选题报告要求：

1. 选题报告的字数一般不少于五千字；
2. 选题报告的内容应包括：选题意义、该领域国内外研究动态（文献综述）、本课题研究的目的、预期成果、研究方案、研究方法及其论证、关键难点拟采取的解决措施、论文工作总体日程安排，预计答辩时间等；
3. 要求查阅一定数量的中、外文献资料。文献综述部分不是将文献内容进行简单的堆砌，而应通过阅读，消化、提炼，对已有的研究成果和动态进行全面的综述；
4. 选题报告的考核由两部分组成：书面报告和口头报告，分别评分后给出一个总成绩；
5. 选题报告和研究工作应具有正确的政治立场，符合学术规范；

填好“研究生选题报告及论文工作计划”表格后，连同书面报告一起交院系研究生管理部门备案。

## 选题意义

流式细胞仪是一种在现代医学分析中常用的仪器，常采取CCD或者单像素成像的方法来采集细胞图像。而细胞种类，细胞特性，细胞状态等就直接通过形态表现出来，如何自动化地根据细胞的形态来进行细胞分析？本课题旨在将在图像分析表现突出的机器学习方法应用在流式细胞分析上面，探索比传统流式细胞分析更快速、更精准的方法。

## 文献综述

在文献综述中，我会先简单回顾一些机器学习方法，随后再讨论机器学习方法在细胞图像分析中的应用，最后再具体到机器学习在流式细胞上的分析。

机器学习是人工智能的一个方面，其目的就是根据大量的数据或者已有的经验，而对新的目标做出预测。可以分为有监督学习和无监督学习，有监督学习需要在训练数据时就对输入进行赋值，往往是人为进行干预，而无监督学习则是根据数据的某种特性，比如聚合离散程度，而对数据进行分类或者回归预测。

在机器学习中，核方法是很重要的一类，如Vapnik V等（1963）最早提出的支持向量机方法、Schölkopf B等（1998）提出的核化的主成分分析方法以及由Gonen M等（2011）提出的核化k均值算法。对于核方法，观测的特征是基于正半定对称核函数固有地提供的，该函数可以解释为希尔伯特空间中的相似性度量，可以让特征空间以一种更为直观的方式呈现出来，便于对转化后的特征空间进行数据分析。此外，增强（boost）方法也是常用的一类，它可以将多个弱学习器转化为一个强学习器，从而提升预测性能。Friedman J H等（2001）提出的梯度增强算法是在机器学习竞赛如kaggle中应用最普遍的一种。这些增强算法会反复迭代学习弱分类器，然后将它们组合为强分类器，例如让弱分类器的输出进行加权，这在统计上也定义明确且易于实现，并且其过拟合现象也很容易分析。LeCun Y（2015）中提到的深度学习方法则是机器学习的一类，它通过大量数据训练多层神经网络来对目标进行预测，深度这个词就来自于神经网路的层数之多，往往达到几十上百层，这被认为可以有效地表征数据。深度学习不是一种特定的方法，而是用于学习复杂神经网络的一组各种技术，例如网络体系结构，激活函数，优化方法，损失函数，评估指标，预处理，数据扩充和许多启发式方法。而深度学习中最具代表性的就是LeCun Y等（1999）提到的卷积神经网络（CNN），这个方法发展了几十年，现在有一些比较好的成果，如Krizhevsky A等（2012）提出的Alex Net、Simonyan K等（2015）提出的VGGNet、Szegedy C等（2015）提出的GoogleNet和He K等（2016）提出的ResNet，这些方法都有很广泛的应用。近些年来深度学习方法发展迅猛，但是在不同的数据集上，其表现能力也不一定都比传统机器学习方法好，如Eulenberg P等（2017）在对F1 巨集细胞作重建时，发现增强算法表现更好一些。所以应该根据不同的数据集而选择不同的机器学习方法。

机器学习对于细胞的分析有以下几点应用：细胞图像处理、细胞图像分割、细胞图像分类、细胞目标跟踪和细胞图像重建。

Su H等（2015）提出了一个细胞检测和分割算法，其使用稀疏重建和堆叠去噪自动编码器，首个使用自动编码器（使用自适应词典和模板）来减少噪声并重建组织病理学图像中的细胞核。Rivenson Y等（2018）提出了一种经过改进的U-Net架构（具有跳过连接的自动编码器），以计算方式提高用40×/ 0.95NA物镜采集的组织样本的明场图像的分辨率，如实现与使用100×/ 1.4NA油浸物镜获取的图像等效的分辨率。

Ronneberger O等（2015）提出的U-Net方法和Valen V等（2016）提出的DeepCell方法，可将一整张图像分割为多张只含单一细胞的图像，这两种方法分割出的细胞在细胞分类和细胞目标跟踪中都表现不错。

对于细胞图像分类，Felix B等（2017）提出将荧光标记的细胞作为真值，以此训练分类器进行细胞分类，而Ciurte A等（2018）则提出了一种无需对细胞染色的自动化的检测方法，Pärnamaa T等（2017）也将有监督学习方法应用于荧光标记的细胞分类，Rumetshofer E等（2019）也发现机器学习算法在高通量荧光显微成像中对特定蛋白质的识别表现突出。

细胞的目标跟踪则用于分析活细胞的动态移动，Kraus O等（2016）提出一种有监督学习方法从相衬图像中可靠地分割细胞核的荧光图像以及无序荧光标记的哺乳动物的细胞质图像，从而同时鉴定不同的哺乳动物生长类型。Mathis A等（2018）提出了一个用于细胞目标跟踪的工具包，可用于分析神经科学实验中的细胞行为。

图像重建涉及到生成模型，如生成对抗网络（GAN），Kingma D等（2014）提出了一种自编码网络、Anand N等（2018）搭建了一套生成模型，用于快速设计蛋白质的结构。

机器学习方法在流式细胞仪中的分析应用，按数据集的类型来分，可分为对细胞进行特征提取后进行训练的和直接对采集数据进行训练的，按数据的采集方式来分，可分为用CCD阵列像素传感器采集细胞的和用单像素传感器采集细胞的。机器学习在流式细胞分析中的应用，以有监督学习为例，涉及几个方面：构建训练数据集、根据数据集训练模型、给模型投喂新数据进行预测分析。

而用于细胞的分析有实时和非实时之分，由于单位时间细胞流量都很大，所以实时的细胞分析，对于算法的复杂度和硬件的算力都有较高要求。Ota S等（2018）在用单像素传感器采集细胞时，没有进行细胞重建，直接用原始信号训练分类器，得到了不错的结果。Gu Y等（2019）则提出了一种实时的基于实时图像处理和机器学习的细胞计数和分类方法。Girault M等（2017）研制了一种基于液滴细胞图像识别的无标签微流体液滴分选系统，这种集成的液滴成像细胞分选方法可以提供一种互补的分选方法，可以在不染色的情况下从混合细胞中分离出单一的靶细胞，具有较高的准确率。Nitta N等（2018）基于多层的卷积神经网络，集成了高通量细胞显微镜、聚焦和分类在一个混合的软硬件数据管理基础设施，使数据采集、数据处理、决策和驱动能够实时自动化操作。Gu et等（2019）报道了根据图像进行指导细胞分选和分类系统的另一项发展，其基于目标分子和颗粒的定位分析实现了细胞分选。对原始数据直接进行训练可以减少一些因为特征提取或者细胞重建中带来的复杂度，同时也减少了额外操作带来的信息损失。Doan M等（2019）描述了一种直接用深度学习分析细胞仪采集的原始数据的过程：直接用原始数据进行训练，预测的时候也不对原始数据采取过多处理。Felix B等（2017）直接将明场成像得到的原始细胞图像进行训练，用荧光标记作为真值，训练得到一个有效的有监督分类器。Adachi H等（2019）也发现不对细胞图像进行重建，而直接使用采集的原始一维信号，可以进行更快的以至于实时的细胞形态分析。Takahashi K等（2004）利用微流体和静电力的优势，开发了一种新的基于单细胞培养的非破坏性片上细胞分选系统，这在细胞纯化实验中可被有效使用。Li Y等（2019）进一步演示了深度卷积网络的应用，以直接处理单个像素成像器的一维时间序列信号，而无需进行图像重建以实现实时处理。这种方法的显著技术优势是，省去了诸如图像重建和取之类的耗时步骤，大大减少了细胞形态分析的处理，Ota S等（2018）的研究成果中，这个处理时间低于10us，与进行图像重建的方法相比，速度提升了几个数量级，这就缓解了流式细胞仪的高通量与实时处理之间的矛盾。Ugawa等（2019）使用无局部标记的压缩波形和通过荧光染色的生物标记细胞训练了一个模型，可以直接从原始采集的一维波形中预测细胞，而消除了荧光染色的影响，证明这项技术能够对多种类型和状态的细胞进行高速无标签分类，可应用在细胞制造和再生医学中。

## 研究目的

探索鲁棒的、泛化能力强的机器学习算法，使之能应用于流式细胞仪的细胞分析，如细胞图像分割、细胞分选、细胞图像重建等方面。

## 预期成果

1. 实现快速响应的机器学习算法用于细胞分析，以适应流式细胞仪的高通量
2. 实现鲁棒的、低误差、泛化能力强的机器学习算法，用于流式细胞分析
3. 将稳定的机器学习算法落地，尽可能与实际产品耦合

## 研究方案

1. 对最新的常用的机器学习算法进行调研，分析各种算法的适用性
2. 基于实验室使用流式细胞仪采集的细胞数据集，选择合适的机器学习算法进行训练，并进行性能评估和对比
3. 对于表现较好的几种算法进行深入研究，尝试用不同数据集进行训练，以评估泛化能力
4. 在保证算法高精度、较好泛化能力的情况下，尽可能对机器学习算法的实施过程进行优化，避免不必要的计算量
5. 结合流式细胞的特性，对已有机器学习算法进行修改，只为了更加适应流式细胞的分析

## 研究方法

主要是研究阶段前期的调研和理论分析、中期的尝试调参、后期的性能分析和优化

## 关键难点

1. 一般而言，算法的复杂度与精确度是相矛盾的。可能解决方案：优化算法处理流程，避免不必要的处理。如在文献综述中说到，一些研究者将采集得到的信号先恢复图像，再进行特征提取，才开始训练学习器，而另一些研究者直接对恢复后的图像进行模型训练，或者不恢复图像，直接对采集得到的原始信号进行分析。这样直接就省去了恢复图像的时间和特征提取的时间。
2. 从众多的算法中选择适合流式细胞分析的几种算法不易。可能解决方案：前期大量调研，分析已有文献中，各个研究者们对于不同应用场景不同数据集，采取的不同算法，进而分析各个算法的适用性。并投入时间对各种算法进行尝试，采取合理的方式进行评估。
3. 机器学习算法涉及很多参数调节，且泛化能力不太好把握。可能解决方案：深入把握算法的原理，了解各个参数含义，从而有方向地调节。并且通过交叉验证或者自适应调参等方法，找到合适的且泛化能力不错的参数，通过由于数据集过大且算法复杂度过高，这个调参环节会耗费大量时间。
4. 对已有机器学习算法的修改较难。已有机器学习算法往往都是经过多重考验，而留下来的经典模型。可能解决方案：从应用场景出发，如对于流式细胞分析，就结合流式细胞的相关特性进行修改，不要求最终的算法对其他领域有较好的适用性，只要能在流式细胞分析方面表现良好即可。

## 总体日程安排

## 预计答辩时间