细胞的差异表达分析结果

对 GMB 与 PCPG 中正常细胞的差异表达分析结果进行可视化处理，可以更为清晰直观地看到其在基因表达中的差异：

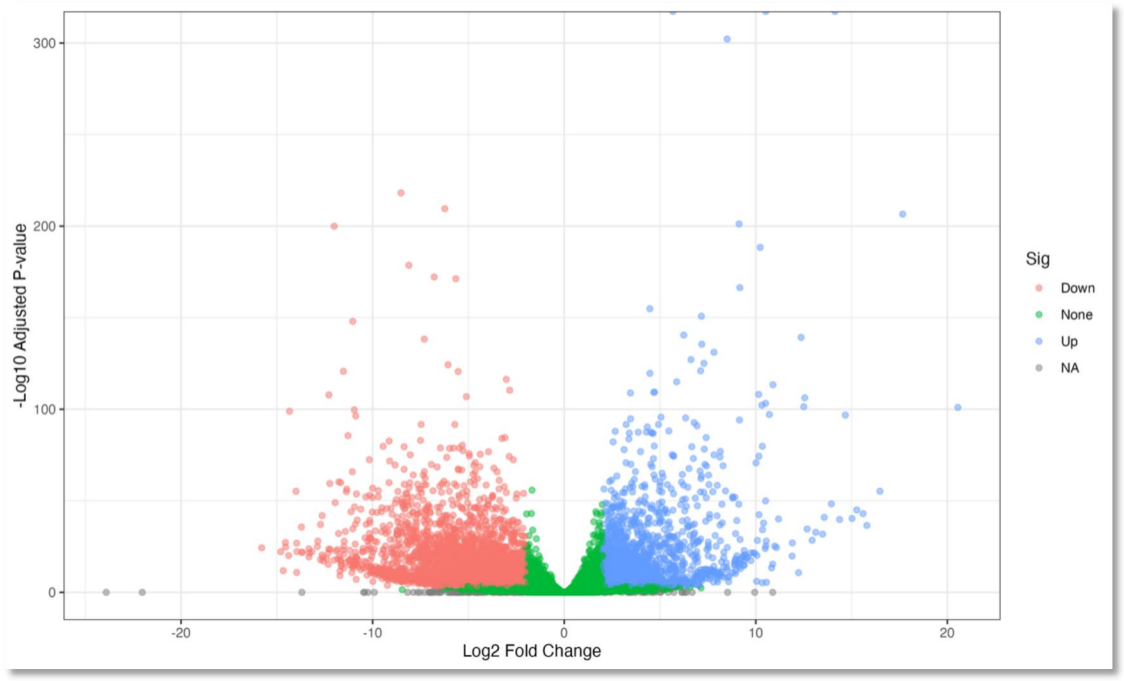

Figure10. GBM 与 PCPG 中正常细胞的差异表达分析火山图

在火山图中，X 轴表示 GBM 和 PCPG 正常细胞之间基因表达变化的幅度。正值表示在 PCPG 中表达较高，负值表示在 GBM 中表达较高；Y 轴表示观察到的基因表达变化的统计显著性，值越高，变化越具有统计显著性，使用调整后的 p 值来考虑多个比较的影响。不同颜色的点表示不同的含义：红色点所代表的基因在 GBM 正常细胞中显著下调（倍数变化小于-2，p 值<0.0001）；蓝色点所代表的基因在 PCPG 正常组织中显著上调（倍数变化大于 2，p 值<0.0001）；绿色点所代表的基因不满足显著的倍数变化和 p 值阈值；灰色点所代表的基因的 p 值没有进行调整或没有达到可报告的水平（NA）。由图可知，在 GBM 和 PCPG 中，即使是正常细胞的基因表达也受到了不同的影响，存在许多基因在两个条件下的表达水平显著不同，表示为红色和蓝色的点，尤其在图的两侧，红色和蓝色点显示出基因表达的极端变化，暗示了两种癌症的正常组织在分子水平上的显著不同。

6. GBM 与 PCPG 中肿瘤细胞的转录组主成分分析(PCA)

同样，为了能够更好地我们对两种癌症的正常组织间的差异，我们对 GBM 和 PCPG 中的正常细胞基因表达进行了主成分分析（PCA）：

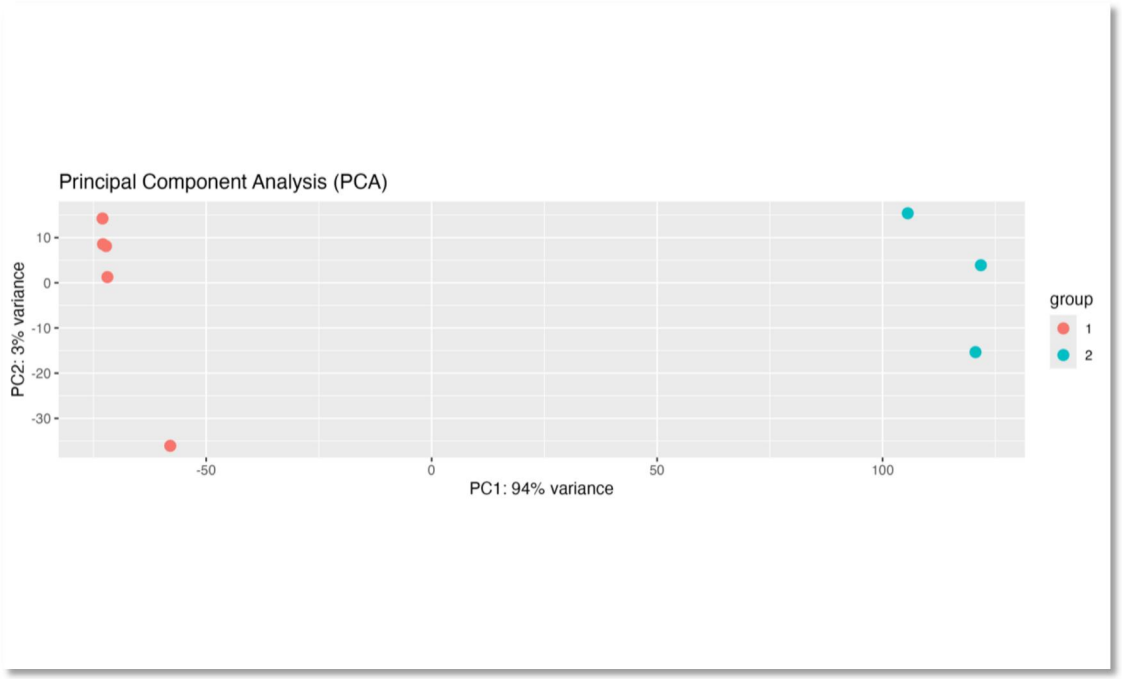

Figure11. GBM 和 PCPG 中正常细胞转录组 PCA 图

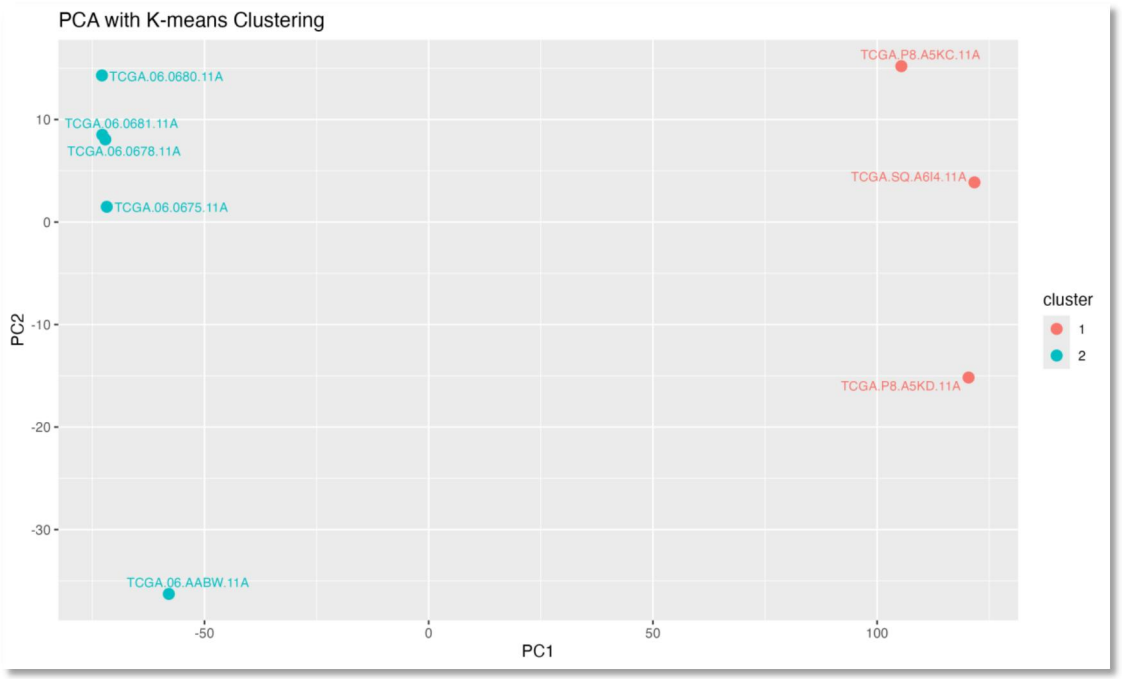

Figure12. GBM 和 PCPG 中正常细胞转录组 PCA K-means clustering 聚类图

在 PCA 图中，蓝色的点表示 PCPG，红色的点表示 GBM，由图可知，虽然数据数量较少，但红蓝两色的点明显聚成了两堆，K-means 聚类与 PCA 标签吻合，同样可以证明在 GBM 和 PCPG 中，非肿瘤细胞的基因表达有显著差异，由此可以得出结论，GBM 和 PCPG 两种肿瘤类型对个体正常细胞的基因表达产生了不同的影响。

结合上述结论得出，在 GBM 和 PCPG 两种肿瘤类型中，无论是肿瘤细胞，还是正常细

胞，彼此之间都有显著的基因表达差异。

三、利用 FNN&CNN 解决基于转录组数据的多分类问题

在对 PCA 标签和 K-means 聚类标签进行比较后我们发现传统的分类方法能够达到很高的准确度，但仍然会有一些掺入点被错分类（详见 Figure7&8 的标签差异）；于是我们想要借助近年来发展出的神经网络方法完成这个分类问题，以此达到更高的准确率；

我们首先搭建了一个简单的全连接网络结构（算法位于 FNNclassification.py 文件中）：首先，我们从两个文件中读取转录组数据（TCGA-GBM.htseq_counts.tsv 和 TCGA-PCPG.htseq_counts.tsv），这些文件的每一行代表一个基因，每一列代表一个样本。读取数据后，我们将其从对数空间转换回原始表达量空间，即通过计算 2^data - 1 并四舍五入来处理数据；

为了确保数据的有效性，我们选择了在所有样本中表达量均不为 0 的基因，并从中随机选取了 1000 个基因。这些基因将作为特征，用于后续的分类任务。接下来，我们为每个样本创建标签：对于 GBM 数据，如果样本编号的第 14 个字符为“0”，则标记为肿瘤（1）；否则为正常（0）；对于 PCPG 数据，肿瘤样本标记为（2），正常样本标记为（0）；

在数据处理方面，我们对选取的 1000 个基因的数据进行归一化处理，使其值在 0 到 1 之间。我们使用 MinMaxScaler 对数据进行归一化，并将数据分为训练集和验证集，比例为 8:2；然后我们将 GBM 和 PCPG 的数据分别进行训练集和验证集的合并，并使用 LabelEncoder 将标签转换为数值形式；接下来用 to_categorical 函数将标签转换为 one-hot 编码，GBM 肿瘤样本为 [1, 0, 0]，PCPG 肿瘤样本为 [0, 1, 0]，正常样本为 [0, 0, 1]；

为了使数据适应全连接神经网络的输入格式，我们保持数据为二维数组形式，即每个样本的数据形状为 (样本数，基因数)，其中基因数为 1000；

接下来构建全连接神经网络模型，这个模型包括以下层：

第一层全连接层：使用 512 个神经元，激活函数为 ReLU，输入形状为 (1000)；

第一层 Dropout 层：防止过拟合，丢弃 50%的神经元；

第二层全连接层：使用 256 个神经元，激活函数为 ReLU；

第二层 Dropout 层：防止过拟合，丢弃 50%的神经元；

第三层全连接层：使用 128 个神经元，激活函数为 ReLU；

第三层 Dropout 层：防止过拟合，丢弃 50%的神经元；

输出层：使用 3 个神经元（对应 3 个类别），激活函数为 Softmax；

我们使用 categorical_crossentropy 作为损失函数，adam 作为优化器，accuracy 作为评估指标；在训练模型时，我们设置了 30 个训练周期，每个批次处理 32 个样本，并在每个周期结束时使用验证数据评估模型性能；

训练完成后，我们使用 Matplotlib 绘制训练和验证的准确率变化曲线，并使用验证集评估模型性能，输出测试准确率和混淆矩阵；最终，我们将训练好的模型保存为 fcnn_model.h5 文件，以便后续使用；

Training logs 和混淆矩阵可视化如下：

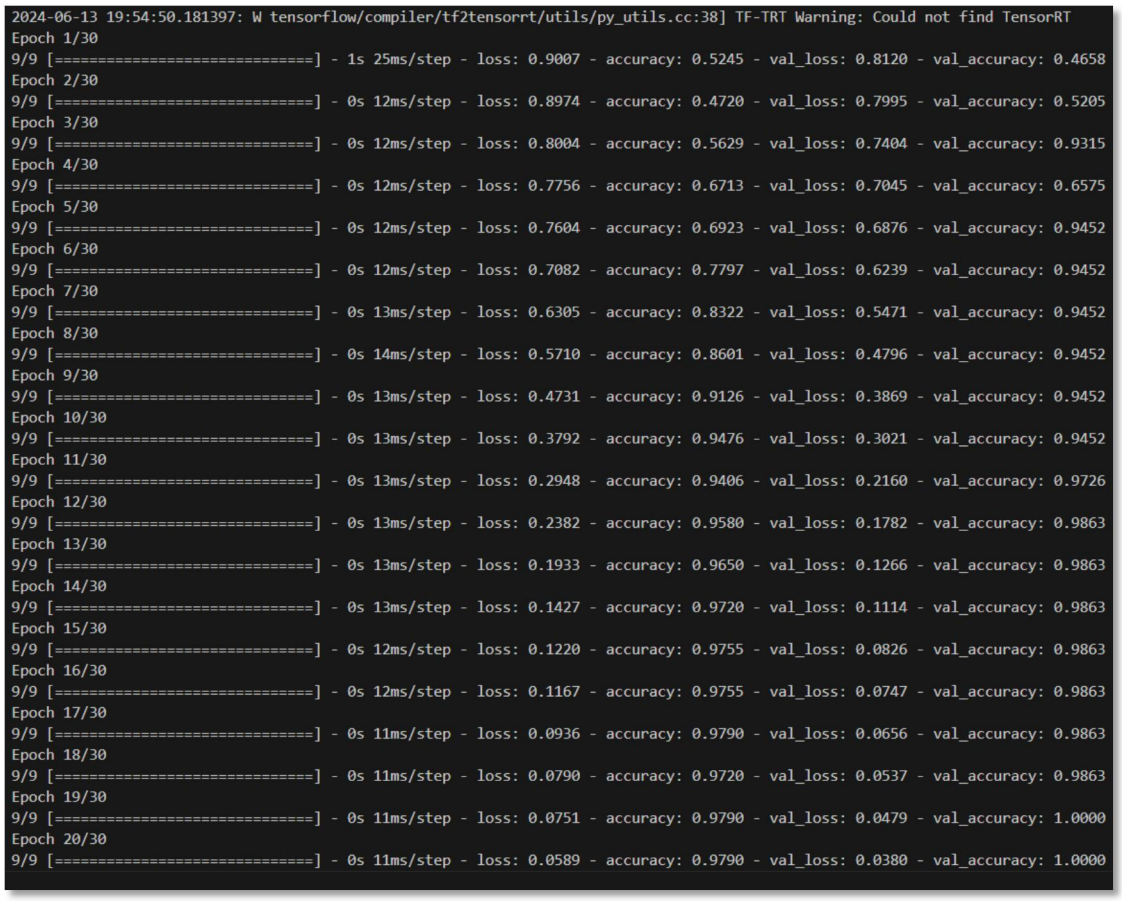

Figure13.FNN training log.

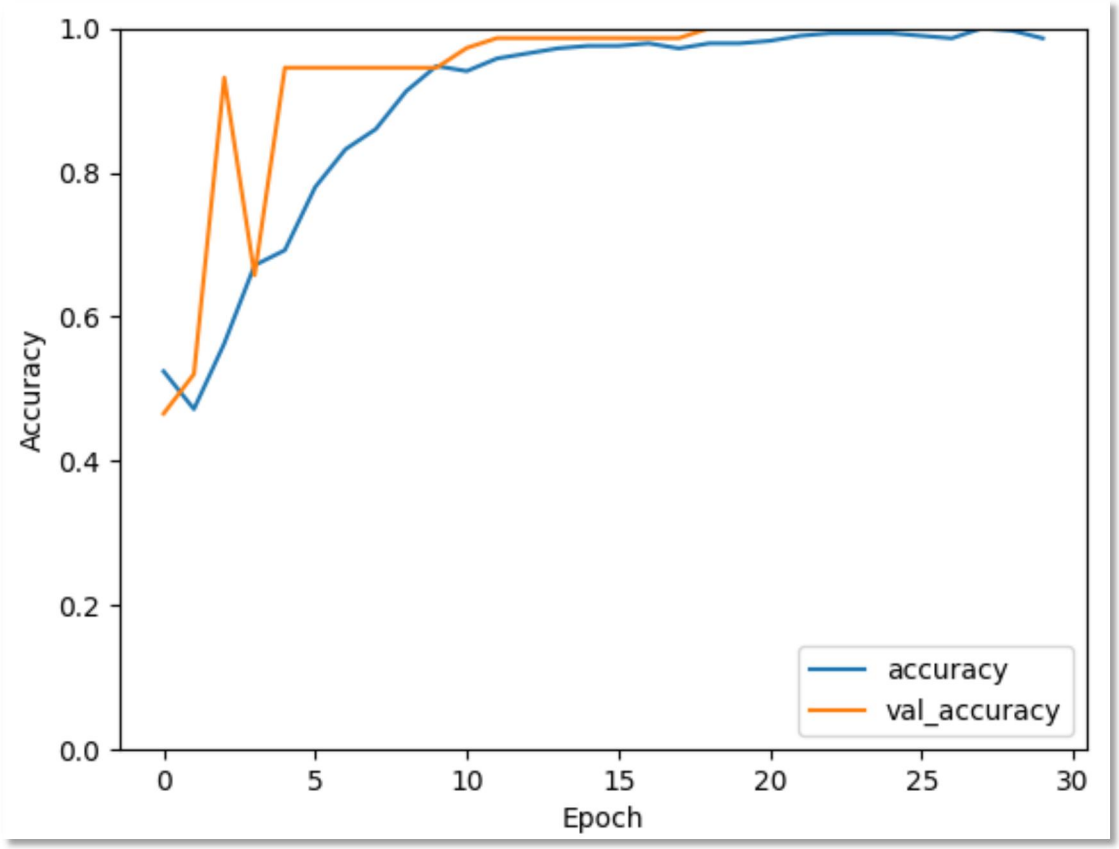

Figure14.FNN training log visualization.

```
Test accuracy: 1.0
3/3 [==============================] - 0s 3ms/step
[[ 1  0  0]
 [ 0 34  0]
 [ 0  0 38]]
```

Figure15.FNN classification task (confusion matrix).

可以看到随着训练 epoch 的增加，全连接神经网络模型在训练集和验证集上的准确率都逐步提升，并且最后的混淆矩阵分类准确率达到了 100%（confusion matrix 为对角矩阵）；

但是我们并不满足于这一点，因为训练初期模型的震荡较为严重，表现出了不稳定性，于是我们打算使用卷积神经网络（CNN）来处理这个较为复杂的分类任务，期望达到更好的分类效果（算法位于 CNNclassification.py 文件中）；

数据处理与前面的 FNN 大致一致；为了使数据适应卷积神经网络的输入格式，我们将

训练和验证数据重塑为四维张量，形状为 (样本数，高度，宽度，通道数)。具体来说，每个样本的数据被重塑为 10x100 的二维数组，其中 10 表示高度，100 表示宽度，1 表示通道数（类似于灰度图像）。

接下来，我们构建了一个卷积神经网络模型。这个模型包括以下层：

第一层卷积层：使用 32 个 3x3 的卷积核，激活函数为 ReLU，输入形状为 (10, 100, 1)。

第一层池化层：使用 2x2 的最大池化层，减少特征图的尺寸。

第二层卷积层：使用 64 个 3x3 的卷积核，激活函数为 ReLU。

第二层池化层：使用 2x2 的最大池化层。

展平层：将多维特征图展平成一维向量。

全连接层：使用 128 个神经元，激活函数为 ReLU。

Dropout 层：防止过拟合，丢弃 50%的神经元。

输出层：使用 3 个神经元（对应 3 个类别），激活函数为 Softmax。

我们使用 categorical_crossentropy 作为损失函数，adam 作为优化器，accuracy 作为评估指标。在训练模型时，我们设置了 30 个训练周期，每个批次处理 32 个样本，并在每个周期结束时使用验证数据评估模型性能。

训练完成后，我们使用 Matplotlib 绘制训练和验证的准确率变化曲线，并使用验证集评估模型性能，输出测试准确率和混淆矩阵。最终，我们将训练好的模型保存为 cnn_model.h5 文件，以便后续使用；

Training logs 和混淆矩阵可视化如下：

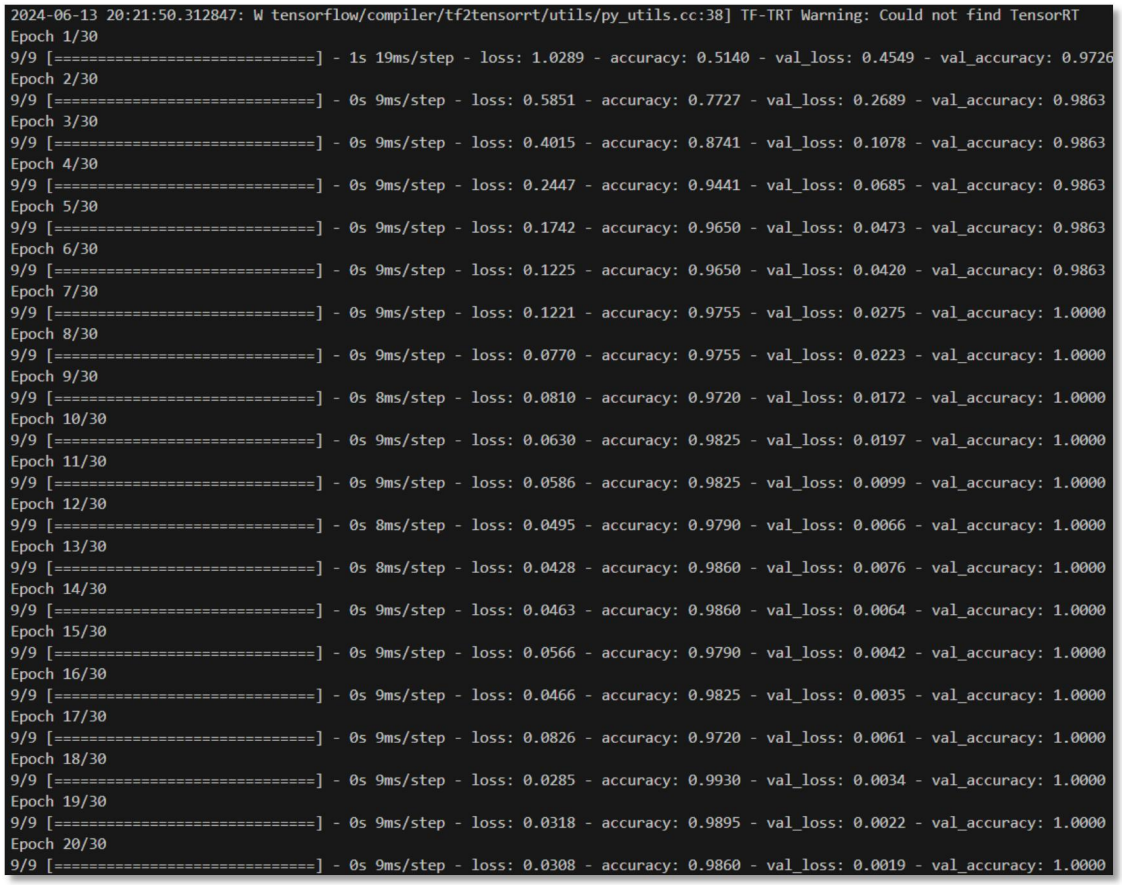

Figure16.CNN training log.

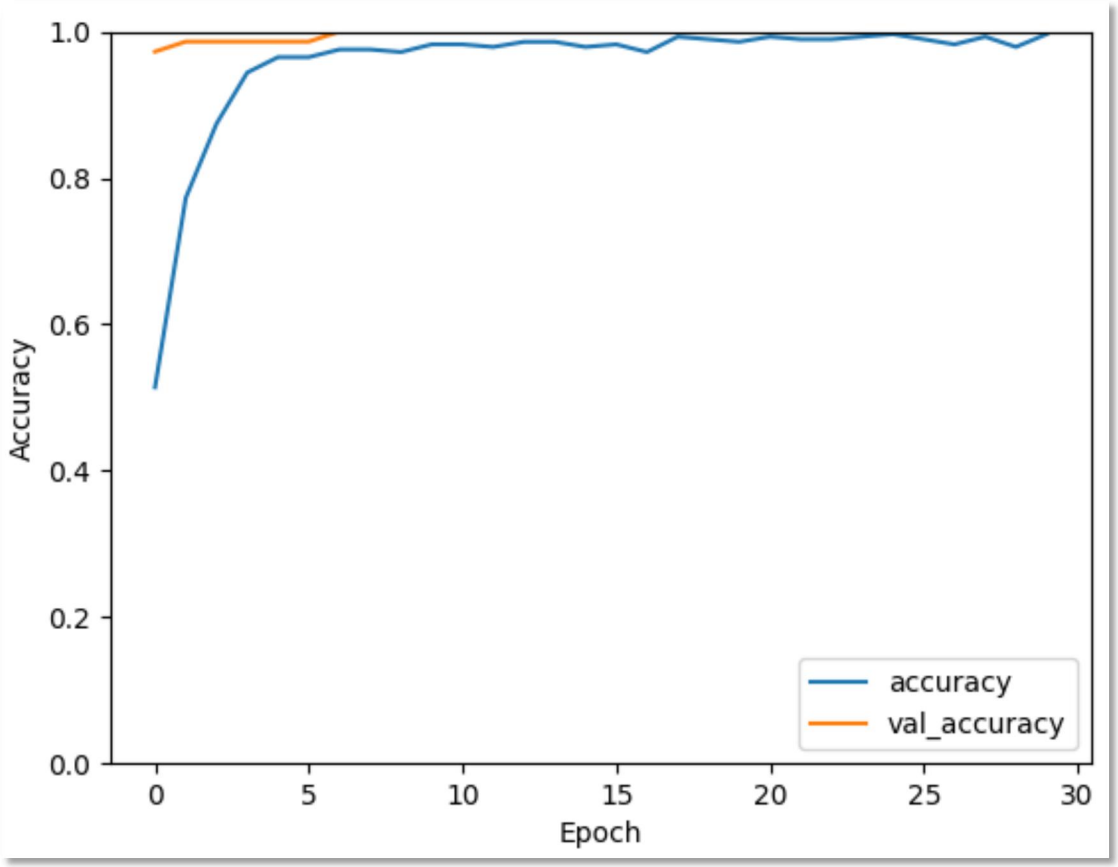

Figure17.CNN training log visualization.

```
Test accuracy: 1.0
3/3 [==============================] - 0s 2ms/step
[[ 1  0  0]
 [ 0 34  0]
 [ 0  0 38]]
```

Figure18.CNN classification task (confusion matrix).

可以看到随着训练 epoch 的增加，卷积神经网络模型在训练集和验证集上的准确率都逐步提升，并且最后的混淆矩阵分类准确率达到了 100%（confusion matrix 为对角矩阵）；并且相对于前面的全连接神经网络，CNN 在训练初期就表现出了很高的准确性和稳定性，这是我们模型优化成功的证明。

综上所述，我们首先通过全连接神经网络完成了较为粗糙的分类，再利用卷积神经网络对分类算法进行了优化，最后达到了很好的效果；表明通过神经网络来癌症组织拟合分类器是可行的，并且相对传统的分类有更高的准确率。

四、总结与讨论

本实验通过利用 R 语言进行差异表达分析，揭示了胶质母细胞瘤（GBM）和嗜铬细胞瘤及副神经节瘤（PCPG）在基因表达水平上的显著差异。实验结果表明，无论是肿瘤细胞还是正常细胞，GBM 和 PCPG 之间都有明显的基因表达差异变化。这些差异反映了两种肿瘤在分子机制上的本质区别，也揭示了肿瘤对正常细胞基因表达的影响。通过火山图和 PCA 图的可视化分析，我们能够直观地观察到这些差异，并识别出在肿瘤细胞中显著变化的基因，这些基因可能在肿瘤的发生和发展中起到关键作用，对这些基因的进一步研究将可能揭示肿瘤发病的关键基因，并提供进一步寻找治疗靶点的可能性。

本实验仍然存在一些局限性：首先，实验样本量相对较小，可能影响结果的代表性和统计显著性。其次，数据多样性有所欠缺，仅使用了一种数据来源（TCGA）进行分析，缺乏对多种数据类型（如蛋白质组学、代谢组学等）的整合分析，对肿瘤的分子机制的揭示较为片面。另外，模型复杂度方面也有所限制，卷积神经网络的应用虽然展示了其潜力，但当同一个网络结构用于不同的肿瘤数据集的时候需要考虑模型的泛化能力，这一点目前来说是较难做到的。

对于未来的研究，我们认为有以下可以继续改进和推进的方向：第一，通过增加更多的样本量和引入不同类型的数据（如蛋白质组学、代谢组学和表观遗传学数据等），可以更为全面地分析肿瘤的分子机制，提升研究结果的可信度和应用性；第二，可以对差异表达分析中发现的关键基因进行功能验证和机制研究，探索其在肿瘤发生、发展中的具体作用及其潜在的治疗价值；第三，增强卷积神经网络的泛化能力，提供给临床部门 pre-training 的模型，临床部门再根据自己的肿瘤数据集进行一定的 fine-tuning，从而达到广泛应用的效果。