# Down Syndrome Prediction/Screening Model Based on Deep Learning and Illumina Genotyping Array

**基于深度学习和Illumina基因分型阵列的唐氏综合征预测/筛选模型**

# 摘要：

唐氏综合征（DS）是一种遗传性疾病，与21号染色体的基因组剂量失衡和微复制有关。它通常与一组严重疾病有关，包括智力残疾，心脏疾病，身体异常和其他异常。目前，由于人类DS无法治愈，筛查和早期发现已成为DS预防的最有效方法。在这项研究中，我们使用深度学习技术，基于新引入的Illumina基因分型阵列的分析，构建准确的DS预测/筛选模型。具体而言，我们根据Vanderbilt大学医学中心收集的临床基因分型数据构建染色体SNP图谱。然后我们提出了一个卷积神经网络（CNN）架构，其中有10层和两个合并的CNN模型，它们将两个输入染色体SNP图组合在一起。我们的CNN DS预测/筛选模型的平均准确度达到99.3％以上，并且假阳性和假阴性率非常低，这对于医疗实践中的疾病预测和筛查至关重要。与三种常规机器学习算法相比，它在所有评估指标方面也有更好的表现。最后，我们从我们训练的CNN模型的中间层可视化特征映射和训练过的滤波器权重。我们进一步讨论了我们的方法的优势以及其强大性能的基本原因。

关键词 - 深度学习，卷积神经网络，唐氏综合症，Illumina基因分型

# 简介

唐氏综合征（DS）是由人类染色体21（HSA21）的基因组剂量不平衡和微复制引起的遗传疾病[1]。 它通常与智力残疾，先天性心脏缺陷，儿童期白血病，阿尔茨海默病，早期衰老，身体异常和其他异常有关[1-3]。 尽管世界范围内DS的发生率很高（每1,000例活产1例）[4]，并且已经得到很好的研究，但研究人员尚未发现任何有效的治疗方法[5]。 没有发现环境因素或父母的行为因素导致人类DS [6]。 目前，人类DS治疗研究主要集中在早期干预，教育治疗[7,8]，物理治疗[9,10]以及情绪和行为治疗[8,11]。 这些疗法只具有有限的效果，并且从未从根本上治疗DS [5]。 因此，筛查和早期检测已成为人类DS预防的最有效方法。

DS筛查自20世纪60年代开始研究。 一些DS生物标志物已被发现，如甲胎蛋白水平，人绒毛膜促性腺激素和非共轭雌三醇[12,13]。 目前，DS筛查研究包括超声测量胎儿颈部半透明[14]，血液测试[15]，测序测试[16]和组合基因测试[17]。 然而，每16名DS筛查试验阳性妇女中就有1人仍然患有进一步的高风险侵入性诊断程序，如羊膜穿刺术和绒毛膜绒毛取样，其中1/200的机会会导致胎儿丧失[16,18]。

人类第21号染色体（Hsa21）编码超过500个基因[19,20]，包括蛋白质修饰剂，转录因子，RNA剪接因子/修饰剂，细胞表面受体和粘附分子以及各种生化途径成分[20,21]。 然而，只有165个基因被注释为蛋白质编码基因或microRNA。 超过350个基因具有未分配的功能[22]。 最近的GWAS研究发现单核苷酸多态性（SNPs）变异，拷贝数变异和未鉴定的遗传变异与人类DS基因组的遗传疾病高度相关[23-26]。 Illumina推出了新的外显子组基因分型阵列技术，该技术针对人类基因组中罕见的SNP。 范德比尔特大学医学中心和范德比尔特流行病学中心已经开发了用于处理Illumina基因分型阵列的芯片处理协议，并收集了各种疾病的临床数据[24]。

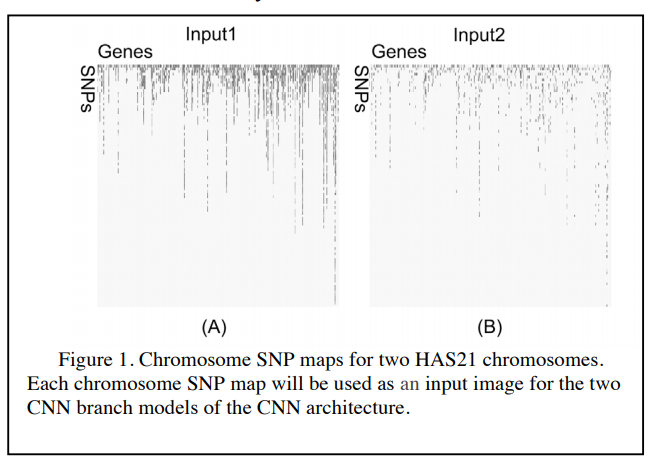
机器学习技术已被广泛应用于疾病预测，疾病诊断和生物标志物鉴定[27-29]。常用的机器学习算法包括传统的人工神经网络，深度学习神经网络，支持向量机（SVM）[27]，随机森林[29]，决策树[30]和贝叶斯分类器[31]。最近，深层卷积神经网络（CNN）和递归神经网络（RNN）已经成功地用于各种疾病筛查，诊断和预测问题[27,32-34]。据我们所知，只有少数传统的机器学习技术已应用于DS研究[20]。他们大多数是在小鼠DS模型上进行[20,24,35]。克拉拉等人。分析了77个蛋白质的表达水平，并设计了一个自组织图来鉴定DS小鼠模型Ts65Dn中的生物学差异[20]。曹等人。使用朴素贝叶斯分类器预测Ts65Dn和Ts1Cje小鼠模型中N-甲基-Daspartate受体处理下的运动活动水平[35]。赵等人。提出了一种独立成分分析的层次约束局部模型来检测年轻儿科患者的DS [36]

**在这项研究中，我们使用深度学习技术从新引入的Illumina外显子组基因分型阵列数据构建准确的人类DS预测/筛选模型。首先，我们根据范德比尔特大学医学中心收集的临床基因分型阵列数据构建了染色体SNP图谱。接下来，我们提出了九层卷积神经网络（CNN）架构和两个合并分支CNN模型，它们将两个染色体SNP图作为输入。我们还使用了三种传统的监督学习算法SVM，随机森林和决策树来构建具有相同基因分型数据的DS预测/筛选模型。我们进一步评估和比较了不同模型的性能，并得出结论：CNN模型在所有模型的所有评估指标中取得了最佳性能。最后，我们从我们训练的CNN模型的中间层可视化特征映射和训练过的滤波器权重。我们进一步讨论了我们方法的优点以及其强大性能的根本原因。**

# 方法

## 数据

我们分析了Illumina外显子组基因分型阵列数据集，该数据集针对罕见的单核苷酸多态性（SNPs）。 Vanderbilt大学医学中心收集了该基因分型数据集，其中包含378个样本，315个非DS患者和63个DS患者。每个样品包含来自321个HSA21编码基因的总共5458个SNP的强度信息。对于CNN模型构建，我们构建了两个染色体SNP图并将它们用作每个样本的输入图像。如图1所示，每个染色体SNP图代表HSA21单个染色体上的所有SNP位点的强度。染色体SNP图的每个像素代表来自一个基因的一个SNP位点的强度。图2给出了两个HSA21染色体的两个输入染色体SNP图谱。染色体SNP图的列表示染色体上的所有相邻基因。行表示相同基因内的相邻SNP位点。对于传统的机器学习模型构造，我们使用了具有5458个SNPs特征的原始Illumina基因分型阵列数据集进行模型训练和测试。所有的SNPs数据被归一化为区间[0,1]。为了构建健壮和可靠的模型，我们通过随机选择75％的样本进行训练，剩下的进行测试，生成了10个并行的训练和测试数据集。然后，我们建立了10个并行DS预测/筛选模型，并计算了平均性能指标，以对本研究中使用的所有模型进行系统评估。



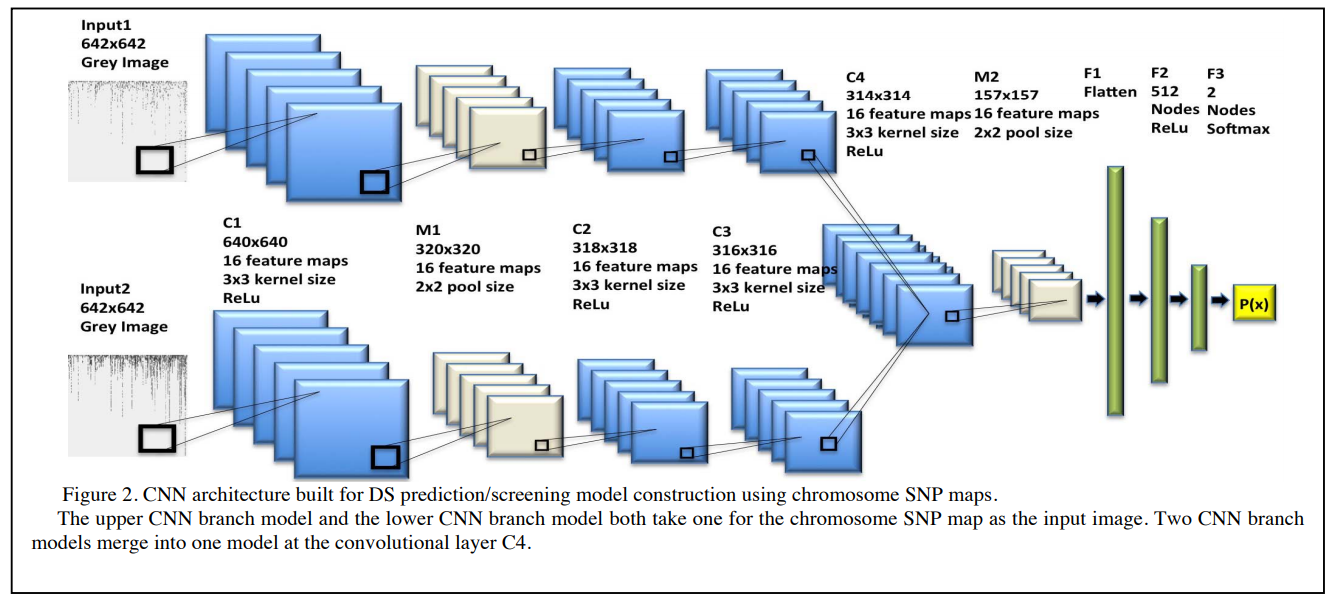
## CNN结构

本研究中构建的CNN架构由两个分支CNN模型合并而成。 每个分支模型包含五个层，包括一个输入层，三个卷积层和一个最大池层。 每个分支模型将一个染色体SNP图作为输入图像。 我们进一步将两个分支CNN模型合并到第6层中的一个模型中，这是另一个卷积层。 图2显示了我们的CNN架构的结构。

本研究中构建的CNN架构由两个分支CNN模型合并而成。 每个分支模型包含五个层，包括一个输入层，三个卷积层和一个最大池层。 每个分支模型将一个染色体SNP图作为输入图像。 我们进一步将两个分支CNN模型合并到第6层中的一个模型中，这是另一个卷积层。 图2显示了我们的CNN架构的结构。 利用二元交叉熵损失函数和随机梯度下降优化器（SGD）编制模型，学习率为0.01,1e-6衰减，动量参数为0.9。 Keras和Tensorflow作为后端实施CNN模型构建和培训，在Ubuntu 14.04.5 LTS上使用NVIDIA GeForce GTX TITAN Pascal 12GB GPU。

## 传统的机器学习算法。

随机森林，SVM和决策树算法由Python和Scikit Learn软件包[37]实现。随机森林算法使用基尼杂质和“熵”来测量特征的分割质量。每个子树的最大深度没有限制，直到其所有的叶子都是纯的或剩余的样本少于两个。决策树是通过分类和回归树算法实现的，该算法与C4.5算法类似，并使用每个节点具有最大信息增益的特征和阈值构建二叉树。它还使用基尼杂质和“熵”来衡量特征的分割质量。最大数量的要素设置为数据集中要素的总数。此外，构建的模型没有深度限制。SVM算法由C-Support Vector Classification算法实现。



# 结果

## 卷积神经网络算法

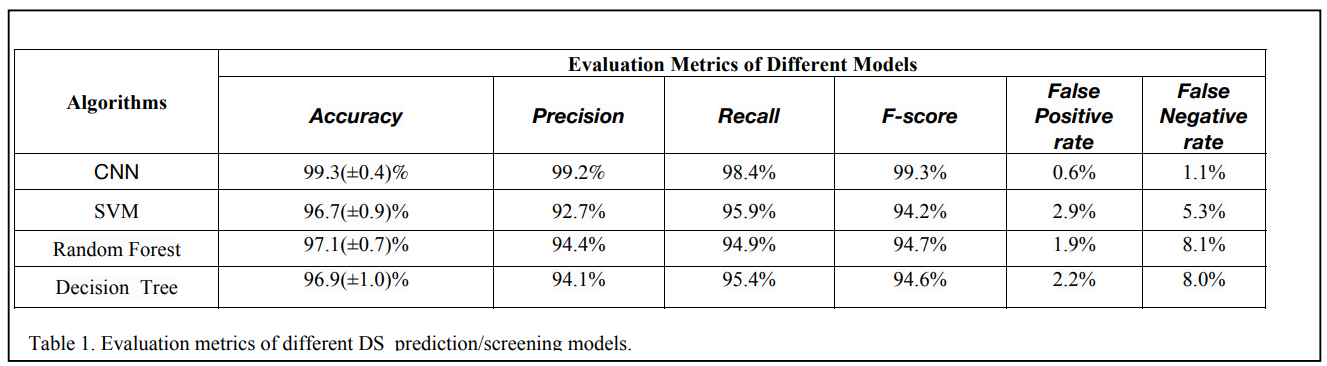
我们的CNN架构从Illumina外显子组基因芯片阵列中提取的染色体SNP图中提取了基因组特征。 对于每位患者，我们构建了HSA21不同染色体的两个染色体SNP图谱。 染色体SNP图的每个像素代表一个SNP位点的强度。 图2显示了用于人类DS预测/筛选的CNN架构。 我们的CNN架构将两个染色体SNP图作为输入图像，并将它们组合到两个分支CNN模型中。 每个分支模型有五个层次，包括一个输入层，三个卷积层和一个maxpooling层。 然后，我们将两个分支模型合并到第四个卷积层（图2，C4）中的一个模型中，其后是一个最大池层和三个完全连接的层。 所有隐藏的图层后面都有Dropout以减少过度拟合。

## 卷积神经网络的DS预测/筛选模型

在本节中，我们使用CNN体系结构（如图2所示）构建具有Illumina基因分型阵列数据的人类DS预测/筛选模型。基因分型阵列数据集包含378个样本，包括63名DS患者和315名非DS患者。首先，我们将所有SNP的强度归一化为区间[0,1]。接下来，针对每个样品，我们基于两个HSA1染色体的SNP构建了两个染色体SNP图谱。如图1所示，染色体SNP图的列表示染色体上的相邻基因。行表示相同基因内的相邻SNP位点。每个像素表示一个基因的一个SNP位点的强度。然后，我们将两个输入图像输入到我们的CNN架构中，进行模型训练和测试。为了建立一个稳健和可靠的模型，我们通过随机选择前75％的样本进行训练并使用剩余的样本进行测试，生成了10个并行的训练和测试数据集。然后，我们建立了10个并行DS预测/筛选模型，并计算出平均绩效指标以提供系统评估。如表1所示，从我们的CNN架构构建的DS预测/分类模型的平均准确率达到了99.3％。它在精度，召回率和F分数方面也取得了非常高的分数，分别为99.2％，98.4％和99.3％。值得注意的是，平均假阳性率和假阴性率仅为0.6％和1.1％。在所有十个并行实验中，只有五个非DS的样品和两个总共DS的样品被误预测为错误的类别。上述结果表明，我们的CNN架构为人类DS预测/筛选构建了鲁棒且准确的CNN模型，其可用于使用SNP染色体图谱进行DS筛选。

## 传统机器学习算法的DS预测/筛选模型

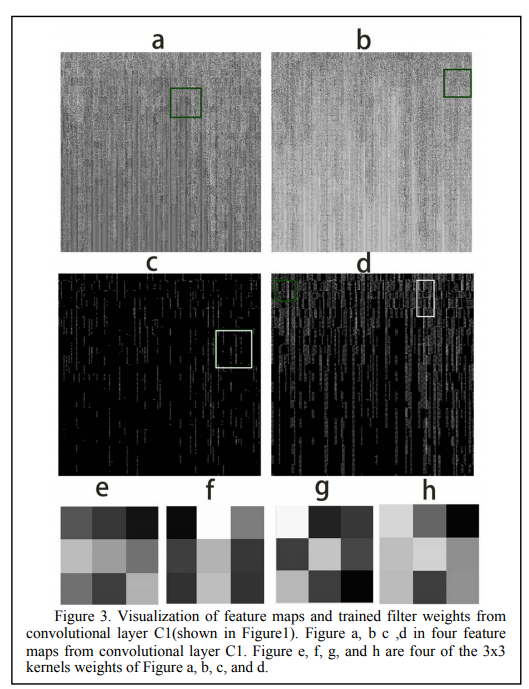
我们评估并比较了CNN模型和三种传统机器学习算法的性能。 我们通过随机选择第一个75％的数据进行训练，剩下的进行测试，从而生成了十个平行的训练和测试数据集。 接下来，我们使用具有总共5458个SNP特征的原始Illumina基因分型阵列数据，使用支持向量机（SVM），随机森林和决策树算法构建DS预测/筛选模型。 学习算法，模型构造和配置由Scikit Learn软件包[37]实现，可以在Methods部分中找到。 最后，我们评估和比较了这三种常规监督学习算法和我们的CNN模型的性能。



如表1所示，从SVM，随机森林和决策树构建的模型可以实现超过96％的平均精度。随机森林模型在所有三种常规监督学习模型中取得了最佳性能，平均准确率为97.1％。 SVM模型的表现最差，平均准确率为96.7％。决策树的表现在其他两个模型之间，平均准确率为96.9％。但是，表1还表明，我们的CNN模型比所有三种常规学习算法获得更高的平均准确率（99.3％）。它也产生了更高的平均精度，召回率和F分数。尽管三种常规模型的准确性得分很高，但其假阴性率却很高，分别为5.3％，8.1％和8.0％。他们高的假阴性率使他们很难进入医疗实践。相比之下，我们的CNN DS预测/筛选模型只有1.1％的平均假阴性率和0.6％的假阳性率。这些研究结果表明，与其他三种常规机器学习模型相比，我们的CNN模型在所有评估指标上都有显着提高的性能。

## 特征映射和训练过滤器权重的可视化

**我们进一步将我们训练的CNN模型中间层的特征映射和相应的训练过滤器可视化。我们讨论了我们的CNN方法的优势以及其强大性能的根本原因。首先，CNN模型以两个染色体SNP图作为输入，并从两个染色体组合提取基因组特征。染色体SNP图的每个像素代表一个SNP位点的强度。我们选择了尺寸为3x3的内核和16个滤镜，这些滤镜不会忽略任何细节信息或相邻像素的图案。图3显示了来自卷积隐藏层C1的输出特征图及其相应的滤波器权重（如图2所示）。如特征图所示，首先，我们的训练模型还可以锐化输入图像，并突出显示信息区和密集区（以绿色矩形显示）。其次，我们的模型可以检测代表连续基因和SNP位点强度的线条（以白色矩形显示）。此外，我们的CNN模型可以突出来自染色体SNP图上小区域的区域视觉模式和独特的局部图案。这些区域模式代表了局部基因组区域内相关的遗传变异，以及相邻基因和SNP位点。因此，我们的CNN模型可以保留最具信息性和关键性的SNP位点，并从输入图像中移除较不重要的位点。它还可以检测综合特征，并从邻近基因组区域提取模式及其同时或因果SNP变异。然而，传统的监督学习算法将所有SNP站点视为独立特征。由于鉴定子集表征和优化的确定性特征的高计算成本，其鉴别力受到限制。此外，他们倾向于从所有可用特征构建具有全球视野的模型，并可能忽略邻近基因和SNP之间的区域模式和相关性。**



# 4．讨论

目前的DS筛查研究包括胎儿颈部半透明[14]，血液测试[15]，测序测试[16]和组合基因测试[17]的超声测量。 DS筛查的主要优点之一是提供无创检测以确定DS的风险。患有低风险的患者可能会决定避免可能导致胎儿丧失的进一步侵入性诊断程序。目前，胎儿基因组的SNP阵列分析可以在妊娠五周时通过非侵入性方法在胎儿滋养层细胞上进行[25,26]。在本研究中，我们的方法将每个SNP位点作为染色体SNP图上的一个像素，这显着降低了图像大小和训练复杂度。 CNN架构的两个CNN分支模型可以将两个输入染色体SNP图谱中相邻基因和SNPs的局部基因组模式特征和相关性结合起来。然而，传统的机器学习算法将所有SNP作为独立特征进行处理。另外，这些算法的判别能力由于在识别子集表征和优化的确定特征方面的计算成本而受到限制。因此，传统的机器学习算法难以检测特征之间的区域模式和相关性。在未来的工作中，CNN和其他基于Illumina基因分型阵列的深度学习技术可以进一步用于疾病预测和基因组模式检测。

先前的研究表明，局部染色体区域中人类疾病基因组中局部基因组和SNP的表达和变异高度相关[38]。 GWAS研究也证明SNPs变异，拷贝数变异和未被鉴定的遗传变异与人类DS高度相关[24-26]。在这项研究中，我们的CNN DS预测/筛选模型利用了综合表征和区域遗传特征提取。如图3所示，卷积层和滤波器可以检测染色体SNP图谱中的遗传模式和独特的局部图案。我们的模型可以突出显示局部基因组区域内的相关遗传变异，以及相邻基因和SNP位点之间的同时或因果变异。我们的研究结果表明，CNN模型不仅能够对人类基因分型阵列进行准确的DS预测，还能对其局部基因组特征进行有效分析。它可以从中间卷积层的输出中找出与人类DS疾病相关的最关键和信息丰富的基因和相应的SNP，从而提供对DS病理基因组学机制的深入了解。目前，DS治疗主要集中在教育治疗[7,8]，物理治疗[9,10]，情绪和行为治疗[8,11]，这些治疗效果有限，无法从根本上治愈DS。传统药物没有显示任何明确的效果或人类DS治疗的好处[39]。我们的CNN模型鉴定的遗传模式，相关基因和SNPs变异为研究与人类DS相关的基因组标记和通路组件提供了机会，这进一步促进了DS基因治疗研究和遗传医学发展。