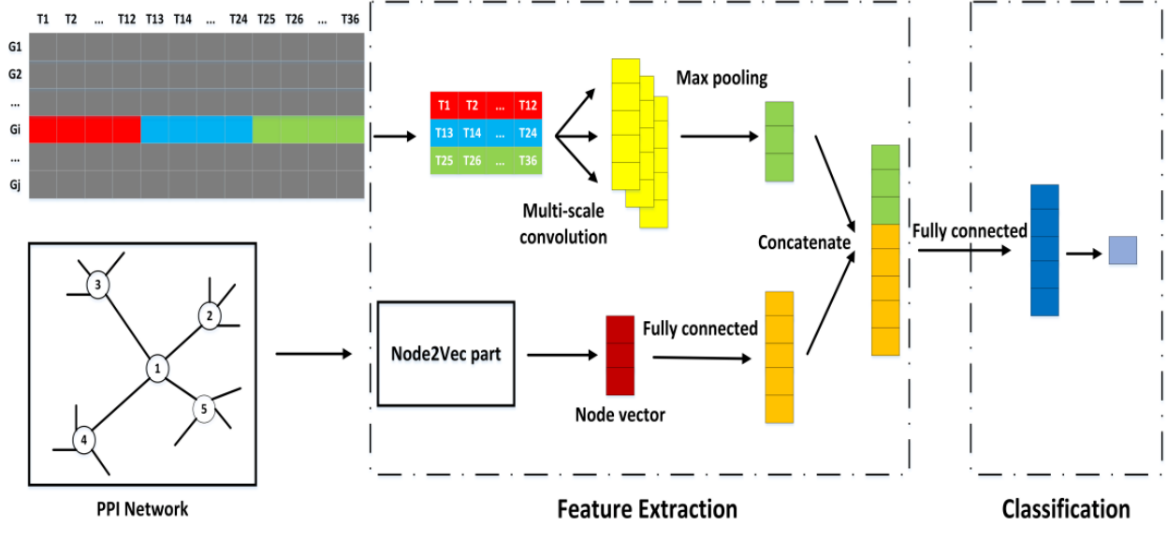
**LSTM model for identification of essential proteins**

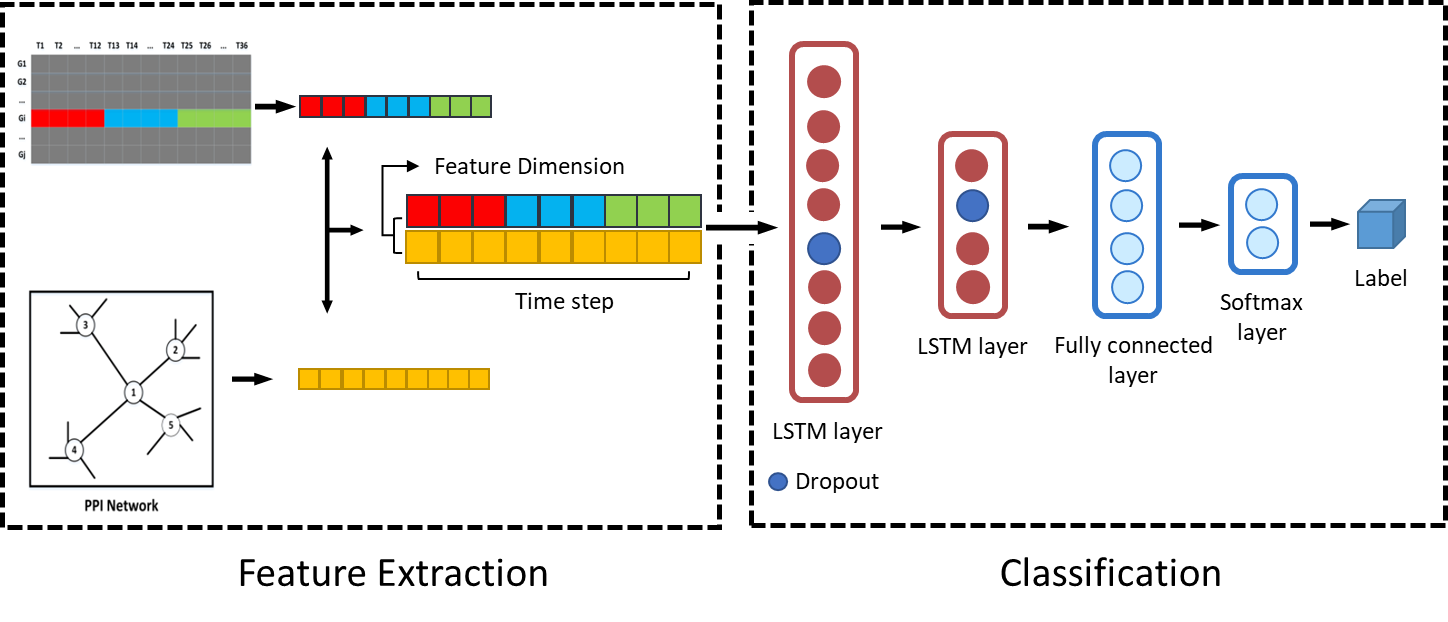
**Our method**

之前我们运用了LSTM模型对关键蛋白质进行了预测，所采用的方法与李敏老师在BIBM中提出的方法类似，只是我们采用了LSTM模型来提取特征。

我们借鉴了李敏老师的思想，通过 Node2vec 算法提取网络拓扑结构特征并与基因表达数据结合用于关键蛋白质的识别，只是这里我们对所有数据采用序列模型进行处理。



**图1. 李敏老师的网络结构**

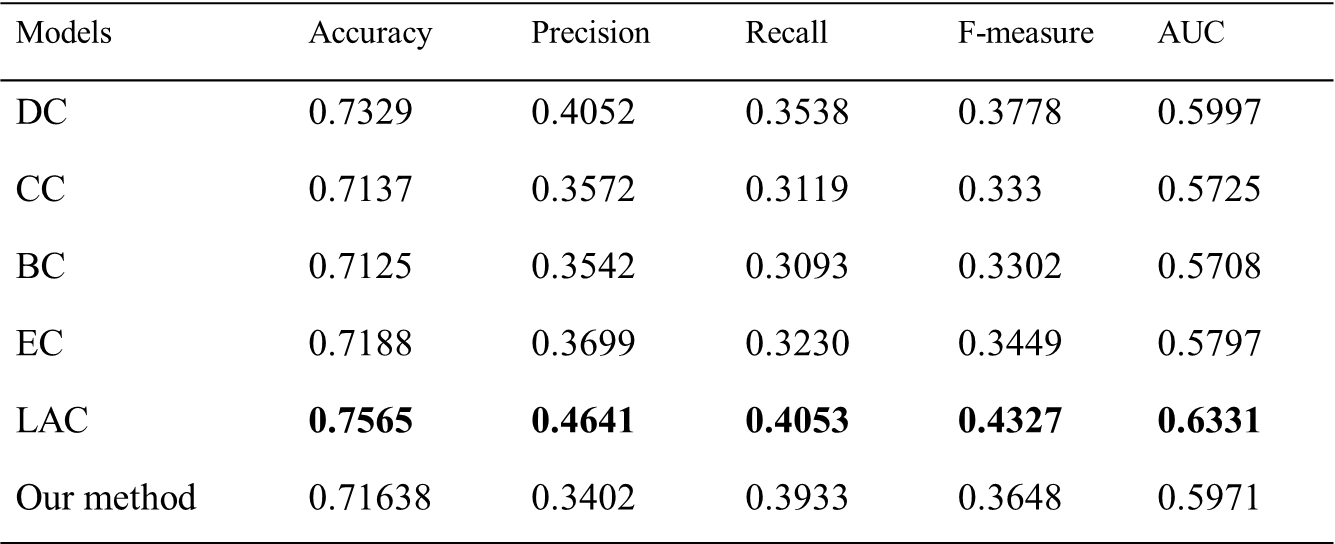
****

**图2. An overview of our proposed deep learning framework for identifying essential proteins.**

在里的序列维度为2，分别为node2vec的网络结构信息和gene expression data，序列的长度为36，也即为基因表达序列的长度，这里的网络结构信息长度也只有36，是为了与基因表达信息长度匹配，而李敏老师的算法里长度是72，所以我们为了与基因表达数据匹配牺牲了一部分信息。模型对比的结果如下。



**表1.** **compared with machine learning methods**

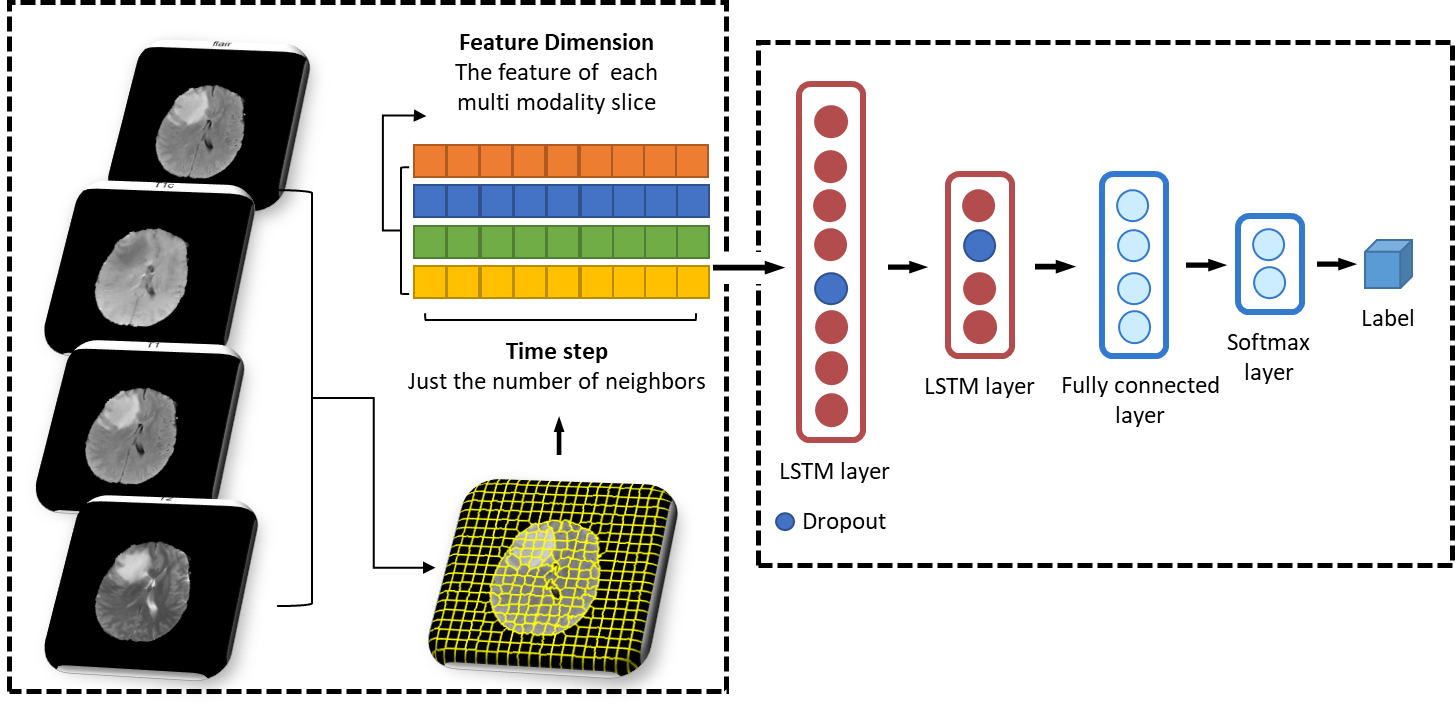


**表2.** **compared with traditional methods**

结果的话比李敏老师的算法（李敏老师的算法我采用pytorch复现的，使用数据与LSTM model 的一样）表现稍微好一点，但是还是很弱，连traditional的方法都比不了。

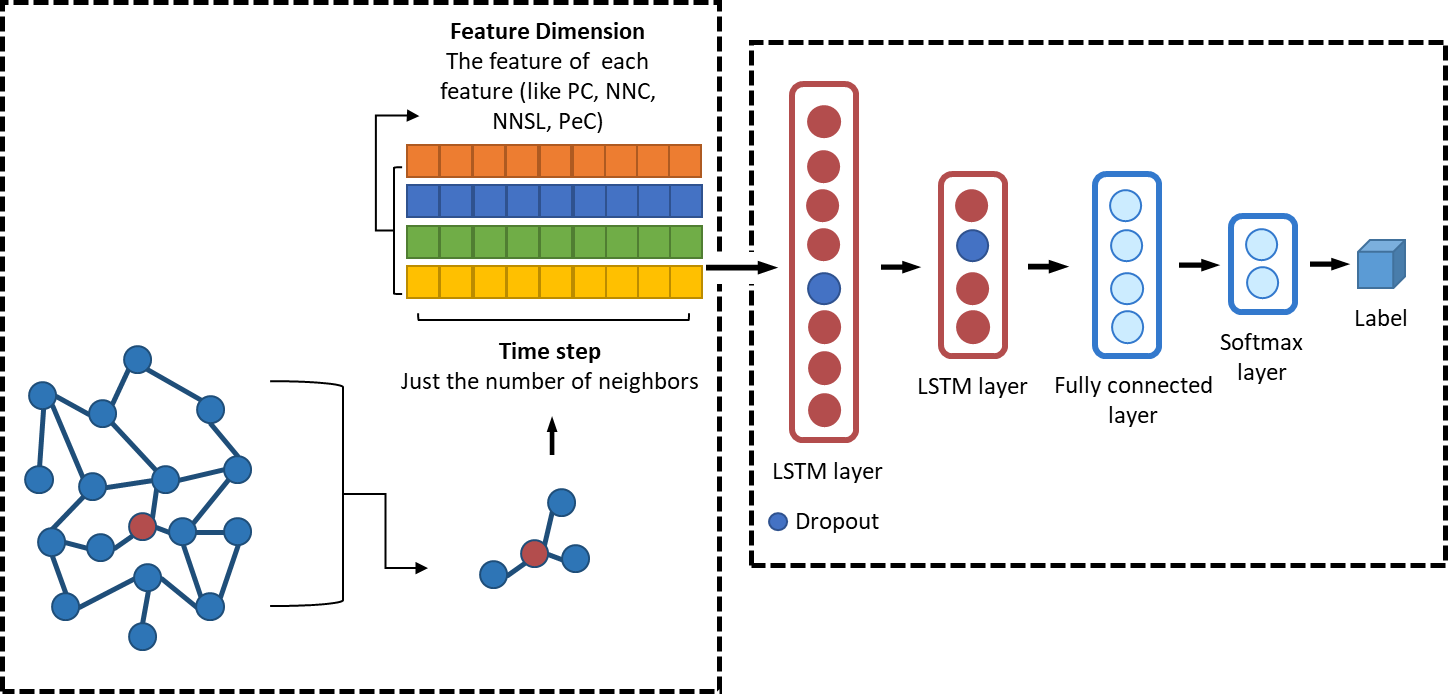
**Now**

最近在改这篇SGC，结合了我们做医学图像的一些经验，我想能不能和用LSTM做多模态分割一样把多种特征信息给融合进来，之前的model为了维度匹配还牺牲了一部分信息，其他的生物信息，像子细胞位置信息，蛋白质模块信息长度都是1，和基因表达信息的长度不匹配，所以我们这里采用了分割的思想对特征进行整合。



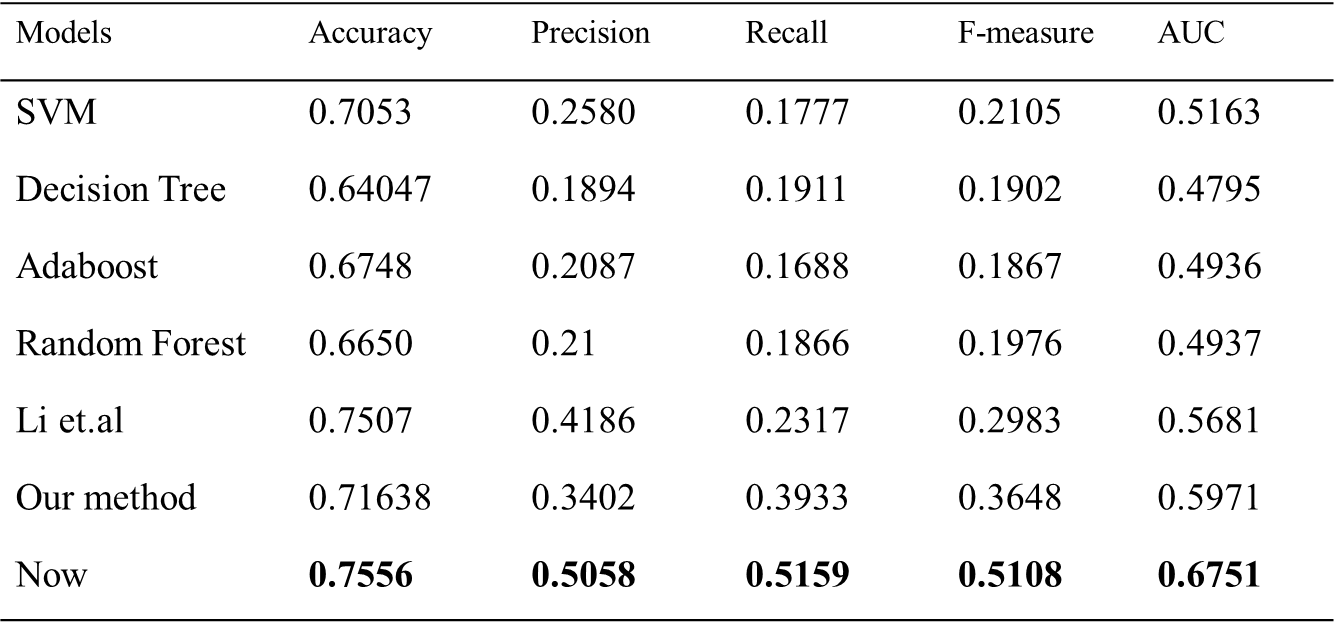
**图3. LSTM model for brain tumor segmentation.**

**思想是我们用SGC中提取得到的特征作为节点特征，然后按照网络信息构建序列，这样序列的维度就是采用的特征的个数（SGC中采用的是NC, PeC, PC, NSL），长度就是每个节点的邻居个数。**

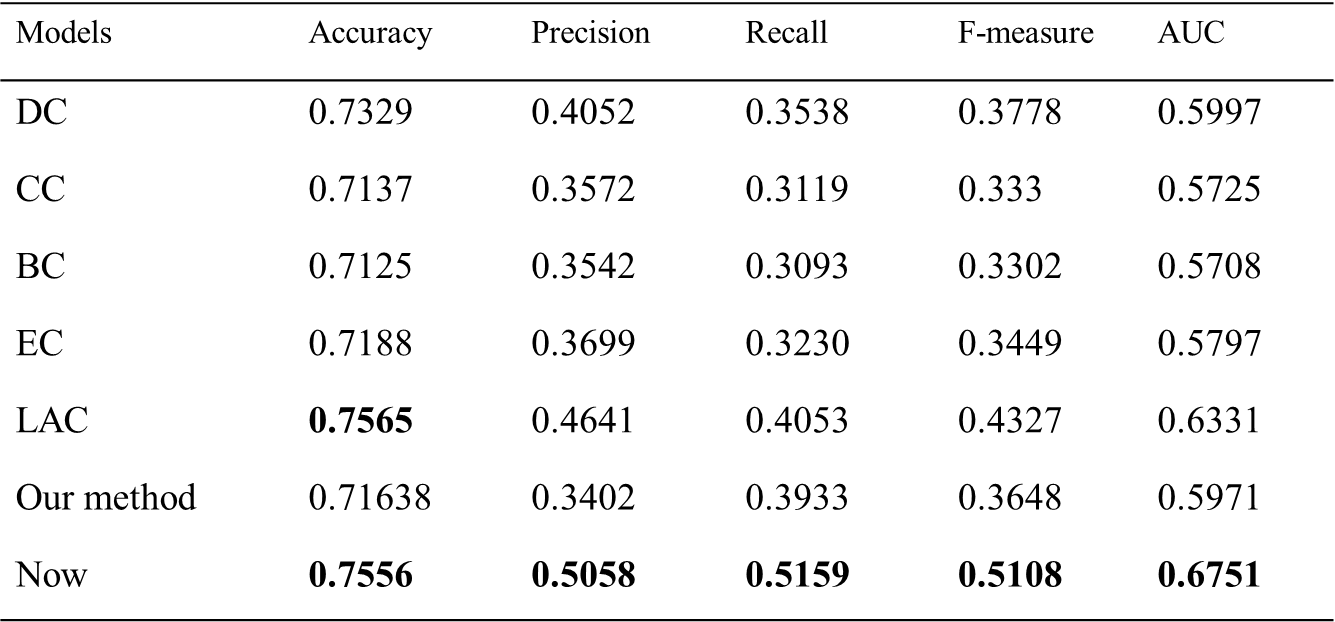


**图4. New LSTM model for brain tumor segmentation.**

**结果的话取得了较大的识别精度提升，可以看到相比于之前的方法，各个评价方法上都取得了较大的提升。比李敏老师的方法提高了不少，相比于传统的算法也比较能看了。**

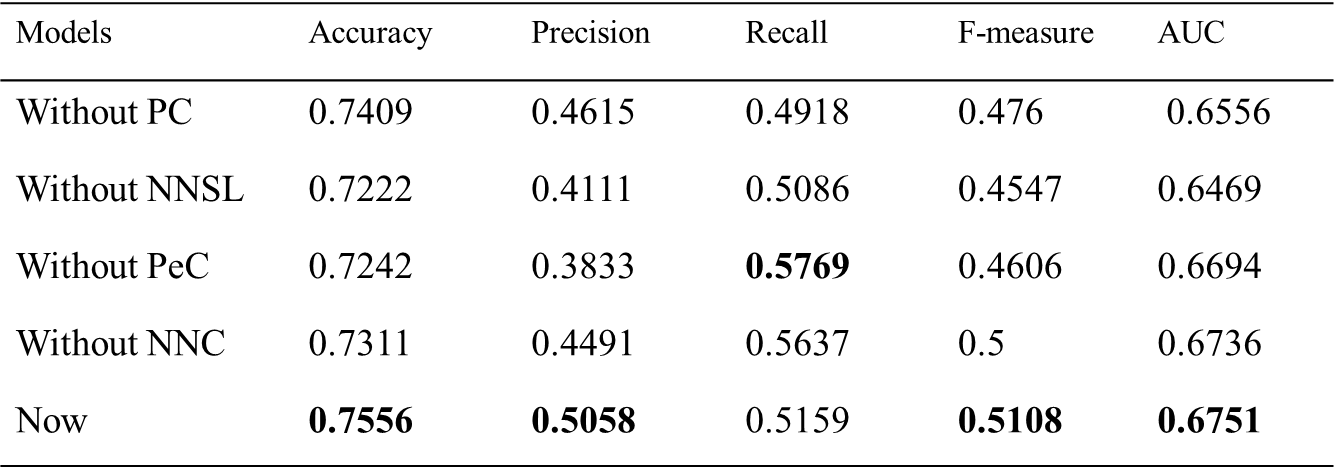


**表3.** **compared with machine learning methods**



**表4. compared with traditional methods**

这里为了进一步明确各个特征的1作用，我们还做了一下消融实验，也就是除掉某个特征，看看整体表现怎么样。这里也可以看出NNSL对算法的影响最大。



**表5. Ablation Experiment**