各位老师上午好，我叫周颖，我的论文题目是 生成式模型在不平衡分类中的应用，接下来我将从五个方面来介绍：研究目的及意义，国内外研究现状，研究内容，已完成的工作以及存在的问题和拟解决方案。

首先介绍一下不平衡分类，是指不同类别的样本数量不等的模式分类问题，在这些问题中，某些重要类别的样本数量远少于其他类，传统的分类器在该问题上难以获得令人满意的性能，但少数类的识别又是非常重要的，因此需要对该类问题进行更加深入的研究。

如图所示，在疾病监测中，橙色的点表示正常，而蓝色的点表示有病，根据传统的逻辑回归方法，会将大部分的样本点划分为正常，但是如果将有病的检查结果判定为正常的，则可能导致严重后果，因此需要提高蓝色样本点即少数类的分类精度。

针对不平衡问题，学者们分析了一下几个原因，包括数据稀缺：指少数类数量少导致分类效果不佳，噪声问题：少数类样本少导致其抗干扰能力差，分类器偏移问题：基于决策面的分类器根据最小化经验误差等，因为多数类的累计误差较大，所以分类面会自然偏，而基于概率估计的分类器，例如贝叶斯等，少数类的先验概率很低，所以样本划分为少数类的可能性也随之下降，评价指标例如全局准确率，在极端情况下，当多数类和少数类的比率为9:1时，则当所有的样本都被预测为多数类时，则准确率仍为90%，但该结果对少数类是无意义的。

以上这些问题都会造成少数类识别低，因此针对这些问题研究者们已经提出了很多方案

例如针对数据量少而产生的采样算法，即在样本数量上平衡不同类别的样本，以及算法层面上，将少数类予以更大的权值，产生的代价敏感算法等，由于在不平衡分类中，少数类的样本数量是训练分类器的瓶颈，因此本文中采用了 对少数类进行过采样的算法，过采样算法发展基本分为三个过程，最开始的简单复制及加入随机噪声，2002年SMOTE算法被提出，之后产生了一系列SMOTE算法的变体，但SMOTE产生的样本有效性和可行度一直被诟病，因此之后有人提出了利用样本分布信息来产生新样本的过采样算法，利用现有的概率密度函数对少数类进行建模并采样，例如高斯函数等，但是精确的概率密度函数能够建模的数据有限，并不能完全覆盖数据的真实分布，因此本文提出利用神经网络作为概率分布函数进行建模，因为神经网络的建模能力强，且其输出维数没有限制，针对样本有效性问题，我们提出采用可视化方法对其定性分析，并利用最终的分类器性能作为其评价标准。

单纯的MLP模型没有监督过程，则

(Kingma and Welling, 2013)提出了变分自编码器的概念，提出将隐层空间未知分布𝑝(𝑧)近似为𝑞(𝑧)，最小化𝑝(𝑧)与𝑞(𝑧)的KL散度，并利用自编码器形成有监督生成网络架构，以获得比较合理的生成样本。这样产生的样本分布相似，且与原始样本间的差距也较小。

存在的问题