(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111652095 A
(43)申请公布日 2020.09.11

(21)申请号 202010438283.3

(22)申请日 2020.05.21

(71)申请人 骏实生物科技(上海)有限公司
地址 201108 上海市闵行区曹建路161号

(72)发明人 金炜翔 温冬 李基

(74)专利代理机构 北京市盈科律师事务所 11344
代理人 陈晨 王津

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G16H 15/00(2018.01)

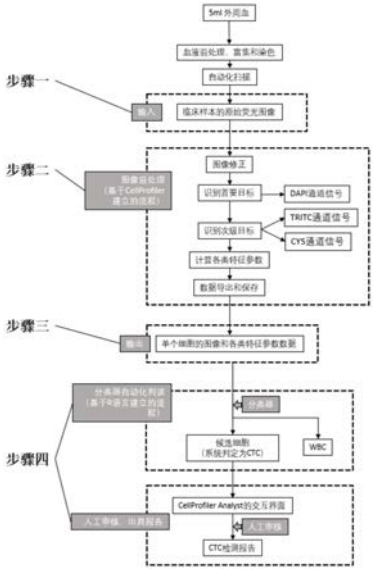
权利要求书3页 说明书23页 附图9页

(54)发明名称

一种基于人工智能的CTC图像识别方法和系统

(57)摘要

本发明公开了一种基于人工智能的CTC图像识别方法和系统,所述方法使用预先建立的分类器进行CTC的自动化判读,包括以下步骤:输入临床样本的原始荧光图像;对所述临床样本的原始荧光图像进行图像前处理;输出临床样本单个细胞的图像和特征参数;采用预先建立的分类器自动化判读,筛选疑似CTC的细胞作为候选细胞,对候选细胞进行审核后,出具CTC报告,本发明还公开了分类器建立的方法和系统,使用人工标注的方法区分CTC,筛选CTC和WBC的特征参数,基于多种机器学习的算法建立多个初步分类器,并优化获得的分类器;本发明采用真实的临床循环肿瘤细胞建立训练集,系统性地对形态学参数进行筛选和验证,保证优化的分类器识别CTC的性能。



CN 111652095 A

1.一种基于人工智能的CTC图像识别方法,使用预先建立的分类器进行CTC的自动化判读,其特征在于,包括以下步骤:

步骤一、输入临床样本的原始荧光图像;

步骤二、对所述临床样本的原始荧光图像进行图像前处理;

步骤三、输出临床样本单个细胞的图像和特征参数;

步骤四、采用预先建立的分类器自动化判读,筛选疑似CTC的细胞作为候选细胞,对候选细胞进行审核后,出具CTC报告。

2.如权利要求1所述的基于人工智能的CTC图像识别方法,其特征在于,所述分类器的建立方法包括如下步骤:

步骤I、输入临床样本的原始荧光图像;

步骤Ⅱ、对所述临床样本的原始荧光图像进行图像前处理;

步骤Ⅲ、输出临床样本单个细胞的图像和特征参数,人工判断临床样本中是否含有CTC,如含有CTC即为临床CTC样本,继续步骤IV,如人工判断临床样本中没有CTC,则换一个样本重复步骤I;

步骤Ⅳ、人工标注CTC样本中的CTC和WBC,筛选CTC和WBC的特征参数,作为训练集;基于多种机器学习的算法建立多个初步分类器;通过参数调优、交叉验证、平行比较优化分类器。

3.如权利要求1或2所述的基于人工智能的CTC图像识别方法,其特征在于,所述图像前处理包括如下步骤:

(1)图像修正;

(2)识别首要目标;

(3)识别次级目标;

(4)计算各类特征参数;

(5)数据导出和保存。

4.如权利要求2所述的基于人工智能的CTC图像识别方法,其特征在于,所述训练集包括超过1700个CTC图像、超过13000个WBC图像,和超过200个特征参数。

5.如权利要求2所述的基于人工智能的CTC图像识别方法,其特征在于,步骤Ⅳ具体包括:

(1)数据中心化和归一化;

(2)基于每个特征参数的散点图,手动筛选能显著区分两种类别细胞的特征参数;

(3)剔除高度相关的特征参数(cutoff>.75);

(4)计算特征参数重要性(RFE),并最终确认用于模型建立的特征参数集合;

经过上述步骤后获得新训练集。

6.如权利要求5所述的基于人工智能的CTC图像识别方法,其特征在于,步骤Ⅳ还包括:在新训练集内,分别使用多种监督式机器学习算法以及融合模型算法、多种不平衡训练集的前处理方法、多种评估方法,进行交叉验证,优化参数,建立初步分类器。

7.如权利要6所述的基于人工智能的CTC图像识别方法,其特征在于,所述监督式机器学习算法包括K-Nearest Neighbors(KNN)、Stochastic Gradient Boosting(GBM)、AdaBoost Classification Trees(ADABOOST)、Support Vector Machines(SVM)、Random

Forest(RF)、Naïve Bayes(NB)、Extreme Gradient Boosting(XGB),所述融合模型算法为多种监督式机器学习算法融合在一起的算法。

8.如权利要求6所述的基于人工智能的CTC图像识别方法,其特征在于,所述不平衡训练集的前处理方法包括:Original、Up-sampling、Down-sampling、SMOTE、ROSE。

9.如权利要求6所述的基于人工智能的CTC图像识别方法,其特征在于,所述评估方法包括:ROC和PR。

10.如权利要求6所述的基于人工智能的CTC图像识别方法,其特征在于,步骤Ⅳ还包括:对初步分类器的性能评估和泛化能力测试,优化分类器,所述性能进行评估指标包括AUC、F1 score、Accuracy、Precision、Recall、TPR、FPR,所述泛化能力评估指标包括:Positive sample concordance,CTC concordance,Screening efficiency。

11.如权利要求3所述的基于人工智能的CTC图像识别方法,其特征在于,所述图像前处理步骤中,识别首要目标为识别DAPI通道有信号的目标;识别次级目标为在识别首要目标的基础上分别识别TRITC通道有信号的目标和CY5通道有信号的目标。

12.如权利要求3所述的基于人工智能的CTC图像识别方法,其特征在于,所述计算各类特征参数,包括计算首要目标和次级目标的形态学参数、各通道荧光信号强度;所述形态学参数包括大小&形状(Area&Shape)、信号强度(Intensity)、表面结构(Texture)、相关性(Correlation)。

13.一种基于人工智能的CTC图像识别系统,其特征在于,包括输入模块、图像前处理模块、输出模块、自动化判读模块;

所述输入模块用于输入临床样本的原始荧光图像;

所述图像前处理模块用于对临床样本的原始荧光图像进行处理后获得单个细胞的图像和特征参数;

所述输出模块用于输出临床样本中单个细胞的图像和特征参数,

所述自动化判读模块用于识别CTC细胞,使用预先建立的分类器进行筛选,筛选出疑似CTC的细胞作为候选细胞,对CTC候选细胞进行人工审核后,出具CTC检测报告。

14.根据权利要求13所述的基于人工智能的CTC图像识别系统,其特征在于,所述自动化判读模块包括初审模块和终审模块,所述初审模块使用预先建立的分类器进行筛选,筛选出疑似CTC的细胞作为候选细胞;所述终审模块采用专业人员判断候选细胞是否确实为CTC,确认后出具CTC检测报告。

15.根据权利要求13所述的基于人工智能的CTC图像识别系统,其特征在于,所述分类器通过分类器建立系统建立,所述分类器建立系统包括输入模块二、图像前处理模块二、输出模块二、分类器建立模块;

所述输入模块二用于输入临床CTC样本的原始荧光图像;

所述图像前处理模块二用于对临床CTC样本的原始荧光图像进行处理后获得单个细胞的图像和特征参数,作为训练集;

所述输出模块二用于输出临床CTC样本中单个细胞的图像和特征参数;

所述分类器建立模块用于建立并优化分类器,人工标注CTC和WBC,筛选CTC和WBC的特征参数;基于多种机器学习的算法建立多个初步分类器;通过参数调优、交叉验证、平行比较优化分类器。

16.根据权利要求13所述的基于人工智能的CTC图像识别系统，其特征在于，所述图像前处理模块包括图像修正模块、识别首要目标模块、识别次级目标模块、计算各类特征参数模块、数据导出和保存模块；所述图像修正模块用于修正不均一光照强度导致的不均一的图像信号和背景；所述识别首要目标模块用于识别DAPI通道有信号的目标；所述识别次级目标模块用于在识别DAPI通道有信号的目标基础上，分别识别TRITC通道有信号的目标和CY5通道有信号的目标，并分别获得单个细胞图像；所述计算各类特征参数模块用于计算单个细胞的形态学参数、各通道荧光信号强度参数；所述数据导出和保存模块用于导出和保存单个细胞的图像和特征参数。

17.根据权利要求15所述的基于人工智能的CTC图像识别系统，其特征在于，所述图像前处理模块二包括图像修正模块二、识别首要目标模块二、识别次级目标模块二、计算特征参数模块二、数据导出和保存模块二；所述图像修正模块二用于修正不均一光照强度导致的不均一的图像信号和背景；所述识别首要目标模块二用于识别DAPI通道有信号的目标；所述识别次级目标模块二用于在识别DAPI通道有信号的目标基础上，分别识别TRITC通道有信号的目标和CY5通道有信号的目标，并分别获得单个细胞图像；所述计算特征参数模块二用于计算单个细胞的形态学参数、各通道荧光信号强度参数；所述数据导出和保存模块用于导出和保存单个细胞的图像和特征参数。

一种基于人工智能的CTC图像识别方法和系统

技术领域

[0001] 本发明涉及循环肿瘤细胞识别技术领域，具体而言，涉及一种基于人工智能的荧光细胞图像识别技术，用于临床循环肿瘤细胞自动化检测。

背景技术

[0002] 循环肿瘤细胞(Circulating Tumor Cell,CTC)指自发或因诊疗操作由实体瘤或转移灶释放进入外周血循环的肿瘤细胞，恶性肿瘤的侵袭转移是患者复发转移的关键环节，常导致肿瘤治疗失败，危及患者生命。由于入侵发生在肿瘤早期阶段，此时血液中的循环肿瘤细胞数量非常稀少，这意味着准确有效的循环肿瘤细胞识别与分析对于尽早地确诊肿瘤并采取有效的治疗措施具有重要意义。

[0003] 通过免疫荧光染色来识别标有肿瘤标志物的CTC是一种常见的临床CTC检测方法，这种方法除了依赖于全自动扫描的荧光显微镜之外，还需要对扫描得到的荧光图像进行判读来识别临床样本中的 CTC。

[0004] 现在临床CTC检测主要还是依赖专业人员对荧光图像进行人工判读，这种判读方式带来的问题包括不同操作人员间的判读误差、不同样本间的判读误差，以及判读效率不高等。现有的CTC判读方法还包括基于各形态学参数threshold的图像筛选，根据经验值，对CTC 的各形态学参数进行限定，设定固定threshold来筛选CTC；这个方式最大缺点是临床样本中CTC的形态各异，基于经验值设定threshold 的方式CTC筛选的灵敏度和特异性均很差。基于机器学习的CTC图像识别方法也在现有文献中有过报道，但现有方法还存在以下缺点：

[0005] (1)现有CTC图像识别方法中涉及的多个形态学参数没有系统性地筛选和比较，不同形态学参数集合会最终影响到识别的灵敏度、特异性等；

[0006] (2)现有CTC图像识别方法中分类器建立所使用的训练集不是来自于真实的临床CTC，而是采用的掺入肿瘤细胞系细胞在血液中来模拟CTC，基于这样的训练集训练出来的分类器，在临床CTC检测和识别的性能并不可靠；

[0007] (3)现有CTC图像识别方法中训练集由于是来自于肿瘤细胞系细胞，所以训练集的细胞比例(CTC和WBC的比例)并不能反映真实情况(实际经过富集后的临床样本中CTC与WBC比例为$1:10^3$)，这样的训练集训练出来的分类器，无法保证对实际临床样本中极度不平衡的样本比例进行有效地识别；

[0008] (4)现有CTC图像识别方法只是简单地比较了几种常见的机器学习算法建立的分类器，对于CTC图像识别方法的性能的优化有限；

[0009] (6)现有CTC图像识别方法的泛化能力评估只是在极少量的临床样本中进行，并不能有效反映CTC图像识别方法的实际临床性能；

[0010] 本专利通过开发基于机器学习的临床CTC图像识别方法和系统来完成对临床CTC图像高效、准确的判读，可以让机器通过对过往的CTC图像数据自我学习，来不断修正改善自身性能，可以有效地保证临床CTC图像识别的准确性，再由专业人员对筛选后的CTC图像

进行终审来确认检测结果。

发明内容

[0011] 本发明提供了一种基于人工智能的CTC图像识别方法和系统。

[0012] 本发明中用到的技术名词和含义如下：

[0013] CTC:循环肿瘤细胞,Circulating Tumor Cell,指自发或因诊疗操作由实体瘤或转移灶释放进入外周血循环的肿瘤细胞。

[0014] WBC:白细胞,White Blood Cell,是无色、球形、有核的血细胞,白细胞的主要功能是防卫作用。

[0015] 临床CTC样本:检测到至少一个CTC的临床样本。

[0016] 分类器:数据挖掘中对样本进行分类的方法的统称。

[0017] DAPI:是一种蓝色荧光DNA染色剂,与dsDNA的AT区结合后,荧光增强约20倍。它被紫罗兰色(405nm)激光线激发,通常用作荧光显微镜,流式细胞仪和染色体染色中的核复染。

[0018] TRITC:是罗丹明染料的高性能衍生物,经活化可轻松可靠地标记用作荧光探针的抗体,蛋白质和其他分子。

[0019] Cy5:是一种明亮的,远红色荧光染料,具有激发光,非常适合 633nm或647nm激光线;用于标记蛋白质和核酸偶联物。

[0020] TP:True Positive,真阳,预测为正样本,实际也为正样本。

[0021] FN:False Negative,假阴,预测为负样本,实际为正样本。

[0022] FP:False Positive,假阳,预测为正样本,实际为负样本。

[0023] TN:True Negative,真阴,预测为负样本,实际为负样本。

[0024] Accuracy:准确率,正确预测正负样本例数/样本总数,accuracy＝(TP+TN)/(TP+FN+FP+TN)。

[0025] Precision:精确率,正确预测正样本例数/预测正样本总数, precision＝TP/(TP+FP)。

[0026] Recall:召回率,正确预测正样本例数/实际正样本总数, recall＝TP/(TP+FN)。

[0027] F1 score:F1＝2*(prescision*recall)/(precision+recall),精确率和召回率的调和值,更接近两个数较小的那个,所以精确率和召回率接近时,F1值最大。

[0028] F2 score:F2＝5*(precision*recall)/(4*precision+recall),相比于F1 score,F2 score中,召回率权重大于精确率。

[0029] TPR:True Positive Rate,真阳性率,正确预测正样本例数/实际正样本总数,TPR＝TP/(TP+FN)。

[0030] FPR:False Positive Rate,假阳性率,错误预测为正样本例数/实际负样本总数,FPR＝FP/(FP+TN)

[0031] PR:由精确率Precision(Y轴)和召回率Recall(X轴)组成的曲线。

[0032] ROC:Receiver Operating Characteristic,度量分类器好坏的一个标准,其主要分析工具是一个画在二维平面上的曲线——ROC曲线。平面的横坐标是FPR,纵坐标是TPR。对某个分类器而言,可以根据其在测试样本上的表现得到一个TPR和FPR点对。这样,此分类

器就可以映射成ROC平面上的一个点。调整这个分类器分类时候使用的阈值,可以得到一个经过(0,0),(1,1)的曲线,这就是此分类器的 ROC曲线。一般情况下,这个曲线都应该处于(0,0)和(1,1)连线的上方,ROC曲线下方的那部分面积越大,分类器效果越好。

[0033] AUC:Area Under Curve,是一种用来度量分类器好坏的一个数值化的标准,AUC越大,分类器分类效果越好。AUC为ROC曲线下方的那部分面积的值。

[0034] CTC concordance:图像识别系统识别到的CTC与人工判读CTC 结果的一致性,CTC concordance=100%*(图像系统识别与人工判读的同一CTC的数量/人工判读得到的CTC数量),CTC concordance 值越高,图像识别系统识别CTC灵敏度越高。

[0035] Positive sample concordance:图像识别系统识别到的阳性样本(≥1 个CTC)与人工判读结果的一致性,positive sample concordance=100%*(人工判读为阳性,图像识别系统判读也为阳性的样本数量/人工判读的阳性样本数量),positive sample concordance 值越高,图像识别系统识别阳性样本灵敏度越高。

[0036] Screening efficiency:100%*(图像识别系统排除掉的非CTC数量/该样本中的细胞总数),screening efficiency值越高,图像识别系统识别CTC特异性越高。

[0037] CellProfiler:用于基于图像的复杂屏幕的数据探索和分析软件,是一种开源软件,用于交互式探索和分析多维数据,特别是来自基于图像的高通量实验数据。该系统可以对基于图像的屏幕进行交互式数据探索,并可以对复杂的表型进行自动评分,这些复杂的表型需要每个单元结合多个测量特征。

[0038] 特征参数相关性计算:采用Pearson相关性分析,计算任意2个特征参数的Pearson相关系数,本专利中相关系数大于.75(cutoff>.75) 的2个参数属于高度相关。

[0039] 特征参数重要性(RFE):即递归式特征消除法,recursive feature elimination,在每轮迭代中,选取不同特征子集,进行模型训练并评估模型,通过计算其决策系数之和,最终得到不同特征的重要程度,然后保留最佳的特征组合。

[0040] 本发明中,血液前处理、CTC富集和染色制作为原始荧光图像的方法为现有技术,具体如下:

[0041] 将5ml外周血液与红细胞特异性抗体和白细胞特异性抗体组合进行孵育,使全血样本中红细胞和白细胞耦联在一起,再通过密度梯度离心方法使血液中细胞根据自己的密度达到分离分层的目的,经过梯密度离心的血液样本会分成4层:从上往下分别为血浆、单核细胞、密度梯度离心液以及红细胞和白细胞;CTC做为单核细胞会处于单核细胞层,将单核细胞层提取出来,达到血液中CTC的富集目的;

[0042] 单核细胞层被提取出来后,进行清洗,然后进行后续的免疫荧光染色流程,包括固定、透化、荧光抗体染色等一系列流程,在染色流程中使用的荧光抗体包括细胞核染料DAPI(DAPI通道)、特异性识别CTC的染色试剂EpCAM/CK(TRITC通道)、特异性识别WBC 的染色试剂CD45(CY5通道),最后剩余约200ul的样本,样本中会包含约0-100个CTC、10^3-10^4个白细胞(WBC)、血小板以及杂质等。经过上述荧光抗体染色后,CTC的判定标准是DAPI+且 TRITC+且CY5-,WBC的判定标准是DAPI+且CY5+;其中“+”指有荧光信号,“-”指没有荧光信号;

[0043] 染好色的样本会转入96孔板中的其中一个孔内,然后进行扫描,扫描可以采用商业化扫描仪ThermoFisher CX5,由于孔底的面积大于单次扫描仪拍照面积,所以扫描仪需要自动移动载物台进行拍照,通过对不同位置进行拍照,然后再进行后期拼接,实现对单个

孔内的样本完整扫描;由于是免疫荧光染色样本,所以在每个区域拍照时,会切换不同荧光通道进行拍照,再进行后期叠加。

[0044] 染好色的样本通过扫描后会生成169组图像覆盖整个样本区域,每组图像包含来自DAPI、TRITC、CY5三个荧光通道的图像,即为本发明中输入图像前处理模块的临床样本的原始荧光图像。

[0045] 本发明的目的可以通过以下技术方案来实现:

[0046] 一种基于人工智能的CTC图像识别方法,使用预先建立的分类器进行CTC的自动化判读,包括以下步骤:

[0047] 步骤一、输入临床样本的原始荧光图像;

[0048] 步骤二、对所述临床样本的原始荧光图像进行图像前处理;

[0049] 步骤三、输出临床样本单个细胞的图像和特征参数;

[0050] 步骤四、采用预先建立的分类器自动化判读,筛选疑似CTC的细胞作为候选细胞,对候选细胞进行审核后,出具CTC报告。

[0051] 本发明上述技术方案中,分类器的建立方法包括如下步骤:

[0052] 步骤I、输入临床样本的原始荧光图像;

[0053] 步骤Ⅱ、对所述临床样本的原始荧光图像进行图像前处理;

[0054] 步骤Ⅲ、输出临床样本单个细胞的图像和特征参数,,人工判断临床样本中是否含有CTC,如含有CTC即为临床CTC样本,继续步骤IV,如人工判断临床样本中没有CTC,则换一个样本重复步骤I;

[0055] 步骤Ⅳ、人工标注临床CTC样本中的CTC和WBC,筛选CTC 和WBC的特征参数,作为训练集;进一步的,所述训练集包括超过 1700个循环肿瘤细胞图像、超过13000个白细胞图像,和超过200个特征参数。;基于多种机器学习的算法建立多个初步分类器;通过参数调优、交叉验证、平行比较优化分类器。

[0056] 进一步的,步骤Ⅳ具体包括:

[0057] (1)数据中心化和归一化;

[0058] (2)基于每个特征参数的散点图,手动筛选能显著区分两种类别细胞的特征参数;

[0059] (3)剔除高度相关的特征参数(cutoff>.75);

[0060] (4)计算特征参数重要性(RFE),并最终确认用于模型建立的特征参数集合;

[0061] 经过上述步骤后获得新训练集。其中,(cutoff>.75)是指对每两个特征参数计算Pearson相关系数,相关系数值大于.75的两个参数认为是高度相关参数,因此会剔除其中一个参数,特征参数重要性 (RFE)的计算方式为每轮迭代中,选取不同特征子集,进行模型训练并评估模型,通过计算其决策系数之和,最终得到不同特征的重要程度,然后保留最佳的特征组合。

[0062] 进一步的,步骤Ⅳ还包括:在新训练集内,分别使用多种监督式机器学习算法以及融合模型算法、多种不平衡训练集的前处理方法、多种评估方法,进行交叉验证,优化参数,建立初步分类器。

[0063] 进一步的,所述监督式机器学习算法采用了总共8种:7种监督式机器学习算法包括K-Nearest Neighbors(KNN)、Stochastic Gradient Boosting(GBM)、AdaBoost Classification Trees(ADABOOST)、Support Vector Machines(SVM)、Random Forest

(RF)、Naïve Bayes(NB)、Extreme Gradient Boosting(XGB),还有1种融合模型算法(Stack),即 7种监督式机器学习算法融合在一起的算法。

[0064] 进一步的,所述不平衡训练集的前处理方法包括5种:Original、Up-sampling、Down-sampling、SMOTE、ROSE。

[0065] 进一步的,所述评估方法包括2种:ROC和PR。

[0066] 在新的训练集内,使用上述7种监督式机器学习的算法以及1个融合模型算法(共8种算法)、5种不平衡训练集的前处理方法、2种测量指标评估方法,进行交叉验证,优化参数,最终生成8*5*2=80个分类器。

[0067] 进一步的,步骤Ⅳ还包括:用训练集对分类器的性能进行评估,分析比较了AUC、F1 score、Accuracy、Precision、Recall、TPR、FPR 参数在80个分类器中的表现,,针对不平衡数据集而言(不平衡数据集是指CTC数量在训练集中占比小于10%),在各项指标均表现良好情况下,F1 score、Recall、TPR指标更关注于识别CTC的灵敏度,所以,F1 score、Recall、TPR数值越高,系统识别CTC的灵敏度越高。

[0068] 进一步的,步骤Ⅳ还包括:用建立调优好的80个分类器在临床样本中进行泛化能力测试,测试范围涵盖了200例临床CTC检测样本 (200例临床CTC检测样本中人工判读结果约含1000个CTC)。基于 CTC自动化判读需求,泛化能力评估指标包括:CTC concordance,Positive sample concordance,Screening efficiency。针对于临床CTC检测应用,首先,期望分类器的Positive sample concordance达到100%,在满足前述条件的情况下,尽可能选择CTC concordance或Screening efficiency较高的。

[0069] 综上,优化的分类器的筛选原则为:F1 score、Recall、TPR数值尽量高,Positive sample concordance达到100%,在前述条件都满足的情况下,优选CTC concordance高于90%的,或者Screening efficiency高于95%的,如有多个符合条件的分类器,根据实际场景选择其中一个,例如,可以选择几个分类器中CTC concordance最高的一个。

[0070] 本发明上述技术方案中,所述图像前处理包括如下步骤:

[0071] (1)图像修正;

[0072] (2)识别首要目标;

[0073] (3)识别次级目标;

[0074] (4)计算各类特征参数;

[0075] (5)数据导出和保存。

[0076] 进一步,所述图像前处理步骤中,识别首要目标为识别DAPI通道有信号的目标;识别次级目标为在识别首要目标的基础上分别识别 TRITC通道有信号的目标和CY5通道有信号的目标。

[0077] 进一步的,所述图像前处理步骤中,识别首要目标为识别DAPI 通道有信号的目标;识别次级目标为在识别首要目标的基础上分别识别TRITC通道有信号的目标和CY5通道有信号的目标。

[0078] 进一步的,所述计算各类特征参数,包括计算首要目标和次级目标的形态学参数、各通道荧光信号强度;所述形态学参数包括大小& 形状(Area&Shape)、信号强度(Intensity)、表面结构(Texture)、相关性(Correlation)。

[0079] 进一步的,自动化判读是基于R语言的libraries建立的。

[0080] 本发明的另一种技术方案为：

[0081] 一种基于人工智能的CTC图像识别系统，包括输入模块、图像前处理模块、输出模块、自动化判读模块；

[0082] 所述输入模块用于输入临床样本的原始荧光图像；

[0083] 所述图像前处理模块用于对临床样本的原始荧光图像进行处理后获得单个细胞的图像和特征参数；

[0084] 所述输出模块用于输出临床样本中单个细胞的图像和特征参数；

[0085] 所述自动化判读模块用于识别CTC细胞，使用预先建立的分类器侵权进行筛选，筛选出疑似CTC的细胞作为候选细胞，对CTC候选细胞进行审核后；出具CTC检测报告。

[0086] 进一步的，所述图像前处理模块包括图像修正模块、识别首要目标模块、识别次级目标模块、计算各类特征参数模块、数据导出和保存模块；所述图像修正模块用于修正不均一光照强度导致的不均一的图像信号和背景；所述识别首要目标模块用于识别DAPI通道有信号的目标；所述识别次级目标模块用于在识别DAPI通道有信号的目标基础上，分别识别TRITC通道有信号的目标和CY5通道有信号的目标，并分别获得单个细胞图像；所述计算各类特征参数模块用于计算单个细胞的形态学参数、各通道荧光信号强度参数；所述数据导出和保存模块用于导出和保存单个细胞的图像和特征参数。

[0087] 进一步的，所述自动化判读模块包括初审模块和终审模块，所述初审模块使用预先建立的并且优化后的分类器进行筛选，筛选出疑似 CTC的细胞作为候选细胞；所述终审模块采用专业人员判断候选细胞是否确实为CTC，确认后出具CTC检测报告。

[0088] 本发明上述技术方案中，所述分类器通过分类器建立系统建立和优化，所述分类器建立系统包括输入模块二、图像前处理模块二、输出模块二、分类器建立模块；

[0089] 所述输入模块二用于输入临床CTC样本的原始荧光图像；

[0090] 所述图像前处理模块二用于对临床CTC样本的原始荧光图像进行处理后获得单个细胞的图像和特征参数，作为训练集；

[0091] 所述输出模块二用于输出临床CTC样本中单个细胞的图像和特征参数；

[0092] 所述分类器建立模块用于建立并优化分类器，人工标注CTC和 WBC，筛选CTC和WBC的特征参数；基于多种机器学习的算法建立多个初步分类器；通过参数调优、交叉验证、平行比较优化分类器。

[0093] 本发明上述技术方案中，所述图像前处理模块二包括图像修正模块二、识别首要目标模块二、识别次级目标模块二、计算各类特征参数模块二、数据导出和保存模块二；所述图像修正模块二用于修正不均一光照强度导致的不均一的图像信号和背景；所述识别首要目标模块二用于识别DAPI通道有信号的目标；所述识别次级目标模块二用于在识别DAPI通道有信号的目标基础上，分别识别TRITC通道有信号的目标和CY5通道有信号的目标，并分别获得单个细胞图像；所述计算各类特征参数模块二用于计算单个细胞的形态学参数、各通道荧光信号强度参数；所述数据导出和保存模块二用于导出和保存单个细胞的图像和特征参数。

[0094] 本发明具有以下有益效果：

[0095] (1)本发明系统性地对形态学参数进行筛选和验证，可以保证优化的分类器识别CTC的性能；

[0096] (2)现有技术中CTC图像识别流程建立使用的训练集采用的掺入肿瘤细胞系细胞在血液中来模拟CTC,本发明采用的是申请人建立的超过1000个临床CTC的建立训练集,验证和测试结果更能直观反映实际临床检测中的性能;

[0097] (3)现有技术中CTC图像识别流程建立使用的训练集因为采用的掺入细胞模拟实验,未涉及不平衡训练集问题,本发明针对临床 CTC检测建立的训练集由于真实反映CTC在血液中的稀有情况,所以系统性地对不平衡训练集的不同处理方式进行了测试和评估,寻找了最优的处理方式;

[0098] (4)本发明系统性比较了80种分类器的性能,远远超过现有技术比较分类器种类的数量;

[0099] (5)现有技术中使用极少量的临床样本对分类器进行泛化能力评估,本发明对前述分类器进行了大规模临床样本的泛化能力评估,检测时间跨度半年,样本量约200例,人工判读的临床CTC数量约 1000个,更真实反映分类器的临床使用性能。

附图说明

[0100] 图1为本发明CTC图像识别方法的流程图;

[0101] 图2为本发明分类器建立方法的流程图;

[0102] 图3为本发明CTC图像识别系统的示意图;

[0103] 图4为本发明分类器建立系统的示意图;

[0104] 图5为原始荧光图像的示意图;

[0105] 图6为单个细胞的图像示意图;

[0106] 图7为分类器在临床样本中的泛化能力评估示意图,包括:图7A CTC concordance计算结果示意图,图7B Positive sample concordance 计算结果示意图,图7C Screening efficiency计算结果示意图;

[0107] 图8为从候选细胞生成CTC报告的流程示意图。

[0108] 附图标记:

[0109] 输入模块1,图像处理模块2,图像修正模块201、识别首要目标模块202、识别次级目标模块203、计算各类特征参数模块204、数据导出和保存模块205,输出模块3,自动化判读模块4,初审模块401,终审模块402,输入模块二5,分类器图像处理模块6,图像修正模块601、识别首要目标模块602、识别次级目标模块603、计算各类特征参数模块604、数据导出和保存模块605,输出模块二7,分类器建立模块8。

具体实施方式

[0110] 结合附图所示,本发明的技术方案作进一步的描述:

[0111] 本发明的第一个实施例如图1所示,一种基于人工智能的CTC图像识别方法,使用预先建立的分类器进行CTC的自动化判读,包括以下步骤:

[0112] 步骤一、输入临床样本的原始荧光图像;临床样本为外周血经过 CTC富集处理,扫描后获得;

[0113] 步骤二、对临床样本的原始荧光图像进行图像前处理,图像前处理是基于Cellprofiler建立的流程;图像前处理包括步骤:(1)图像修正,修正不均一光照强度导致

的不均一的图像信号和背景;(2)识别首要目标,识别DAPI通道有信号的目标,并将图片切割成以该目标为中心的图像,即关于首要目标的单个细胞的图像;(3)识别次级目标;在识别DAPI通道有信号的目标基础上,分别识别TRITC通道有信号的目标和CY5通道有信号的目标,并分别将图片切割成以该目标为中心的图像,即关于次级目标的单个细胞的图像;(4)计算各类特征参数,分别计算识别到的首要目标和次级目标的形态学参数,各通道荧光信号强度参数;(5)数据导出和保存,导出并保存各类形态学参数的数据,以及关于首要目标和次级目标的单个细胞图像;

[0114] 步骤三、输出临床样本单个细胞的图像和特征参数,分别输出关于首要目标、次级目标的单个细胞图像和特征参数;

[0115] 步骤四、如图8,采用预先建立的分类器自动化判读,筛选疑似 CTC的细胞作为候选细胞,对候选细胞进行打分,将候选细胞按分数从高到低排序,通过CellProfiler Analyst的交互界面呈现给人工;专业人员在交互界面中,对候选细胞进行审核,确认最终CTC的数量,并出具报告。

[0116] 进一步的,自动化判读是基于R语言的libraries建立的。

[0117] 本发明的第二个实施例如图2所示,本发明还公开了用于CTC进行自动化判的分类器的建立方法,包括如下步骤:

[0118] 步骤I、输入临床样本的原始荧光图像,临床样本为外周血经过 CTC富集、扫描后获得;

[0119] 步骤Ⅱ、对临床样本的原始荧光图像进行图像前处理,图像前处理是基于Cellprofiler建立的流程,图像前处理包括步骤:(1)图像修正,修正不均一光照强度导致的不均一的图像信号和背景;(2)识别首要目标,识别DAPI通道有信号的目标,并将图片切割成以该目标为中心的图像,即关于首要目标的单个细胞的图像;(3)识别次级目标;在识别DAPI通道有信号的目标基础上,分别识别TRITC通道有信号的目标和CY5通道有信号的目标,并分别将图片切割成以该目标为中心的图像,即关于次级目标的单个细胞的图像;(4)计算各类特征参数,分别计算识别到的首要目标和次级目标的形态学参数,各通道荧光信号强度参数;(5)数据导出和保存,导出并保存各类形态学参数的数据,以及关于首要目标和次级目标的单个细胞图像;

[0120] 步骤Ⅲ、输出临床样本单个细胞的图像和特征参数,具体包括分别输出关于首要目标、次级目标的单个细胞图像和特征参数,人工判断临床样本中是否含有CTC,如含有CTC即为临床CTC样本,继续步骤IV,如人工判断临床样本中没有CTC,则换一个样本重复步骤I,;

[0121] 步骤Ⅳ、人工标注CTC样本中分别属于CTC和WBC的细胞图像,借助R语言的libraries,对特征参数进行筛选,将单个细胞图像以及各类特征参数作为训练集,将不重要的特征参数剔除,筛选最优特征参数集合;基于多种机器学习的算法建立多个初步分类器;通过参数调优、交叉验证、平行比较优化分类器最终筛选出优化分类器。

[0122] 进一步的,步骤Ⅲ中Cellprofiler导出的形态学参数如表1所示,包括:DAPI通道的Area&Shape,Intensity,Texture;TRITC通道的Area&Shape,Intensity,Texture;CY5通道的Area&Shape,Intensity, Texture;TRITC通道和CY5通道的Correlation,共计778个特征参数。

[0123]

特征参数类型	荧光通道	形态学参数名称（举例）	特征参数数量	
			筛选前	筛选后
Area&Shape	DAPI	Area, Perimeter, Radius, Formfactor, Eccentricity, Compactness, etc.	49	21
	TRITC		49	29
	CY5		49	29
Intensity	DAPI	MinIntensity, MaxIntensity, MeanIntensity, StdIntensity, EdgeIntensity, etc.	15	6
	TRITC		15	15
	CY5		15	15
Texture	DAPI	RadialDistribution, Contrast, Entropy, Correlation, Variance, etc.	192	53
	TRITC		192	71
	CY5		192	71
Correlation	CY5 & TRITC	Correlation, Costes, RWC, K, Overlaps, etc.	10	8

[0124] 表1图像前处理导出的细胞全部特征参数，以及筛选后的特征参数列表

[0125] 优选的，步骤Ⅲ中收集了超过1700个CTC和13000个WBC的图像，同时每个细胞图像分别采集778个特征参数作为训练集。

[0126] 进一步的，步骤Ⅳ中，筛选最优特征参数集合的流程包括；

[0127] (1)数据中心化和归一化；

[0128] (2)基于每个特征参数的散点图，手动筛选能显著区分两种类别细胞的特征参数；

[0129] (3)剔除高度相关的特征参数(cutoff>.75)；

[0130] (4)计算特征参数重要性(RFE)，并最终确认用于模型建立的特征参数集合；

[0131] 经过上述步骤将不重要的特征参数剔除掉，获得一个新的训练集。优选的，778个特征参数筛选后数量降到318个。

[0132] 进一步的，步骤Ⅳ还包括：在新训练集内，分别使用多种监督式机器学习算法以及融合模型算法、多种不平衡训练集的前处理方法和多种评估方法用于测试和比较，进行交

叉验证，优化参数，建立初步分类器。多种监督式机器学习算法包括7种算法：K-Nearest Neighbors (KNN)、Stochastic Gradient Boosting(GBM)、AdaBoost Classification Trees(ADABOOST)、Support Vector Machines(SVM)、Random Forest (RF)、Naïve Bayes (NB)、Extreme Gradient Boosting(XGB)，以及1种融合模型算法，即7种监督式机器学习算法融合在一起的算法(Stack)。不平衡训练集的前处理方法包括5种处理方式：Original、Up-sampling、 Down-sampling、SMOTE、ROSE。评估方法包括2种：ROC和PR。最终生成的分类器个数为80个，计算方法：(7+1)*5*2＝80个分类器，80 个分类器见表2。

[0133]

序号	监督式机器学习算法	评估方法	不平衡训练集的前处理方法	训练后分类器
1	KNN	PR	Original	KNN_PR_Original
2	KNN	PR	Up-sampling	KNN_PR_Up-sampling
3	KNN	PR	Down-sampling	KNN_PR_Down-sampling
4	KNN	PR	SMOTE	KNN_PR_SMOTE
5	KNN	PR	ROSE	KNN_PR_ROSE
6	KNN	ROC	Original	KNN_ROC_Original
7	KNN	ROC	Up-sampling	KNN_ROC_Up-sampling
8	KNN	ROC	Down-sampling	KNN_ROC_Down-sampling
9	KNN	ROC	SMOTE	KNN_ROC_SMOTE
10	KNN	ROC	ROSE	KNN_ROC_ROSE
11	GBM	PR	Original	GBM_PR_Original
12	GBM	PR	Up-sampling	GBM_PR_Up-sampling
13	GBM	PR	Down-sampling	GBM_PR_Down-sampling
14	GBM	PR	SMOTE	GBM_PR_SMOTE
15	GBM	PR	ROSE	GBM_PR_ROSE
16	GBM	ROC	Original	GBM_ROC_Original
17	GBM	ROC	Up-sampling	GBM_ROC_Up-sampling
18	GBM	ROC	Down-sampling	GBM_ROC_Down-sampling
19	GBM	ROC	SMOTE	GBM_ROC_SMOTE
20	GBM	ROC	ROSE	GBM_ROC_ROSE
21	ADABOOST	PR	Original	ADABOOST_PR_Original
22	ADABOOST	PR	Up-sampling	ADABOOST_PR_Up-sampling
23	ADABOOST	PR	Down-sampling	ADABOOST_PR_Down-sampling
24	ADABOOST	PR	SMOTE	ADABOOST_PR_SMOTE
25	ADABOOST	PR	ROSE	ADABOOST_PR_ROSE
26	ADABOOST	ROC	Original	ADABOOST_ROC_Original
27	ADABOOST	ROC	Up-sampling	ADABOOST_ROC_Up-sampling
28	ADABOOST	ROC	Down-sampling	ADABOOST_ROC_Down-sampling

[0134]

29	ADABOOST	ROC	SMOTE	ADABOOST_ROC_SMOTE
30	ADABOOST	ROC	ROSE	ADABOOST_ROC_ROSE
31	SVM	PR	Original	SVM_PR_Original
32	SVM	PR	Up-sampling	SVM_PR_Up-sampling
33	SVM	PR	Down-sampling	SVM_PR_Down-sampling
34	SVM	PR	SMOTE	SVM_PR_SMOTE
35	SVM	PR	ROSE	SVM_PR_ROSE
36	SVM	ROC	Original	SVM_ROC_Original
37	SVM	ROC	Up-sampling	SVM_ROC_Up-sampling
38	SVM	ROC	Down-sampling	SVM_ROC_Down-sampling
39	SVM	ROC	SMOTE	SVM_ROC_SMOTE
40	SVM	ROC	ROSE	SVM_ROC_ROSE
41	RF	PR	Original	RF_PR_Original
42	RF	PR	Up-sampling	RF_PR_Up-sampling
43	RF	PR	Down-sampling	RF_PR_Down-sampling
44	RF	PR	SMOTE	RF_PR_SMOTE
45	RF	PR	ROSE	RF_PR_ROSE
46	RF	ROC	Original	RF_ROC_Original
47	RF	ROC	Up-sampling	RF_ROC_Up-sampling
48	RF	ROC	Down-sampling	RF_ROC_Down-sampling
49	RF	ROC	SMOTE	RF_ROC_SMOTE
50	RF	ROC	ROSE	RF_ROC_ROSE
51	NB	PR	Original	NB_PR_Original
52	NB	PR	Up-sampling	NB_PR_Up-sampling
53	NB	PR	Down-sampling	NB_PR_Down-sampling
54	NB	PR	SMOTE	NB_PR_SMOTE
55	NB	PR	ROSE	NB_PR_ROSE
56	NB	ROC	Original	NB_ROC_Original
57	NB	ROC	Up-sampling	NB_ROC_Up-sampling
58	NB	ROC	Down-sampling	NB_ROC_Down-sampling
59	NB	ROC	SMOTE	NB_ROC_SMOTE
60	NB	ROC	ROSE	NB_ROC_ROSE
61	XGB	PR	Original	XGB_PR_Original
62	XGB	PR	Up-sampling	XGB_PR_Up-sampling
63	XGB	PR	Down-sampling	XGB_PR_Down-sampling
64	XGB	PR	SMOTE	XGB_PR_SMOTE
65	XGB	PR	ROSE	XGB_PR_ROSE
66	XGB	ROC	Original	XGB_ROC_Original
67	XGB	ROC	Up-sampling	XGB_ROC_Up-sampling

[0135]

68	XGB	ROC	Down-sampling	XGB_ROC_Down-sampling
69	XGB	ROC	SMOTE	XGB_ROC_SMOTE
70	XGB	ROC	ROSE	XGB_ROC_ROSE
71	Stack	PR	Original	Stack_PR_Original
72	Stack	PR	Up-sampling	Stack_PR_Up-sampling
73	Stack	PR	Down-sampling	Stack_PR_Down-sampling
74	Stack	PR	SMOTE	Stack_PR_SMOTE
75	Stack	PR	ROSE	Stack_PR_ROSE
76	Stack	ROC	Original	Stack_ROC_Original
77	Stack	ROC	Up-sampling	Stack_ROC_Up-sampling
78	Stack	ROC	Down-sampling	Stack_ROC_Down-sampling
79	Stack	ROC	SMOTE	Stack_ROC_SMOTE
80	Stack	ROC	ROSE	Stack_ROC_ROSE

[0136] 表2分类器

[0137] 进一步的，步骤Ⅳ还包括：用训练集对分类器的性能进行评估，分析比较了AUC、F1 score、Accuracy、Precision、Recall、TPR、FPR 参数在80个分类器中的表现，如表3所示，包括：表3A机器学习算法为KNN的分类器建立列表及其调优后性能、表3A机器学习算法为 KNN的分类器建立列表及其调优后性能、表3B机器学习算法为GBM 的分类器建立列表及其调优后性能、表3C机器学习算法为 ADABOOST的分类器建立列表及其调优后性能、表3D机器学习算法为SVM的分类器建立列表及其调优后性能、表3E机器学习算法为 RF的分类器建立列表及其调优后性能、表3F机器学习算法为NB的分类器建立列表及其调优后性能、表3G机器学习算法为XGB的分类器建立列表及其调优后性能。

[0138]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
KNN	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	KNN_PR_Down-sampling	0.995	0.937	98%	93%	94%	94%	1%
		KNN_PR_Original	0.999	0.944	98%	92%	94%	94%	1%
		KNN_PR_ROSE	0.251	0.661	94%	98%	48%	48%	0%
		KNN_PR_SMOTE	0.997	0.949	98%	94%	93%	93%	1%
		KNN_PR_Up-sampling	0.999	0.971	98%	95%	92%	92%	1%
		KNN_ROC_Down-sampling	0.995	0.937	98%	93%	94%	94%	1%
		KNN_ROC_Original	0.999	0.944	98%	92%	94%	94%	1%
		KNN_ROC_ROSE	0.251	0.661	94%	98%	48%	48%	0%
		KNN_ROC_SMOTE	0.997	0.949	98%	94%	93%	93%	1%
		KNN_ROC_Up-sampling	0.999	0.971	98%	95%	92%	92%	1%

[0139] 表3A机器学习算法KNN训练后的分类器的表现

[0140]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
GBM	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	GBM_PR_Down-sampling	0.999	0.966	99%	94%	95%	95%	1%
		GBM_PR_Original	1.000	0.992	99%	95%	96%	96%	1%
		GBM_PR_ROSE	0.994	0.736	95%	99%	62%	62%	0%
		GBM_PR_SMOTE	0.999	0.975	99%	94%	97%	97%	1%
		GBM_PR_Up-sampling	1.000	0.991	99%	95%	97%	97%	1%
		GBM_ROC_Down-sampling	0.999	0.966	99%	94%	95%	95%	1%
		GBM_ROC_Original	1.000	0.989	99%	96%	96%	96%	0%
		GBM_ROC_ROSE	0.994	0.913	98%	95%	90%	90%	1%
		GBM_ROC_SMOTE	0.999	0.975	99%	94%	97%	97%	1%
		GBM_ROC_Up-sampling	1.000	0.991	99%	95%	97%	97%	1%

[0141] 表3B机器学习算法GBM训练后的分类器的表现

[0142]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
ADABOOST	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	ADABOOST_PR_Down-sampling	0.999	0.982	99%	96%	93%	93%	1%
		ADABOOST_PR_Original	1.000	0.997	98%	88%	99%	99%	2%
		ADABOOST_PR_ROSE	0.995	0.921	98%	92%	93%	93%	1%
		ADABOOST_PR_SMOTE	0.925	0.784	93%	65%	96%	96%	7%
		ADABOOST_PR_Up-sampling	0.726	0.507	77%	34%	90%	90%	1%
		ADABOOST_ROC_Down-sampling	0.999	0.981	99%	96%	93%	93%	1%
		ADABOOST_ROC_Original	1.000	0.997	98%	88%	99%	99%	2%
		ADABOOST_ROC_ROSE	0.995	0.921	98%	92%	93%	93%	1%
		ADABOOST_ROC_SMOTE	1.000	0.990	99%	98%	91%	91%	0%
		ADABOOST_ROC_Up-sampling	1.000	0.997	98%	89%	99%	99%	2%

[0143] 表3C机器学习算法ADABOOST训练后的分类器的表现

[0144]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
SVM	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	SVM_PR_Down-sampling	0.996	0.938	98%	93%	91%	91%	1%
		SVM_PR_Original	0.997	0.956	99%	95%	94%	94%	1%
		SVM_PR_ROSE	0.998	0.945	98%	91%	95%	95%	1%
		SVM_PR_SMOTE	0.997	0.957	98%	92%	95%	95%	1%
		SVM_PR_Up-sampling	0.998	0.972	98%	95%	92%	92%	1%
		SVM_ROC_Down-sampling	0.996	0.945	98%	92%	94%	94%	1%
		SVM_ROC_Original	0.997	0.961	99%	94%	97%	97%	1%
		SVM_ROC_ROSE	0.998	0.945	98%	91%	95%	95%	1%
		SVM_ROC_SMOTE	0.997	0.957	98%	92%	95%	95%	1%
		SVM_ROC_Up-sampling	0.998	0.972	98%	95%	92%	92%	1%

[0145] 表3D机器学习算法SVM训练后的分类器的表现

[0146]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
RF	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	RF_PR_Down-sampling	0.999	0.962	98%	94%	92%	92%	1%
		RF_PR_Original	1.000	0.997	98%	92%	95%	95%	1%
		RF_PR_ROSE	0.996	0.918	98%	92%	94%	94%	1%
		RF_PR_SMOTE	1.000	0.981	98%	95%	92%	92%	1%
		RF_PR_Up-sampling	1.000	0.997	98%	91%	95%	95%	1%
		RF_ROC_Down-sampling	0.999	0.967	98%	95%	92%	92%	1%
		RF_ROC_Original	1.000	0.997	99%	94%	97%	97%	1%
		RF_ROC_ROSE	0.996	0.918	98%	92%	94%	94%	1%
		RF_ROC_SMOTE	0.999	0.984	99%	96%	94%	94%	0%
		RF_ROC_Up-sampling	1.000	0.997	99%	95%	96%	96%	1%

[0147] 表3E机器学习算法RF训练后的分类器的表现

[0148]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
NB	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	NB_PR_Down-sampling	0.989	0.915	98%	93%	91%	91%	1%
		NB_PR_Original	0.992	0.913	98%	92%	92%	92%	1%
		NB_PR_ROSE	0.991	0.905	98%	94%	88%	88%	1%
		NB_PR_SMOTE	0.988	0.921	98%	93%	92%	92%	1%
		NB_PR_Up-sampling	0.988	0.918	98%	93%	92%	92%	1%
		NB_ROC_Down-sampling	0.991	0.915	98%	90%	93%	93%	1%
		NB_ROC_Original	0.992	0.913	98%	92%	92%	92%	1%
		NB_ROC_ROSE	0.991	0.905	98%	94%	88%	88%	1%
		NB_ROC_SMOTE	0.991	0.918	98%	91%	93%	93%	1%
		NB_ROC_Up-sampling	0.992	0.920	98%	91%	93%	93%	1%

[0149] 表3F机器学习算法NB训练后的分类器的表现

[0150]

机器学习算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
XGB	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	XGB_PR_Down-sampling	0.999	0.957	99%	93%	96%	96%	1%
		XGB_PR_Original	1.000	0.997	99%	95%	97%	97%	1%
		XGB_PR_ROSE	0.995	0.914	98%	94%	91%	91%	1%
		XGB_PR_SMOTE	0.999	0.964	99%	95%	95%	95%	1%
		XGB_PR_Up-sampling	1.000	0.997	99%	96%	97%	97%	1%
		XGB_ROC_Down-sampling	0.999	0.970	99%	94%	96%	96%	1%
		XGB_ROC_Original	1.000	0.997	99%	94%	98%	98%	1%
		XGB_ROC_ROSE	0.996	0.933	98%	91%	96%	96%	1%
		XGB_ROC_SMOTE	1.000	0.986	99%	96%	95%	95%	0%
		XGB_ROC_Up-sampling	1.000	0.997	99%	93%	99%	99%	1%

[0151] 表3G机器学习算法XGB训练后的分类器的表现

[0152]

融合模型算法	训练参数	训练后分类器	AUC	F1 score	Accuracy	Precision	Recall	TPR	FPR
Stack	1. 基于ROC或PR的度量模型；2. 对不平衡训练集的处理方法：Original, Up-sampling, Down-sampling, ROSE, SMOTE	Stack_PR_Down-sampling	0.999	0.784	99%	96%	95%	94%	1%
		Stack_PR_Original	0.925	0.997	98%	94%	88%	88%	1%
		Stack_PR_ROSE	1.000	0.997	99%	96%	99%	97%	1%
		Stack_PR_SMOTE	0.999	0.966	98%	95%	94%	90%	1%
		Stack_PR_Up-sampling	1.000	0.913	98%	92%	95%	95%	7%
		Stack_ROC_Down-sampling	0.999	0.905	99%	96%	90%	95%	1%
		Stack_ROC_Original	0.991	0.997	93%	94%	94%	96%	0%
		Stack_ROC_ROSE	1.000	0.957	99%	65%	96%	99%	0%
		Stack_ROC_SMOTE	0.994	0.984	99%	91%	95%	94%	1%
		Stack_ROC_Up-sampling	0.997	0.984	98%	93%	97%	95%	1%

[0153] 表3H融合模型算法Stack训练后的分类器的表现

[0154] 针对不平衡数据集而言(不平衡数据集是指CTC数量在训练集中占比小于10%)，在各项指标均表现良好情况下，F1 score、Recall、TPR指标更关注于识别CTC的灵敏度，所以，80个分类器中，F1 score 值最高为97%，F1 score值为97%的分类器包括：

[0155] (1)ADABOOST_PR_Original、

[0156] (2)ADABOOST_ROC_Original、

[0157] (3)ADABOOST_ROC_Up-sampling、

[0158] (4)RF_PR_Original、

[0159] (5)RF_PR_Up-sampling、

[0160] (6)RF_ROC_Original、

[0161] (7)RF_ROC_Up-sampling、

[0162] (8)XGB_PR_Original、

[0163] (9)XGB_PR_Up-sampling、

[0164] (10)XGB_ROC_Original、

[0165] (11)XGB_ROC_Up-sampling、

[0166] (12)Stack_PR_ROSE、

[0167] (13)Stack_ROC_Original、

[0168] (14)Stack_PR_Original;

[0169] Rcall值最高为99%,Rcall值为99%的分类器包括:

[0170] (1)ADABOOST_PR_Original、

[0171] (2)ADABOOST_ROC_Original、

[0172] (3)ADABOOST_ROC_Up-sampling、

[0173] (4)XGB_ROC_Up-sampling、

[0174] (5)Stack_PR_ROSE;

[0175] TPR值最高为99%,TPR值为99%的分类器包括:

[0176] (1)ADABOOST_PR_Original、

[0177] (2)ADABOOST_ROC_Original、

[0178] (3)ADABOOST_ROC_Up-sampling、

[0179] (4)XGB_ROC_Up-sampling、

[0180] (5)Stack_ROC_ROSE;

[0181] 综上,F1 score、Recall、TPR值均表现良好的分类器为:

[0182] (1)ADABOOST_PR_Original、

[0183] (2)ADABOOST_ROC_Original、

[0184] (3)ADABOOST_ROC_Up-sampling、

[0185] (4)XGB_ROC_Up-sampling。

[0186] 进一步的,步骤Ⅳ还包括:用建立调优好的分类器(表2中的80 个分类器)在临床样本中进行泛化能力测试,测试范围涵盖了200例临床CTC检测样本(200例临床CTC检测样本中人工判读结果约1000 个CTC)。基于CTC自动化判读需求,泛化能力评估指标包括:Positive sample concordance,CTC concordance,Screening efficiency。针对于临床CTC检测应用,期望分类器的Positive sample concordance需要达到100%,在Positive sample concordance需要达到100%的前提下,CTC concordance和Screening efficiency尽可能的高,如CTC concordance尽大于90%,Screening efficiency大于95%。将测试结果用图7呈现:

[0187] Positive sample concordance达到100%的包括:

[0188] (1)XGB_ROC_SMOTE、

[0189] (2)XGB_ROC_Down-sampling、

[0190] (3)XGB_PR_SMOTE、

[0191] (4)XGB_PR_Down-sampling、

[0192] (5)SVM_ROC_Down-sampling、

[0193] (6)SVM_PR_Down-sampling、

[0194] (7)Stack_ROC_Original、

[0195] (8)Stack_PR_SMOTE、

[0196] (9)Stack_PR_ROSE、

[0197] (10)Stack_PR_Down-sampling、

[0198] (11)RF_ROC_Down-sampling、

[0199] (12)RF_PR_Down-sampling、

[0200] (13)KNN_ROC_Up-sampling、

[0201] (14)KNN_PR_Up-sampling、

[0202] (15)GBM_ROC_Up-sampling、

[0203] (16)GBM_ROC_SMOTE、

[0204] (17)GBM_ROC_Down-sampling、

[0205] (18)GBM_PR_Up-sampling、

[0206] (19)GBM_PR_SMOTE、

[0207] (20)GBM_PR_Down-sampling、

[0208] (21)ADABOOST_ROC_SMOTE、

[0209] (22)ADABOOST_ROC_ROSE、

[0210] (23)ADABOOST_ROC_Down-sampling、

[0211] (24)ADABOOST_PR_ROSE、

[0212] (25)ADABOOST_PR_Down-sampling;

[0213] 进一步的,在Positive sample concordance达到100%的25个分类器中,选择CTC concordance达到90%的分类器包括8个:

[0214] (1)KNN_ROC_Up-sampling

[0215] (2)KNN_PR_Up-sampling

[0216] (3)XGB_ROC_Down-sampling

[0217] (4)XGB_PR_Down-sampling

[0218] (5)SVM_ROC_Down-sampling

[0219] (6)SVM_PR_Down-sampling

[0220] (7)Stack_PR_Down-sampling

[0221] 进一步的,在Positive sample concordance达到100%的25个分类器中,选择Screening efficiency大于95%的分类器包括1个:

[0222] (1)Stack_ROC_Original、

[0223] 综上,前述9个分类器为优选的分类器,在选择最终使用分类器时,可以:(1)选择优选分类器中任何一个;(2)根据实际应用需求选择;(3)选择优选分类器中CTC concordance最高的一个。

[0224] 本发明的第三个实施例，如图3所示，一种基于人工智能的CTC 图像识别系统，包括输入模块1、图像前处理模块2、输出模块3、自动化判读模块4；

[0225] 输入模块1用于输入临床样本的原始荧光图像；

[0226] 图像前处理模2块用于临床样本的原始荧光图像进行处理后获得单个细胞的图像和特征参数；

[0227] 输出模块3用于输出临床样本中单个细胞的图像和特征参数，

[0228] 自动化判读模块4用于识别CTC细胞，自动化判读模块4包括初审模块401和终审模块402，初审模块401使用预先建立的分类器进行筛选，筛选出疑似CTC的细胞作为候选细胞；终审模块402采用专业人员判断候选细胞是否确实为CTC，确认后出具CTC检测报告。

[0229] 进一步的，所述图像前处理模块2包括图像修正模块201、识别首要目标模块202、识别次级目标模块203、计算各类特征参数模块204、数据导出和保存模块205；所述图像修正模201块用于修正不均一光照强度导致的不均一的图像信号和背景；所述识别首要目标模202块用于识别DAPI通道有信号的目标；所述识别次级目标模块203用于在识别DAPI通道有信号的目标基础上，分别识别TRITC通道有信号的目标和CY5通道有信号的目标，并分别获得单个细胞图像；所述计算各类特征参数模块204用于计算单个细胞的形态学参数、各通道荧光信号强度参数；所述数据导出和保存模块205用于导出和保存单个细胞的图像和特征参数。

[0230] 本发明的第四个实施例如图4所示，本发明还公开了分类器建立系统，包括输入模块二5、图像前处理模块二6、输出模块二7、分类器建立模块8；

[0231] 输入模块二5用于输入临床CTC样本的原始荧光图像；

[0232] 图像前处理模二6块用于对临床CTC样本的原始荧光图像进行处理后获得单个细胞的图像和特征参数，作为训练集；

[0233] 输出模块二7用于输出临床CTC样本中单个细胞的图像和特征参数；

[0234] 分类器建立模块8用于建立并优化分类器，人工标注CTC和 WBC，筛选CTC和WBC的特征参数；基于多种机器学习的算法建立多个初步分类器；通过参数调优、交叉验证、平行比较优化分类器。

[0235] 进一步的，CTC图像识别系统和分类器建立系统中用到的图像前处理模块二6包括图像修正模块601、识别首要目标模块602、识别次级目标模块603、计算各类特征参数模块604、数据导出和保存模块 605；图像修正模块601用于修正不均一光照强度导致的不均一的图像信号和背景；识别首要目标模块602用于识别DAPI通道有信号的目标；识别次级目标模块603用于在识别DAPI通道有信号的目标基础上，分别识别TRITC通道有信号的目标和CY5通道有信号的目标，并分别获得单个细胞图像；计算各类特征参数模块604用于计算单个细胞的形态学参数、各通道荧光信号强度参数；数据导出和保存模块605 用于导出和保存单个细胞的图像和特征参数。

[0236] 以上仅是本发明的优选实施方式，本发明的保护范围并不仅局限于实施例，凡属于本发明思路下的技术方案均属于本发明的保护范围。应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理前提下的若干改进和变换，应视为本发明的保护范围。

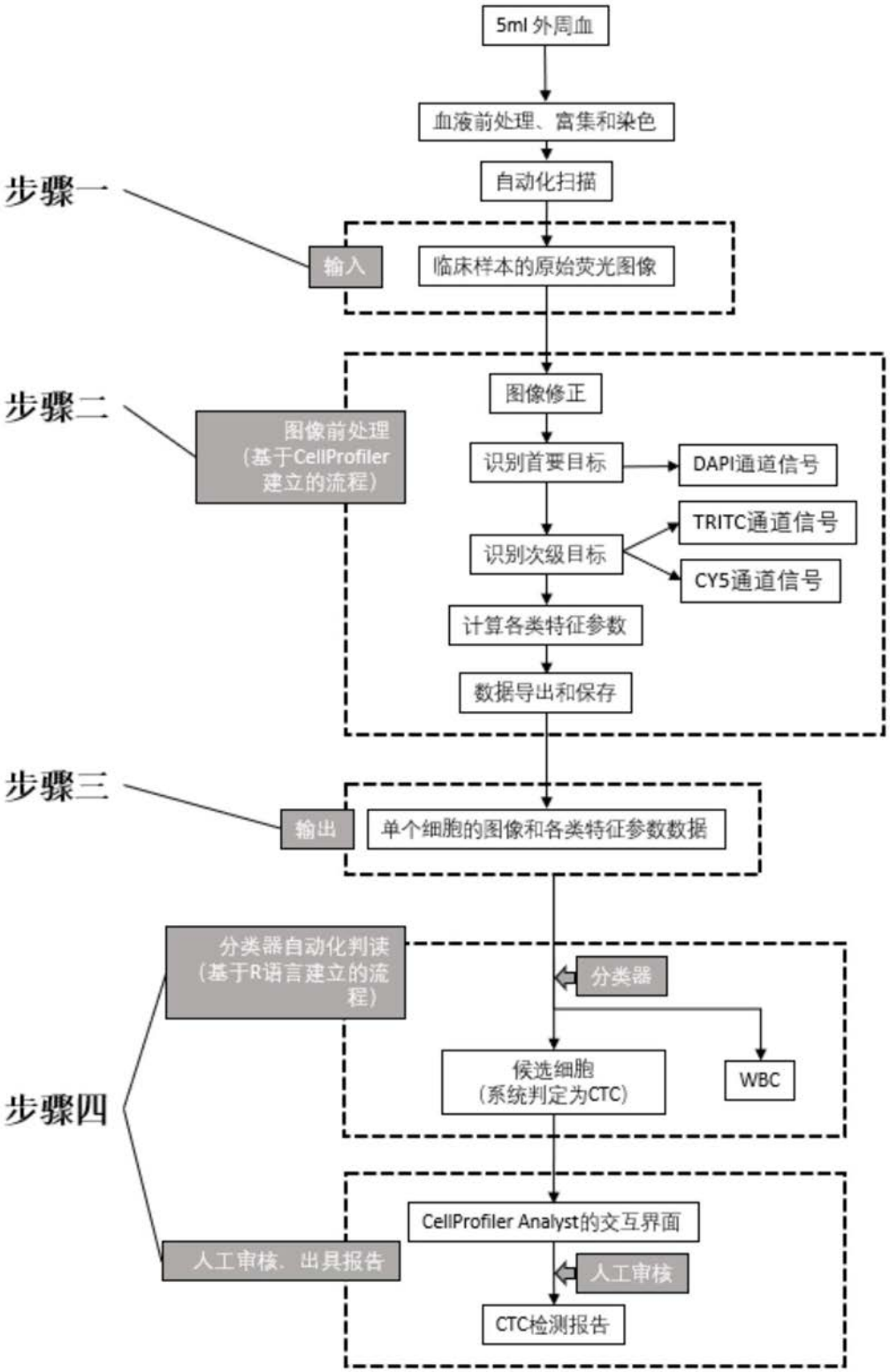

5ml 外周血
血液前处理、富集和染色
自动化扫描
步骤一
输入
临床样本的原始荧光图像
步骤二
图像前处理
(基于CellProfiler
建立的流程)
图像修正
识别首要目标
DAPI通道信号
识别次级目标
TRITC通道信号
CY5通道信号
计算各类特征参数
数据导出和保存
步骤三
输出
单个细胞的图像和各类特征参数数据
步骤四
分类器自动化判读
(基于R语言建立的流
程)
分类器
候选细胞
(系统判定为CTC)
WBC
CellProfiler Analyst的交互界面
人工审核，出具报告
人工审核
CTC检测报告

图1

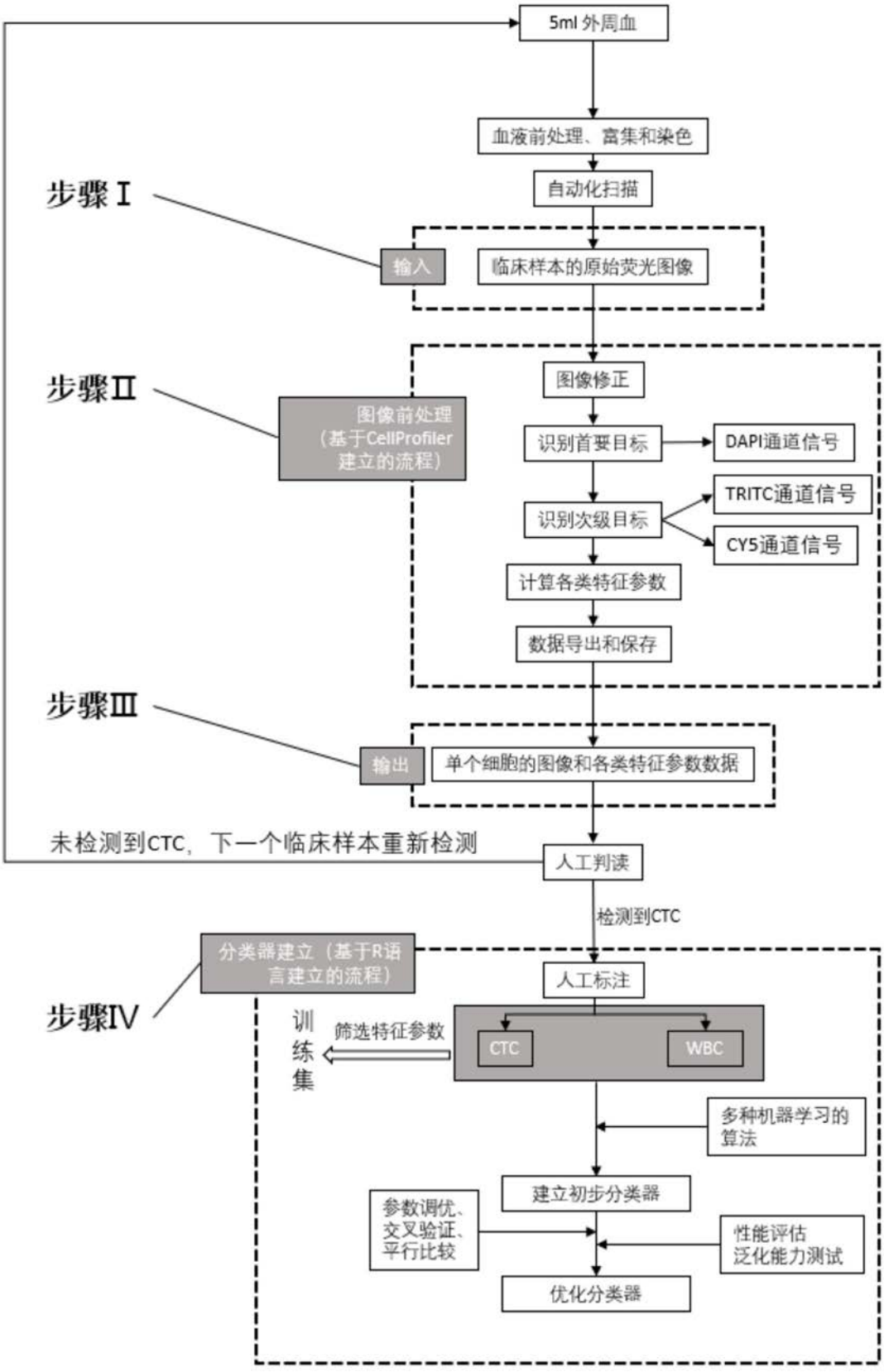

5ml 外周血
血液前处理、富集和染色
自动化扫描
步骤 I
输入
临床样本的原始荧光图像
步骤II
图像前处理（基于CellProfiler建立的流程）
图像修正
识别首要目标
DAPI通道信号
识别次级目标
TRITC通道信号
CY5通道信号
计算各类特征参数
数据导出和保存
步骤III
输出
单个细胞的图像和各类特征参数数据
未检测到CTC，下一个临床样本重新检测
人工判读
检测到CTC
步骤IV
分类器建立（基于R语言建立的流程）
人工标注
训练集
筛选特征参数
CTC
WBC
多种机器学习的算法
建立初步分类器
参数调优、交叉验证、平行比较
性能评估泛化能力测试
优化分类器

图2

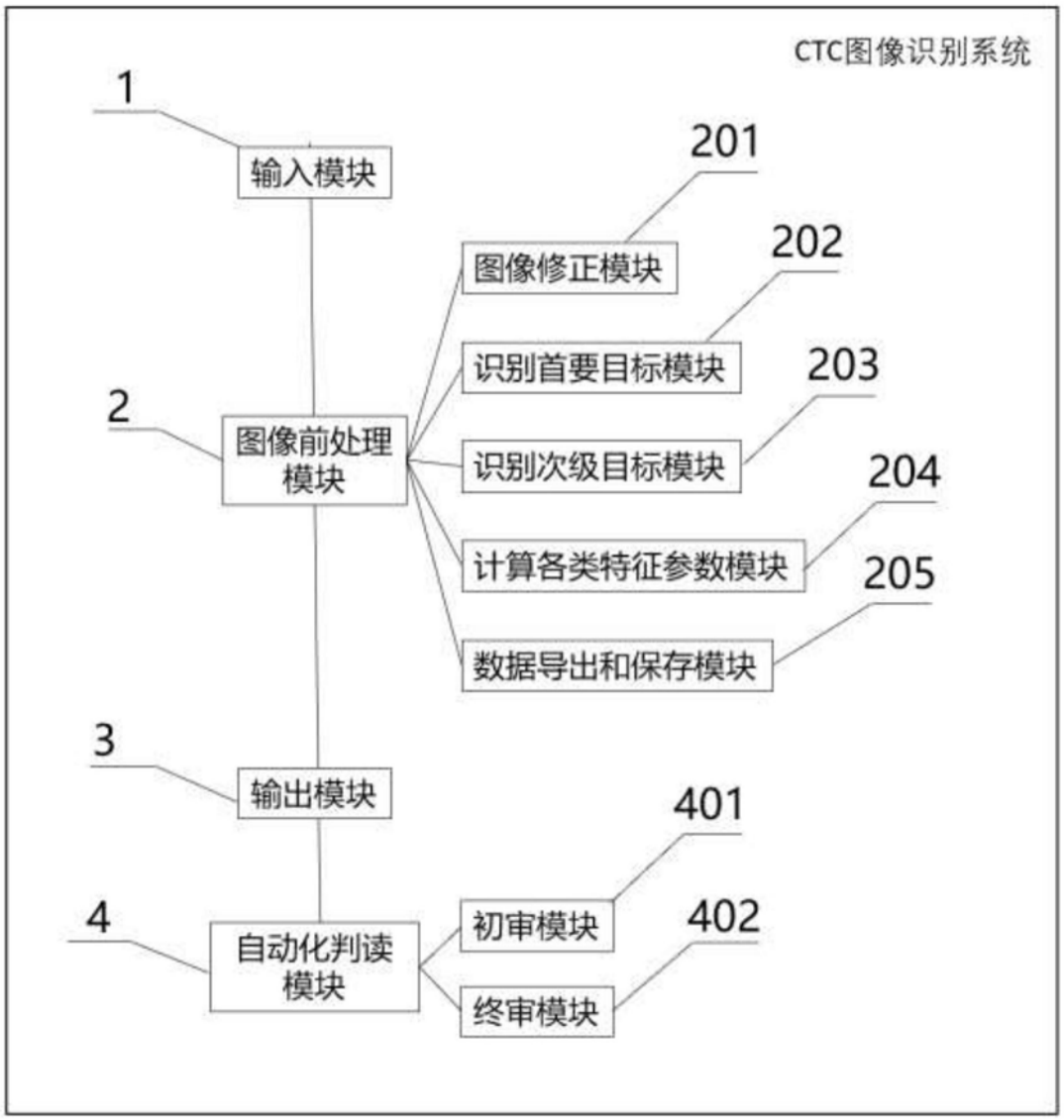

CTC图像识别系统
1
输入模块
2
图像前处理模块
201
图像修正模块
202
识别首要目标模块
203
识别次级目标模块
204
计算各类特征参数模块
205
数据导出和保存模块
3
输出模块
4
自动化判读模块
401
初审模块
402
终审模块

图3

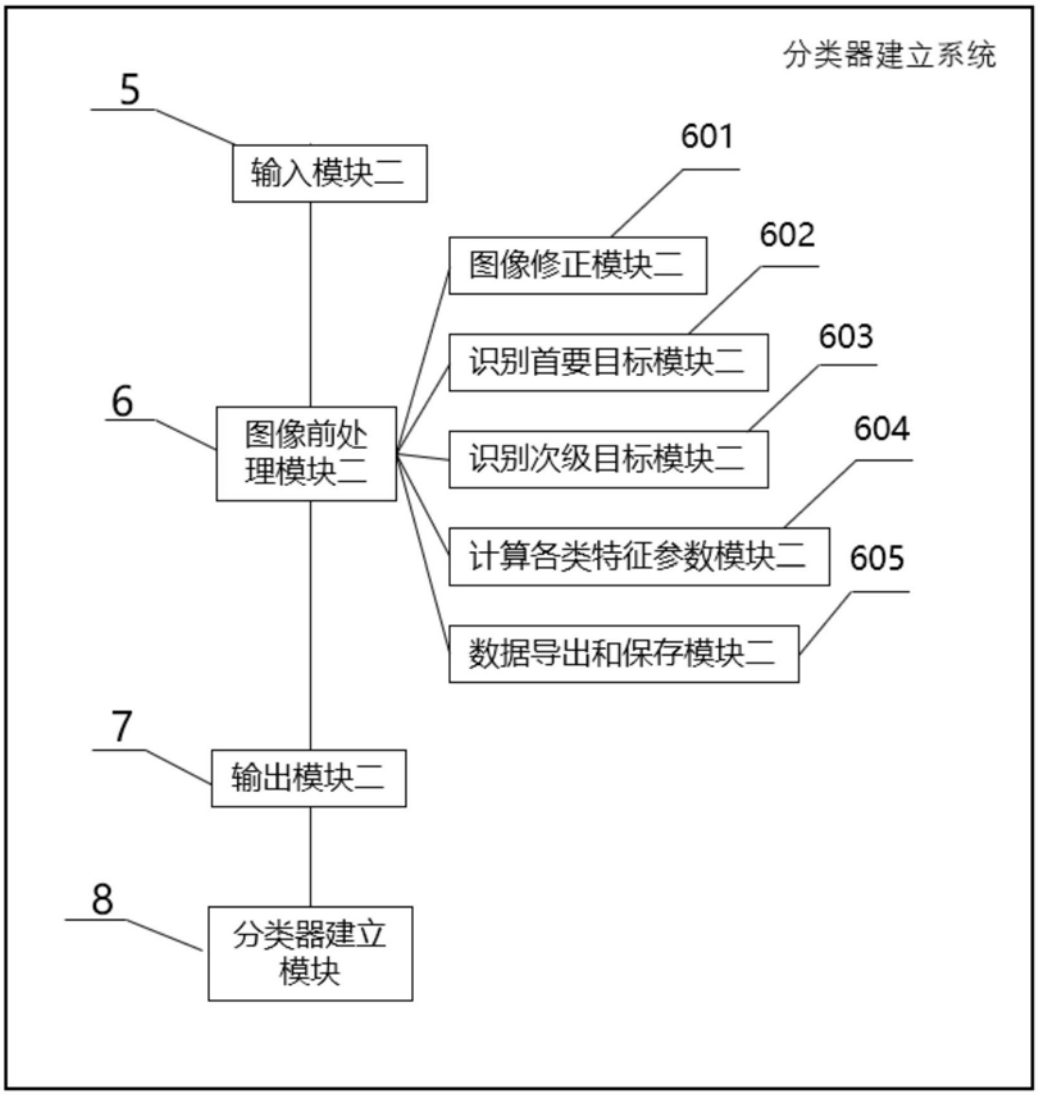

分类器建立系统
5
输入模块二
6
图像前处理模块二
601
图像修正模块二
602
识别首要目标模块二
603
识别次级目标模块二
604
计算各类特征参数模块二
605
数据导出和保存模块二
7
输出模块二
8
分类器建立模块

图4

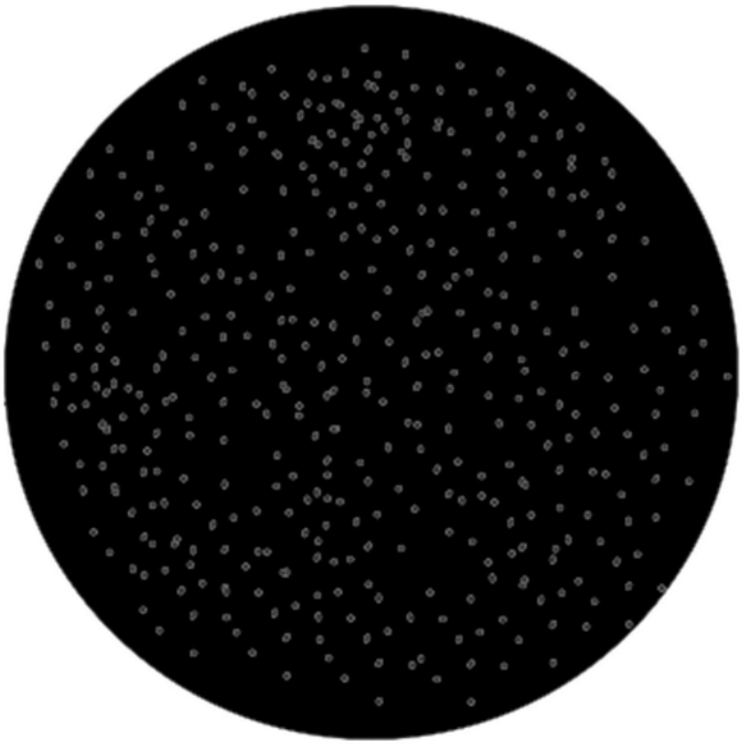


图5

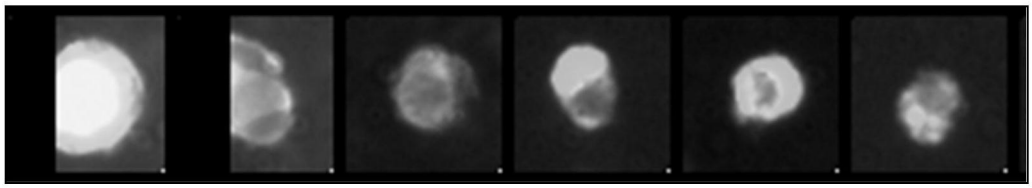


图6

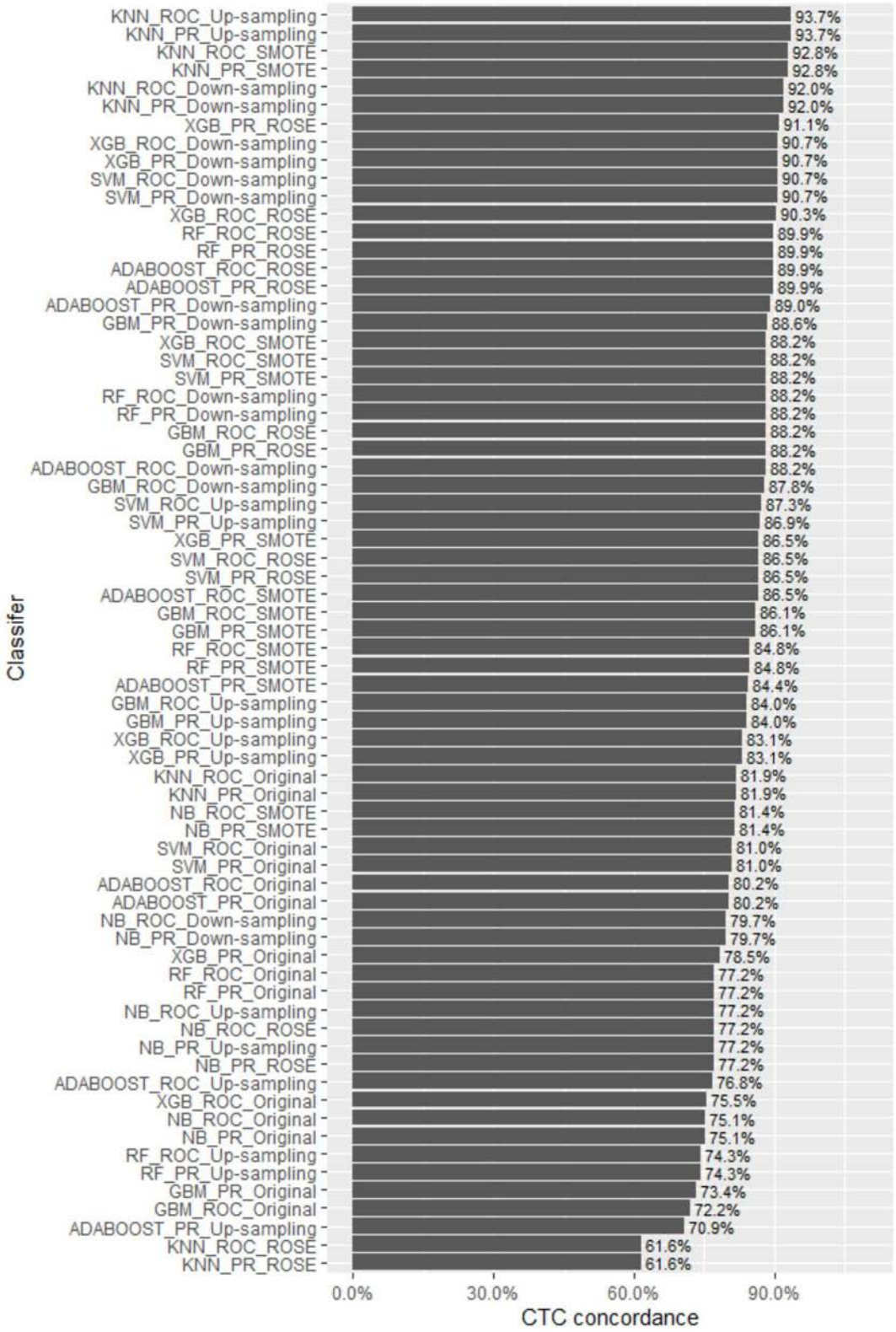

KNN_ROC_Up-sampling 93.7%
KNN_PR_Up-sampling 93.7%
KNN_ROC_SMOTE 92.8%
KNN_PR_SMOTE 92.8%
KNN_ROC_Down-sampling 92.0%
KNN_PR_Down-sampling 92.0%
XGB_PR_ROSE 91.1%
XGB_ROC_Down-sampling 90.7%
XGB_PR_Down-sampling 90.7%
SVM_ROC_Down-sampling 90.7%
SVM_PR_Down-sampling 90.7%
XGB_ROC_ROSE 90.3%
RF_ROC_ROSE 89.9%
RF_PR_ROSE 89.9%
ADABOOST_ROC_ROSE 89.9%
ADABOOST_PR_ROSE 89.9%
ADABOOST_PR_Down-sampling 89.0%
GBM_PR_Down-sampling 88.6%
XGB_ROC_SMOTE 88.2%
SVM_ROC_SMOTE 88.2%
SVM_PR_SMOTE 88.2%
RF_ROC_Down-sampling 88.2%
RF_PR_Down-sampling 88.2%
GBM_ROC_ROSE 88.2%
GBM_PR_ROSE 88.2%
ADABOOST_ROC_Down-sampling 88.2%
GBM_ROC_Down-sampling 87.8%
SVM_ROC_Up-sampling 87.3%
SVM_PR_Up-sampling 86.9%
XGB_PR_SMOTE 86.5%
SVM_ROC_ROSE 86.5%
SVM_PR_ROSE 86.5%
ADABOOST_ROC_SMOTE 86.5%
GBM_ROC_SMOTE 86.1%
GBM_PR_SMOTE 86.1%
RF_ROC_SMOTE 84.8%
RF_PR_SMOTE 84.8%
ADABOOST_PR_SMOTE 84.4%
GBM_ROC_Up-sampling 84.0%
GBM_PR_Up-sampling 84.0%
XGB_ROC_Up-sampling 83.1%
XGB_PR_Up-sampling 83.1%
KNN_ROC_Original 81.9%
KNN_PR_Original 81.9%
NB_ROC_SMOTE 81.4%
NB_PR_SMOTE 81.4%
SVM_ROC_Original 81.0%
SVM_PR_Original 81.0%
ADABOOST_ROC_Original 80.2%
ADABOOST_PR_Original 80.2%
NB_ROC_Down-sampling 79.7%
NB_PR_Down-sampling 79.7%
XGB_PR_Original 78.5%
RF_ROC_Original 77.2%
RF_PR_Original 77.2%
NB_ROC_Up-sampling 77.2%
NB_ROC_ROSE 77.2%
NB_PR_Up-sampling 77.2%
NB_PR_ROSE 77.2%
ADABOOST_ROC_Up-sampling 76.8%
XGB_ROC_Original 75.5%
NB_ROC_Original 75.1%
NB_PR_Original 75.1%
RF_ROC_Up-sampling 74.3%
RF_PR_Up-sampling 74.3%
GBM_PR_Original 73.4%
GBM_ROC_Original 72.2%
ADABOOST_PR_Up-sampling 70.9%
KNN_ROC_ROSE 61.6%
KNN_PR_ROSE 61.6%
Classifer
0.0%
30.0%
60.0%
90.0%
CTC concordance

图7A

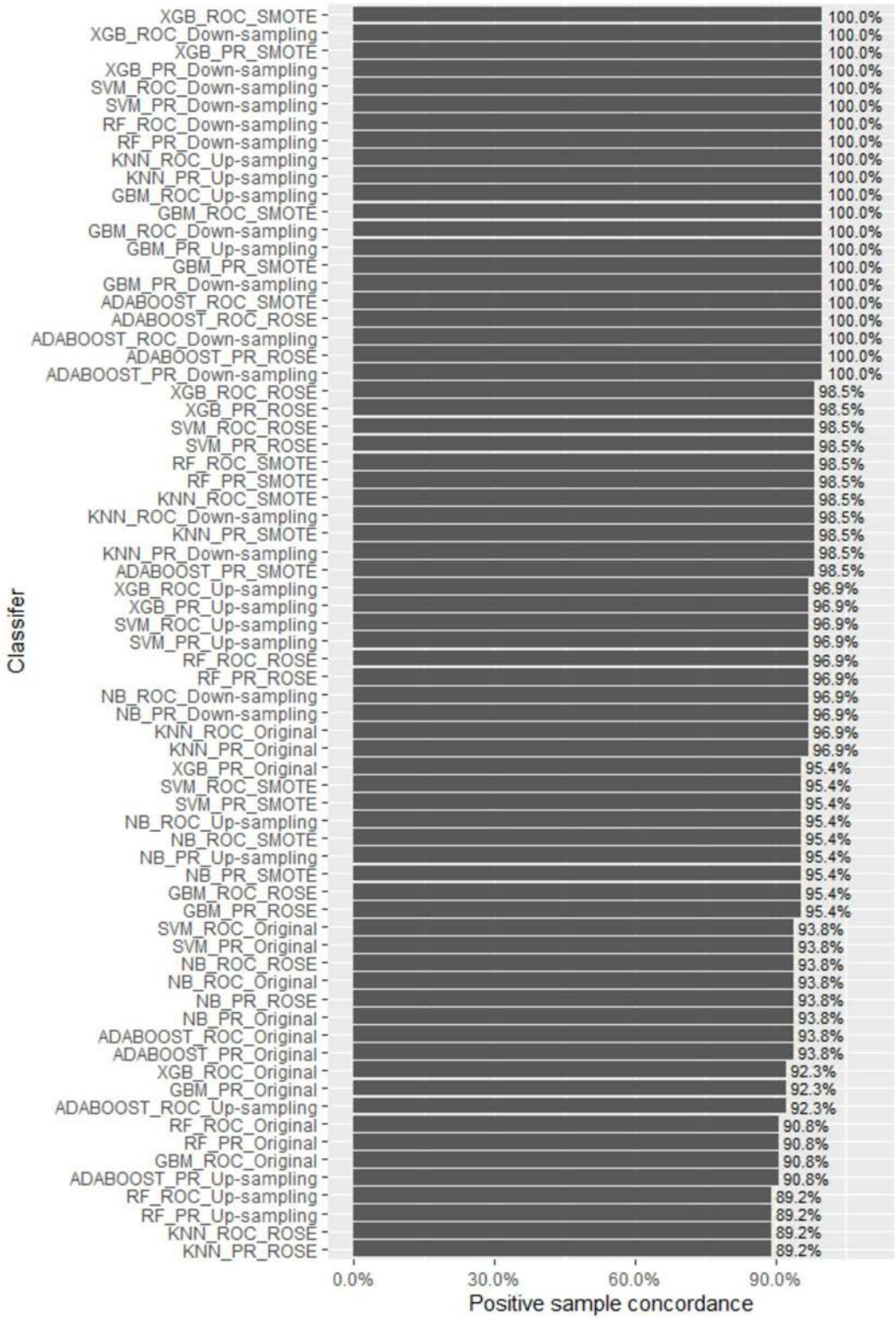

Classifer
XGB_ROC_SMOTE 100.0%
XGB_ROC_Down-sampling 100.0%
XGB_PR_SMOTE 100.0%
XGB_PR_Down-sampling 100.0%
SVM_ROC_Down-sampling 100.0%
SVM_PR_Down-sampling 100.0%
RF_ROC_Down-sampling 100.0%
RF_PR_Down-sampling 100.0%
KNN_ROC_Up-sampling 100.0%
KNN_PR_Up-sampling 100.0%
GBM_ROC_Up-sampling 100.0%
GBM_ROC_SMOTE 100.0%
GBM_ROC_Down-sampling 100.0%
GBM_PR_Up-sampling 100.0%
GBM_PR_SMOTE 100.0%
GBM_PR_Down-sampling 100.0%
ADABOOST_ROC_SMOTE 100.0%
ADABOOST_ROC_ROSE 100.0%
ADABOOST_ROC_Down-sampling 100.0%
ADABOOST_PR_ROSE 100.0%
ADABOOST_PR_Down-sampling 100.0%
XGB_ROC_ROSE 98.5%
XGB_PR_ROSE 98.5%
SVM_ROC_ROSE 98.5%
SVM_PR_ROSE 98.5%
RF_ROC_SMOTE 98.5%
RF_PR_SMOTE 98.5%
KNN_ROC_SMOTE 98.5%
KNN_ROC_Down-sampling 98.5%
KNN_PR_SMOTE 98.5%
KNN_PR_Down-sampling 98.5%
ADABOOST_PR_SMOTE 98.5%
XGB_ROC_Up-sampling 96.9%
XGB_PR_Up-sampling 96.9%
SVM_ROC_Up-sampling 96.9%
SVM_PR_Up-sampling 96.9%
RF_ROC_ROSE 96.9%
RF_PR_ROSE 96.9%
NB_ROC_Down-sampling 96.9%
NB_PR_Down-sampling 96.9%
KNN_ROC_Original 96.9%
KNN_PR_Original 96.9%
XGB_PR_Original 95.4%
SVM_ROC_SMOTE 95.4%
SVM_PR_SMOTE 95.4%
NB_ROC_Up-sampling 95.4%
NB_ROC_SMOTE 95.4%
NB_PR_Up-sampling 95.4%
NB_PR_SMOTE 95.4%
GBM_ROC_ROSE 95.4%
GBM_PR_ROSE 95.4%
SVM_ROC_Original 93.8%
SVM_PR_Original 93.8%
NB_ROC_ROSE 93.8%
NB_ROC_Original 93.8%
NB_PR_ROSE 93.8%
NB_PR_Original 93.8%
ADABOOST_ROC_Original 93.8%
ADABOOST_PR_Original 93.8%
XGB_ROC_Original 92.3%
GBM_PR_Original 92.3%
ADABOOST_ROC_Up-sampling 92.3%
RF_ROC_Original 90.8%
RF_PR_Original 90.8%
GBM_ROC_Original 90.8%
ADABOOST_PR_Up-sampling 90.8%
RF_ROC_Up-sampling 89.2%
RF_PR_Up-sampling 89.2%
KNN_ROC_ROSE 89.2%
KNN_PR_ROSE 89.2%
0.0%
30.0%
60.0%
90.0%
Positive sample concordance

图7B


RF_ROC_Up-sampling 97.0%
RF_PR_Up-sampling 97.0%
ADABOOST_PR_Up-sampling 96.7%
XGB_ROC_Original 96.5%
XGB_PR_Original 96.5%
RF_ROC_Original 96.4%
RF_PR_Original 96.4%
KNN_PR_ROSE 96.2%
KNN_ROC_ROSE 96.2%
GBM_PR_Original 96.1%
GBM_ROC_Original 96.1%
ADABOOST_ROC_Original 95.8%
ADABOOST_PR_Original 95.8%
SVM_ROC_Original 95.8%
SVM_PR_Original 95.8%
KNN_PR_Original 95.6%
KNN_ROC_Original 95.6%
XGB_ROC_Up-sampling 95.4%
XGB_PR_Up-sampling 95.4%
RF_ROC_SMOTE 95.1%
RF_PR_SMOTE 95.1%
ADABOOST_ROC_Up-sampling 94.3%
GBM_ROC_Up-sampling 94.1%
GBM_PR_Up-sampling 94.1%
GBM_ROC_SMOTE 93.8%
GBM_PR_SMOTE 93.8%
ADABOOST_PR_SMOTE 93.5%
XGB_PR_SMOTE 93.4%
SVM_PR_Up-sampling 93.4%
SVM_ROC_SMOTE 93.3%
SVM_PR_SMOTE 93.3%
SVM_ROC_ROSE 93.2%
SVM_PR_ROSE 93.2%
ADABOOST_ROC_SMOTE 93.0%
SVM_ROC_Up-sampling 92.9%
XGB_ROC_SMOTE 92.5%
ADABOOST_ROC_Down-sampling 92.2%
RF_ROC_Down-sampling 92.1%
RF_PR_Down-sampling 92.1%
GBM_ROC_Down-sampling 92.0%
ADABOOST_PR_Down-sampling 91.9%
NB_ROC_Original 91.8%
NB_PR_Original 91.8%
KNN_PR_SMOTE 90.8%
KNN_ROC_SMOTE 90.8%
GBM_PR_Down-sampling 90.7%
KNN_PR_Down-sampling 90.6%
KNN_ROC_Down-sampling 90.6%
NB_ROC_Up-sampling 90.5%
NB_PR_Up-sampling 90.5%
SVM_ROC_Down-sampling 90.5%
SVM_PR_Down-sampling 90.5%
XGB_ROC_Down-sampling 90.0%
XGB_PR_Down-sampling 90.0%
NB_ROC_ROSE 88.8%
NB_PR_ROSE 88.8%
NB_ROC_Down-sampling 88.2%
NB_PR_Down-sampling 88.2%
KNN_PR_Up-sampling 87.7%
KNN_ROC_Up-sampling 87.7%
NB_ROC_SMOTE 87.3%
NB_PR_SMOTE 87.3%
RF_ROC_ROSE 87.3%
RF_PR_ROSE 87.3%
XGB_ROC_ROSE 80.3%
XGB_PR_ROSE 80.1%
GBM_ROC_ROSE 79.6%
GBM_PR_ROSE 79.6%
ADABOOST_ROC_ROSE 77.1%
ADABOOST_PR_ROSE 77.1%
Classifer
0.0%
30.0%
60.0%
90.0%
Screening efficiency

图7C

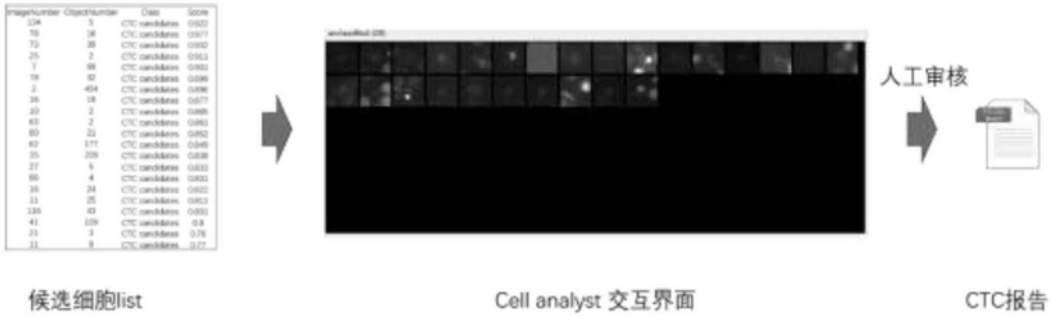


图8