第 32 卷 第 1 期
2014 年 1 月

天 文 学 进 展
PROGRESS IN ASTRONOMY

Vol. 32, No. 1
Jan., 2014

doi: 10.3969/j.issn.1000-8349.2014.01.01

基于机器学习的热亚矮星光谱分类

寇博凯 [1,2]，冯梦奇 [1,2]，肖化平 [1,2]，雷振新 [1,2*]

(1. 湘潭大学 湖南省高校恒星与星际介质重点实验室，湘潭 411105； 2. 湘潭大学 物理与光电工程学院，湘潭 411105)

摘要： 热亚矮星的形成机制仍不清楚。不同光谱类型热亚矮星的大气参数和表面化学丰度存在较大区别，它们可能有不同的形成起源。随着大批巡天望远镜观测数据的相继释放，新发现的热亚矮星将急剧增加。然而，热亚矮星的光谱分类仍依赖传统的人工方法，这种方法在大批量光谱分析上效率不足。本研究使用卷积神经网络（CNN）提取光谱特征，使用随机森林（RF）来对光谱进行分类。该集成模型对具有 LAMOST 光谱的 1 223 颗热亚矮星进行训练， 实现对六种不同光谱类型热亚矮星的自动分类。该模型在测试集中达到了 90.8% 的总体分类准确率，并且所有分类的 ROC 曲线下面积（AUC）均超过 0.9，显示出较高的可靠性。CNN+RF 集成模型对具有 SDSS 光谱的 1 718 颗热亚矮星进行了分类预测，整体准确率达到了 94%，表明模型对新数据具有较好的泛化能力。该模型可以运用于新发现热亚矮星的自动光谱分类，进而为大批量热亚矮星的统计研究提供保障。

关 键 词： 热亚矮星；光谱分类；机器学习

中图分类号： P145.2 **文献标识码：** A

1 引言

热亚矮星是宇宙中一类非常特殊的小质量恒星，它们的质量大约为 0.5 $M_{\odot}$，表面有效温度可达到 20000 - 70000 K，表面重力加速度为 5.0 - 6.5 cm s^{-2} [1−3]。热亚矮星的大气成分主要为氢（H）和氦（He），根据光谱特征大致可以分成 B 型和 O 型两大类。热亚矮星处于恒星演化的晚期阶段，一般认为是红巨星的后续演化。在赫罗图中，热亚矮星位于水平分支的最蓝端，因此也被称为极端水平分支（EHB）星[4]。

热亚矮星在天体物理多个前沿领域都具有重要的研究意义。首先，热亚矮星一般被认为是在双星系统中形成的 [5]，对热亚矮星进行研究有助于理解和认识双星演化的物理过程[6]；其次，热亚矮星的表面化学元素丰度变化非常大，可以从纯氢变化到纯氦，因此它们是理解

收稿日期： ； **修回日期**：

资助项目：国家自然科学基金 (12073020)；湖南省教育厅重点项目（23A0132）

通讯作者：雷振新，leizhenxin2060@163.com

恒星内部元素扩散过程的理想实验室[7, 8]；再次，利用星震学方法可以研究部分具有脉动特征的热亚矮星，进而了解小质量恒星的内部结构和获取精确的物理参数（如：质量、半径等）[9, 10]。最后，热亚矮星 + 白矮星组成的致密双星系统是很好的引力波源[11, 12] 和可能的 Ia 型超新星前身星[13]。

目前，热亚矮星的形成机制仍不清楚。由于大多数热亚矮星被发现处于密近双星系统中[14, 15]，因此双星演化被认为是其形成的主流渠道。Han 等人利用双星族演化合成方法系统性地研究了双星演化在热亚矮星形成方面的作用[16, 17]。他们发现双星演化中的洛希瓣物质转移、共有包层抛射以及双氦白矮星并合等三种渠道可以分别产生长周期热亚矮星双星、短周期热亚矮星双星以及单个热亚矮星。这些模型的理论预测能解释热亚矮星的大部分观测特征。Rodríguez-Segovi 等人同样用双星族演化合成的方法研究热亚矮星的形成，不同之处在于他们在模型中考虑了热亚矮星薄氢包层的存在[18]，从而更加真实地再现恒星的演化过程。Li 等人发现中小质量的恒星在 AGB 阶段的共有包层抛射过程可以形成较大质量的富氦型热亚矮星[19]，该模型可以解释最近发现的通过共有包层抛射形成的 sdO 型热亚矮星双星系统[20]。而 Ji 等人利用 Ia 型超新星爆炸后的残留伴星来解释中等富氦型热亚矮星的形成[21]，但该模型预测的热亚矮星诞生率较低[22]。

近年来，随着大规模测光和光谱巡天数据的不断更新，大量新的热亚矮星被发现。这些观测数据主要来自斯隆数字巡天 (SDSS)[23] [24]、盖亚卫星巡天 (Gaia)[25]、郭守敬望远镜 (LAMOST)[26–30] 光谱巡天以及凌日系外行星勘测卫星 (TESS)[31–36] 等项目。Geier 等人根据文献中发布的热亚矮星编辑了证认热亚矮星星表[37]，其中包含了 5 613 颗热亚矮星的坐标、视差、星等、大气参数（如：有效温度、重力加速度）等重要信息。随着新发现的热亚矮星不断增加，这一数量在 2020 年增长至 5 874 颗[38]，并在 2022 年进一步增加至 6 616 颗[39]。这些观测成果不仅丰富了人们对热亚矮星的认知，也为研究其物理性质和演化机制提供了更加完整的样本来源。

Bu 等人引入分层极限学习机（HELM）从 LAMOST DR1 数据中筛选热亚矮星候选体，通过多层级框架提取光谱特征并高效分类，识别出约 10 000 颗候选体[40]。在此基础上，Bu 等人又提出 CNN+SVM 混合模型，利用 CNN 提取特征、SVM 分类，在 LAMOST DR4 数据上挑选热亚矮星候选体的表现优于传统方法[41]。Tan 等人改进 CNN 方法，结合八分类与二元分类策略，在 LAMOST DR7-V1 数据集上准确率达 87.42%，并发现 25 颗新热亚矮星[42]。

不同光谱类型热亚矮星具有明显不同的大气参数和表面化学丰度，它们很可能是通过不同演化渠道形成的。根据光谱中 H 线和 He 线的相对强弱，热亚矮星可以进一步细分为贫氦类型：sdB、sdO、sdOB 和富氦类型 He-sdB、He-sdO、He-sdOB 等六类[26, 37, 43, 44]。该方法是目前热亚矮星光谱分类的主流方法。为了使热亚矮星的分类与它们的大气参数、光度、表面 He 丰度等联系起来，同时也为了与正常恒星的 MK 分类衔接起来，Drilling 等人设计了一种包含光谱类、光度类以及 He 分类的三维 MK-like 分类方法[45]。利用该方法，Jeffery 等人对发现的 107 颗富氦型热亚矮星进行了详细的光谱分类[46]，而 Zou 等人对 LAMOST 中发现的 1 224 颗热亚矮星进行了 MK-like 分类[47]。这些分类结果对后续通过统计分析来研

究不同光谱类型热亚矮星的形成起源具有重要作用。

面对观测数据的井喷式增长，新发现的热亚矮星数量将急剧增加。然而，依赖人工分类的传统光谱分类方法存在速度慢、效率低以及对光谱质量要求高等缺陷，因此很难运用于大批量热亚矮星的光谱分析。本文设计了一种卷积神经网络（CNN）和随机森林（RF）的集成机器学习算法，通过对已经证认且有 LAMOST 光谱的上千颗热亚矮星进行了训练和测试，得到了泛化能力较强的热亚矮星自动光谱分类模型。该模型实现了超过 90% 的分类准确率，非常适合应用于大批量热亚矮星的分类工作。本文的章节安排如下：第二部分介绍了数据集的来源、神经网络结构以及模型性能评估指标。第三部分给出分类结果和各项指标的数据，并通过 SDSS 光谱验证了模型的泛化能力。第四部分对分类结果进行了讨论和总结。

2　数据集的构造和神经网络模型结构

2.1　数据集

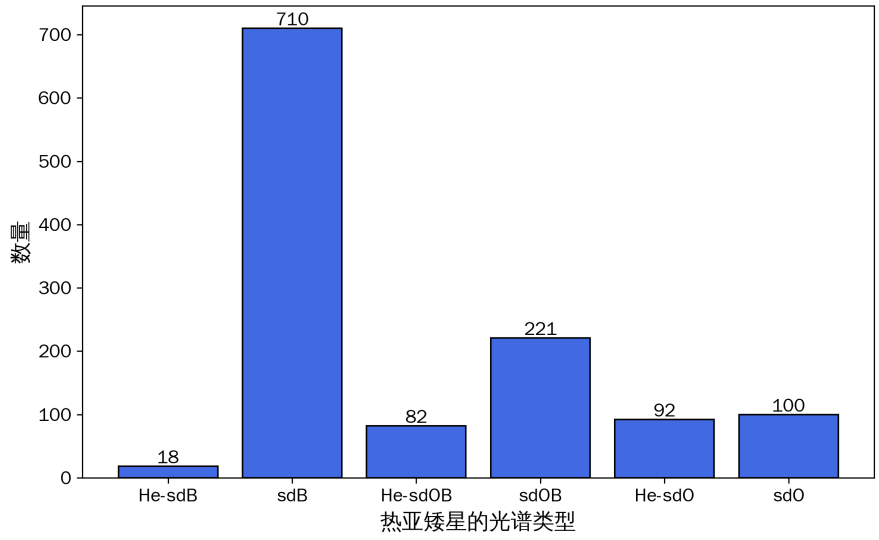

图 1　数据集中不同光谱类型热亚矮星的数量分布。

Culpan 等人基于之前的热亚矮星星表[38]，补充了文献中新发现的热亚矮星，编辑了目前最大的热亚矮星星表[39]。该星表包含了 6 616 颗已经证认的热亚矮星，其中 3 087 颗有大气参数、2 791 颗有视向速度。除此之外，星表还给出了光谱分类、消光、视差等重要参数。该星表已经被广泛运用于热亚矮星领域的研究。

我们将 Culpan 星表与 LAMOST DR9 光谱数据库进行交叉，获得了 1 922 个有 LAMOST 光谱的热亚矮星。为了提升光谱质量，我们选择了信噪比大于 10.0 的光谱进行后续分析，同时剔除了重复观测的源。最终，我们挑选出 1 223 条高质量的热亚矮星的光谱以及它

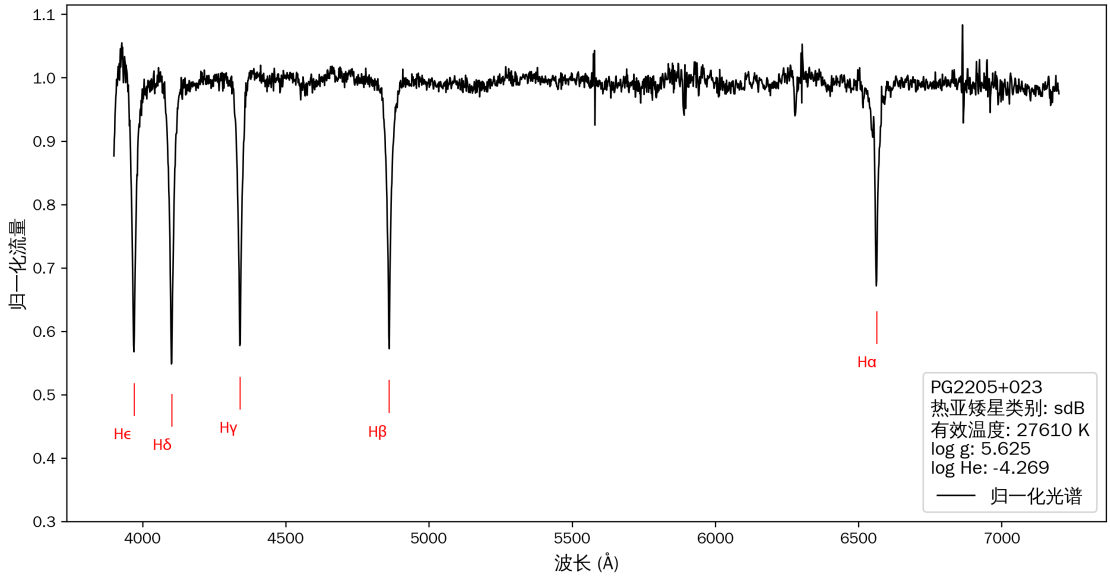

图 2　归一化后的 sdB 光谱及其大气参数。主要的 H 线位置用红色实线标示。

们的光谱分类。我们将其中的 950 条光谱作为训练集，273 条光谱作为测试集。

我们的数据集涵盖了六种光谱类型的热亚矮星，其中 710 颗 sdB、100 颗 sdO、221 颗 sdOB、18 颗 He-sdB、92 颗 He-sdO 和 82 颗 He-sdOB。图1给出了数据集中不同类型热亚矮星的数量分布柱状图。

2.2　光谱预处理

在对数据集进行训练之前，我们采用了样条函数方法对热亚矮星的光谱进行了归一化处理[48]，将光谱流量值统一到 0 - 1 之间。首先，将待处理光谱 $f(w)$ 按波长 w 均匀划分为多个窗口，并在每个窗口内选取流量中值点作为参考点集 $\{(w_i, f_i)\}$。随后，剔除异常点以确保拟合基准点主要来自光谱的本底区域。最终，利用参考点集 $\{(w_i, f_i)\}$ 构建三次样条插值（Cubic Spline Interpolation）来拟合连续谱 $B(w)$，其表达式为：

$$B(w) = a_i + b_i(w - w_i) + c_i(w - w_i)^2 + d_i(w - w_i)^3, \quad w_i \le w \le w_{i+1}, \tag{1}$$

其中 a_i, b_i, c_i, d_i 为样条插值系数。归一化后的光谱由原始光谱与连续谱之比定义：

$$f_{\mathrm{norm}}(w) = \frac{f(w)}{B(w)}. \tag{2}$$

图2展示了一条 sdB 型热亚矮星归一化后的光谱。

2.3　卷积神经网络（CNN）

卷积神经网络是一种在图像识别和特征提取领域广泛应用的深度学习模型[49]。它通过模拟生物视觉系统的分层处理机制，能够自动提取光谱中的关键特征，从而有效提高分类精度。马嘉卉等人通过小波变换寻峰与卷积神经网络相结合的光变曲线自动寻峰方法，实现了对 HXMT 光变曲线中峰的有效检测[50]。

卷积神经网络主要由卷积层、池化层和全连接层构成。其中，卷积层用于提取局部特征，池化层对特征进行降维，以降低计算复杂度并减少过拟合风险。表1给出了本文采用的卷积神经网络结构和参数。它由 2 个卷积层叠加，以充分挖掘光谱数据中的模式信息，并通过 ReLU 激活函数增强非线性建模能力。中间通过 2 个池化层来降低特征维度。此外，为了优化模型性能，我们在网络结构末端加入了一个全连接层，该层对提取到的特征进行整合，并输出用于分类的特征向量。通过这种方式，卷积神经网络能够在大规模光谱数据集中高效提取有价值的特征，为后续的分类任务奠定坚实基础。

表 1　本文采用的卷积神经网络结构和参数设置

层序	层类型	参数

1	卷积层 `conv1` + 激活 `ReLU`	`in_channels=1`
		`out_channels=16`
		`kernel_size=3`

2	最大池化 `pool1`	`kernel_size` = 2

3	卷积层 `conv2` + 激活 `ReLU`	`in_channels=16`
		`out_channels=32`
		`kernel_size=3`

4	最大池化 `pool2`	`kernel_size` = 2
5	展平层 `Flatten`	无参数
6	全连接层 `fc`	`in_features` = 32 × 964 `out_features` = 30848

2.4　随机森林（RF）

随机森林是一种集成学习方法，它通过多个决策树的集成，提升分类的鲁棒性和准确性[51]。相比单一决策树，随机森林能够减少过拟合问题，并对噪声具有更强的抵抗能力。特别是在处理光谱数据时，其随机特征选择机制能够有效提高模型的泛化能力。黄天君等人利用随机森林算法对南极 AST3-2 2016 年观测数据进行分析，挑选出一批变星候选体[52]。

图3展示了随机森林的训练过程，包括 Bootstrap 采样、构建多棵决策树以及最终的投票分类。在训练过程中，每棵决策树基于不同的子数据集进行训练，并通过多数投票的方式决定测试阶段的分类结果。为了优化模型性能，我们通过网格搜索对关键超参数进行了优化，最终确定了树的数量（`n_estimators=150`）、默认的最大深度（`max_depth=None`）以及最小样本分裂数（`min_samples_split=2`）。这些参数的选择显著提升了模型在六分类任务中的稳定性和准确性。这种方式不仅能够提高分类的可靠性，还能有效降低因个别决策树误判带来的影响。

2.5　CNN+RF 集成模型

本文中，我们采用 CNN+RF 集成模型实现高效的热亚矮星光谱自动分类。如图 4 所示，该模型首先利用 CNN 从光谱数据中提取特征，然后将提取的特征输入随机森林进行分类预测。具体而言，模型的输入是归一化后的热亚矮星光谱数据，其特征为形状为 $1 \times 3,862$ 的光谱流量信息。经过多层卷积和池化操作，CNN 将高维光谱数据映射到一个更加紧凑的特

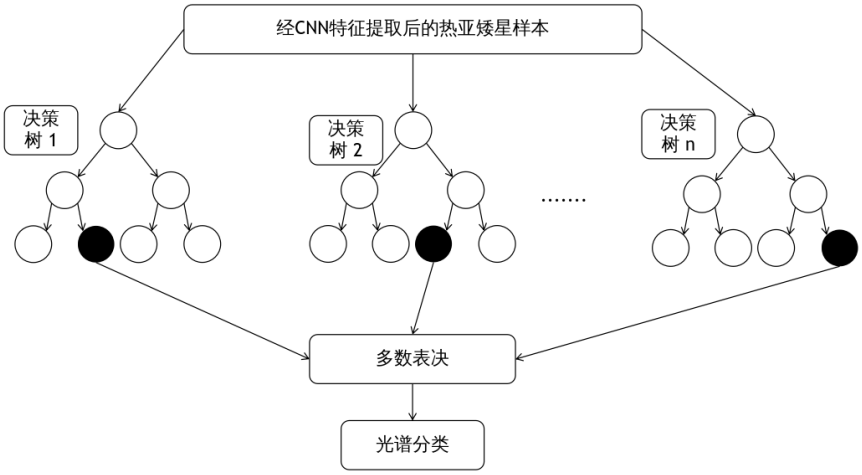

图 3 随机森林分类流程图。

征空间，最终输出形状为 32×964 的特征张量。为了适配随机森林分类器，这些特征被进一步展平。随机森林通过集成多个决策树，对展平后的特征进行分类，并实现对六种类型热亚矮星的精确识别。CNN+RF 集成模型的优势在于：CNN 负责自动提取光谱中的关键特征，避免了人工设计特征的复杂性；同时随机森林通过投票机制提高了分类的稳定性[53]。

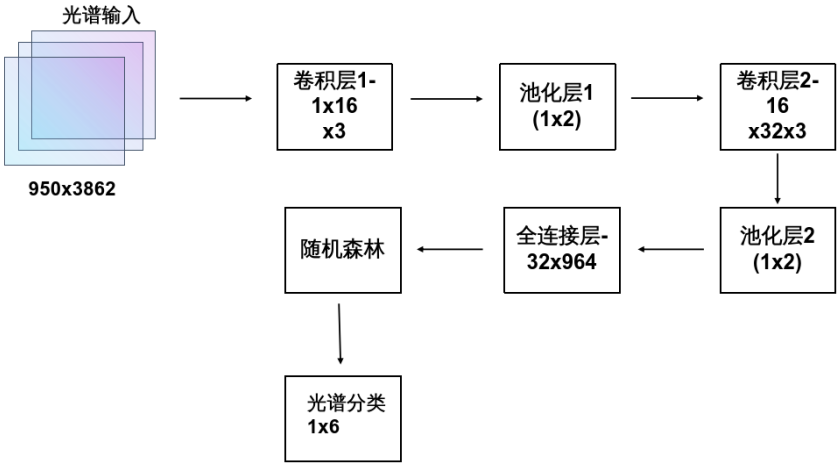

图 4 CNN+RF 集成模型结构

2.6 模型性能指标

为了评估 CNN+RF 集成模型在热亚矮星光谱分类中的性能，我们采用了多种评价指标，包括混淆矩阵、F1 分数、精确率、召回率以及 ROC 曲线下的面积（AUC）。这些指标能够全面反映模型的分类能力。

基于混淆矩阵的基本概念，我们定义以下关键性能指标：

- **精确率（Precision）**：预测为正类的样本中，真正例所占的比例：

$$\text{精确率} = \frac{TP}{TP + FP} \tag{3}$$

- **召回率（Recall）**：真实正类样本中，被正确预测为正类的比例：

$$\text{召回率} = \frac{TP}{TP + FN} \tag{4}$$

- **F1 分数**：精确率和召回率的调和平均数，综合衡量模型的分类效果：

$$F1 = 2 \times \frac{\text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{5}$$

- **总体准确率（Accuracy）**：所有样本中被正确分类的比例：

$$\text{总体准确率} = \frac{\sum TP_i}{\sum (TP_i + FP_i + FN_i)} \tag{6}$$

其中 TP_i、FP_i 和 FN_i 分别表示每个类别的真正例、假正例和假负例。

此外，我们使用 ROC 曲线和 AUC 来评估模型的区分能力。ROC 曲线通过绘制不同阈值下的真正例率（TPR）和假正例率（FPR）来反映模型性能，其中：

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{7}$$

AUC 表示 ROC 曲线下的面积，其值越接近 1，模型的分类能力越强。AUC 的物理意义是随机选择一个正类样本和一个负类样本时，模型正确区分它们的概率。

3　研究结果及讨论

3.1　混淆矩阵分类结果

图 5 通过混淆矩阵清晰地展示了热亚矮星的 6 种光谱类型的真正例（True Positive）数量以及不同类别之间的误分类情况。该图横坐标代表 CNN+RF 集成模型预测的光谱类型，而纵坐标代表真实的光谱类型。右边的颜色条代表不同光谱类型热亚矮星的数量，颜色越深数量越多。

从混淆矩阵可以看出，CNN+RF 集成模型在不同类别的热亚矮星中均具有较高的分类准确率，并且在各类别间的误分类率较低，表明模型的整体识别能力较为可靠。测试集中共有 161 颗 sdB 型热亚矮星，模型准确分类了其中的 157 颗，有 4 颗被误分成了 sdOB。主要原因是 sdB 和 sdOB 的光谱非常相似，都包含了较强的 H 线和较弱的 HeI 线[26, 37, 43]。尽管 sdOB 的光谱中还包含有较弱的 HeII 线，但是对于质量不太好的光谱该特征不容易被模型识别到。这个原因也导致了测试集的 47 颗 sdOB 中有 10 颗被分成了 sdB，2 颗被分成了 He-sdO。其它光谱类别都获得了较高的分类准确率。

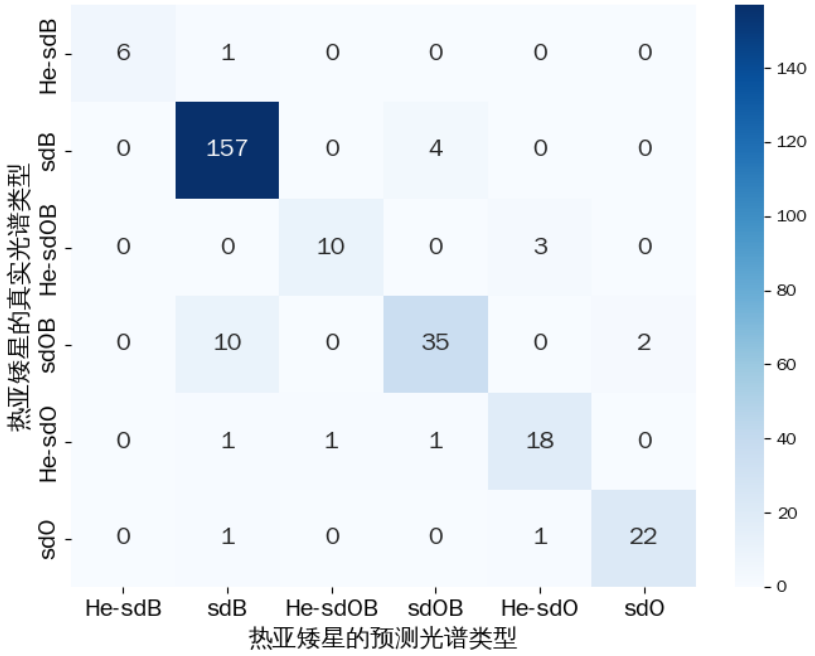

图 5 CNN+RF 集成模型对热亚矮星进行光谱分类的混淆矩阵。横坐标是 CNN+RF 集成模型预测的测试集中热亚矮星的光谱分类情况，而纵坐标是真实的分类情况。右边颜色条表示各类热亚矮星的数量。

表 2 CNN+RF 集成模型给出测试集热亚矮星光谱分类的 F1 分数、精确率和召回率。

类别	F1 分数 (%)	精确率 (%)	召回率 (%)
He-sdB	92.31	100.00	85.71
sdB	94.86	92.35	97.52
He-sdOB	83.33	90.91	76.92
sdOB	80.46	87.50	74.47
He-sdO	83.72	81.82	85.71
sdO	91.67	91.67	91.67

总体准确率：90.84%

3.2 F1 分数、精确率、召回率分类结果分析

为了量化 CNN+RF 集成模型在热亚矮星光谱分类中的能力，表 2 展示了各类别的 F1 分数、精确率(Precision)和召回率(Recall)。整体来看，分类结果的准确率达到了 90.84%。sdB 类型分类的 F1 分数、精确率和召回率基本都排在了所有分类的前面，分别达到了 94.86%、92.35% 和 97.52%。这是由于 sdB 在所有类型的热亚矮星中数目最多（710 颗，见图1），其训练效果最好。令人意外的是，虽然 He-sdB 在训练样本中总共只有 18 颗，但是其分类的 F1 分数、精确率和召回率都达到了 85% 以上。这个主要是因为 He-sdB 型热亚矮型的光谱中以 HeI 线为主，HeII 和 H 线非常弱或者没有。这些特征导致 He-sdB 与其它类型的热亚矮星存在显著区别，因此我们的集成模型能够准确地抓住其特征。与此相反，尽管 sdOB 型热亚矮星的数量在训练样本中居第二（221 颗，见图1），仅次于 sdB 的数量，但是它的 F1 分数、精确率和召回率是最低的，都没有超过 88%。这主要是因为 sdOB 的光谱中的 H 线比较强，同时有较弱的 HeI 和 HeII 线，与 sdB 的光谱相似性较高，从而使得部分 sdOB 被错误地分成了 sdB（见图5和 3.1 节的讨论）。

3.3 ROC 曲线分类结果分析

在 ROC 曲线图中，如果某个类别曲线的面积 AUC 较低（如接近 0.5），说明模型难以正确识别该类别，与其他类别的区分度较低。相反，如果该类别的 ROC 曲线下面积越接近于 1，说明模型对于该类别的光谱特征越能够有效抓取。从分类边界的清晰程度来看，AUC 值越高，表明正负类样本之间的区分能力越强，分类边界越明确。

图6 展示了 CNN+RF 集成模型对测试集中 6 种热亚矮星光谱类型预测的 ROC 曲线图。由图可以看到，该模型对热亚矮星光谱分类的整体准确率达到 90.84%，且所有类别的 AUC 值均超过 0.9。这表明模型不仅能够正确分类样本，还能有效捕捉不同光谱类型之间的细微差异，即使面对样本不均衡的情况，模型仍能保持较高的分类性能，体现了模型的稳健性和泛化能力。

3.4 CNN+RF 集成模型对具有 SDSS 光谱的热亚矮星进行分类

以上结果分析都是基于测试集数据，光谱均来自于 LAMOST 光谱数据库。为了验证 CNN+RF 集成模型的泛化能力，需要通过一些新的光谱数据来进行验证。SDSS 光谱与 LAMOST 光谱具有非常接近的分辨率和光谱质量，非常适合用于对模型预测能力的验证。

我们将 Culpan 星表与 SDSS DR18 光谱数据库进行交叉，获得了 1 718 个有 SDSS 光谱的热亚矮星。其中有 970 颗 sdB、156 颗 sdO、286 颗 sdOB、5 颗 He-sdB、181 颗 He-sdO 和 120 颗 He-sdOB。我们对这 1 718 颗热亚矮星的 SDSS 光谱进行了归一化处理 (见 2.2 节)，并把归一化后的光谱输入到训练好的 CNN+RF 集成模型中进行分类预测。图7 给出了 CNN+RF 集成模型对 1 718 颗热亚矮星分类结果的柱状图。整体来看，所有热亚矮星被分类正确的比例为 94.35%。这说明了该模型在面对新数据时不错的泛化能力。970 颗 sdB 中有 884 颗分类正确，有 86 颗被分成了其它类别，准确率为 91%; 而 172 颗 He-sdOB 热亚矮星中，只有 120 颗分类准确，准确率约为 70%。主要原因是 He-sdOB 的谱线特征与 He-sdO 比较类似，它们的谱线都是 He 线较强，而 H 线比较弱或者没有。其中，He-sdOB 的 HeI 和

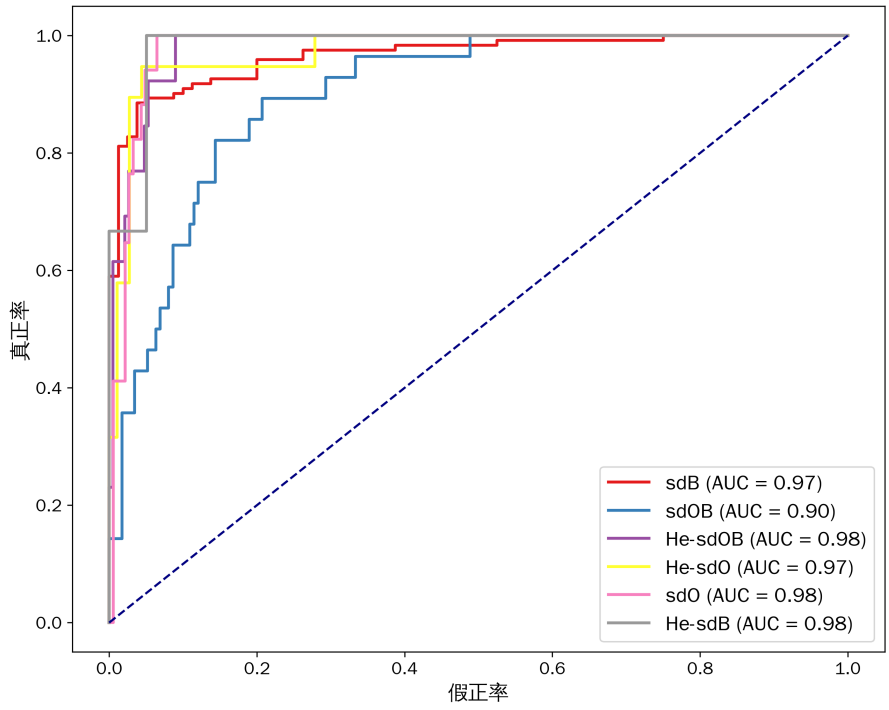

图 6 CNN+RF **集成模型对测试集热亚矮星进行光谱分类的** ROC **曲线和** AUC。

HeII 线都比较明显，而 He-sdO 的 HeII 线要强于 HeI 线[26, 43]。此外，He-sdOB 和 He-sdO 这两种类型热亚矮星在温度上也有区别，他们的连续谱形状会有所不同。而本文对所有光谱进行了连续谱归一化，也会在一定程度上影响这两类热亚矮星的光谱分类准确率。 其它光谱类型如 sdO、sdOB、sdO 和 He-sdB 等都获得 90% 以上的准确率。

4 总结与展望

在本研究中，我们研发了一种专门用于热亚矮星光谱分类的 CNN+RF 集成深度学习模型。我们通过 Culpan 热亚矮星星表和 LAMOST 光谱数据库交叉得到 1223 颗热亚矮星光谱，并构建数据集，其中 950 条光谱为训练集，273 条光谱作为测试集。为了评估模型的分类能力，我们通过光谱分类的混淆矩阵、F1 分数、精确率和召回率等性能指标来评估模型。在测试集中，CNN+RF 集成模型实现了 90.84% 的总体分类准确率。在 SDSS 光谱上的验证结果中，有 94.35% 的热亚矮星被正确分类。

我们的研究结果表明，深度学习能够实现热亚矮星的光谱自动分类，并且达到很高的准确率。相比传统的人工光谱分类方法，我们的模型具备明显的高效率优势，非常适合运用于大批量热亚矮星的光谱分析，从而为统计分析提供重要分类参数。由于不同光谱类型热亚矮

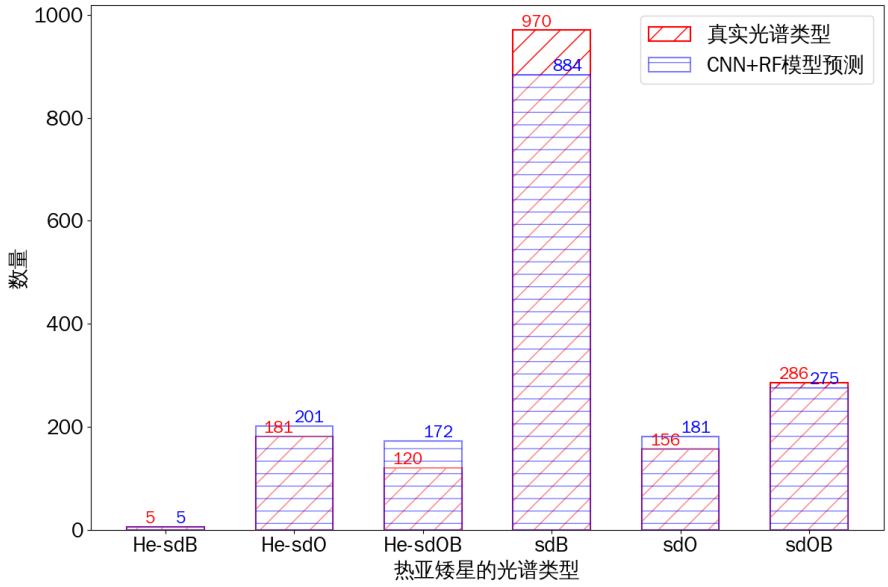

图 7　CNN+RF **集成模型对具有** SDSS **光谱热亚矮星的分类结果统计。红色斜线柱子代表热亚矮星真实光谱分类，而蓝色横线柱子代表** CNN+RF **集成模型预测的光谱分类。**

星的数量有着较大区别，训练样本存在一定程度的不均衡性，导致部分光谱类型的分类准确率不高（如 sdOB）。此外，当前天文学界对热亚矮星的光谱分类尚未形成统一严格的标准，这种分类上的不确定性可能也是导致模型分类精度下降的潜在原因。后续可以通过增加新证认热亚矮星的光谱数据，提升训练样本的均衡性，同时尽量使用具有相同光谱分类标准的训练样本，从而达到更高的分类准确率。

参考文献:

[1] Heber U. A&A, 1986, 155: 33–45
[2] Heber U. ARA&A, 2009, 47(1): 211–251
[3] Heber U. PASP, 2016, 128(966): 082001
[4] Catelan M. Ap&SS, 2009, 320(4): 261–309
[5] Pelisoli I, Vos J, Geier S, et al. A&A, 2020, 642: A180
[6] Ge H, Tout C A, Webbink R F, et al. ApJ, 2024, 961(2): 202
[7] Byrne C M, Jeffery C S, Tout C A, et al. MNRAS, 2018, 475(4): 4728–4738
[8] Li Y, Chen X H, Xiong H R, et al. ApJ, 2018, 863(1): 12
[9] Kawaler S D, Reed M D, Østensen R H, et al. MNRAS, 2010, 409(4): 1509–1517
[10] Zong W, Charpinet S, Fu J N, et al. ApJ, 2018, 853(2): 98
[11] Kupfer T, Korol V, Shah S, et al. MNRAS, 2018, 480(1): 302–309
[12] Kupfer T, Korol V, Littenberg T B, et al. ApJ, 2024, 963(2): 100
[13] Wang B, Meng X, Chen X, et al. MNRAS, 2009, 395(2): 847–854

[14] Maxted P F L, Heber U, Marsh T R, et al. MNRAS, 2001, 326(4): 1391–1402
[15] Copperwheat C M, Morales-Rueda L, Marsh T R, et al. MNRAS, 2011, 415(2): 1381–1395
[16] Han Z, Podsiadlowski P, Maxted P F L, et al. MNRAS, 2002, 336(2): 449–466
[17] Han Z, Podsiadlowski P, Maxted P F L, et al. MNRAS, 2003, 341(2): 669–691
[18] Rodríguez-Segovia N, Ruiter A J, Seitenzahl I R. , 2025, 42: e012
[19] Li Z, Zhang Y, Chen H, et al. ApJ, 2024, 964(1): 22
[20] Li J, Onken C A, Wolf C, et al. MNRAS, 2022, 515(3): 3370–3382
[21] Ji R J, Meng X C, Liu Z W. Research in Astronomy and Astrophysics, 2024, 24(5): 055003
[22] Meng X C, Luo Y P. MNRAS, 2021, 507(3): 4603–4617
[23] Abolfathi B, Aguado D S, Aguilar G, et al. ApJS, 2018, 235(2): 42
[24] Kepler S O, Pelisoli I, Koester D, et al. MNRAS, 2019, 486(2): 2169–2183
[25] Gaia Collaboration, Vallenari A, Brown A G A, et al. A&A, 2023, 674: A1
[26] Lei Z, Zhao J, Németh P, et al. ApJ, 2018, 868(1): 70
[27] Lei Z, Zhao J, Németh P, et al. ApJ, 2019, 881(2): 135
[28] Lei Z, Zhao J, Németh P, et al. ApJ, 2020, 889(2): 117
[29] Lei Z, He R, Németh P, et al. ApJ, 2023, 942(2): 109
[30] Luo Y, Németh P, Wang K, et al. ApJS, 2021, 256(2): 28
[31] Uzundag M, Vučković M, Németh P, et al. A&A, 2021, 651: A121
[32] Krzesinski J, Balona L A. A&A, 2022, 663: A45
[33] Schaffenroth V, Pelisoli I, Barlow B N, et al. A&A, 2022, 666: A182
[34] Krzesinski J, Şener H T, Zola S, et al. MNRAS, 2022, 516(1): 1509–1523
[35] Schaffenroth V, Barlow B N, Pelisoli I, et al. Astronomy & Astrophysics, 2023, 673: A90
[36] Sahoo S K, Baran A S, Worters H L, et al. MNRAS, 2023, 519(2): 2486–2499
[37] Geier S, Østensen R H, Nemeth P, et al. A&A, 2017, 600: A50
[38] Geier S. A&A, 2020, 635: A193
[39] Culpan R, Geier S, Reindl N, et al. A&A, 2022, 662: A40
[40] Bu Y, Lei Z, Zhao G, et al. ApJS, 2017, 233(1): 2
[41] Bu Y, Zeng J, Lei Z, et al. ApJ, 2019, 886(2): 128
[42] Tan L, Mei Y, Liu Z, et al. ApJS, 2022, 259(1): 5
[43] Moehler S, Richtler T, de Boer K S, et al. A&AS, 1990, 86: 53
[44] He R, Meng X, Lei Z, et al. A&A, 2025, 693: A121
[45] Drilling J S, Jeffery C S, Heber U, et al. A&A, 2013, 551: A31
[46] Jeffery C S, Miszalski B, Snowdon E. MNRAS, 2021, 501(1): 623–642
[47] Zou X, Lei Z. PASJ, 2024, 76(5): 1084–1097
[48] 罗锋, 刘超, 赵永恒. 天文研究与技术, 2019, 16(03): 300–311
[49] LeCun Y, Boser B, Denker J S, et al. Neural Computation, 1989, 1(4): 541–551
[50] 马嘉卉, 马森, 邹自明, et al. 天文学进展, 2022, 40(04): 575–589
[51] Breiman L. Machine Learning, 2001, 45(1): 5–32
[52] 黄天君, 孙天瑞, 胡镭, et al. 天文学报, 2019, 60(05): 99–115
[53] Liang D D, Liang D D, Pomeroy M J, et al. In: Chen W, Astley S M, eds. Medical Imaging 2024: Computer-Aided Diagnosis. 2024. . [S.l.]: [s.n.] , Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 12927

Machine learning-based classification of hot subdwarfs

Kou Bokai [1,2], Feng Mengqi[1,2], Xiao Huaping [1,2], Lei Zhenxin[1,2*]

(1. Xiangtan University, Key Laboratory of Stars and Interstellar Medium, Xiangtan 411105; 2. Xiangtan University, College of Physics and Optoelectronic Engineering, Xiangtan 411105)

Abstract: The formation mechanism of hot subdwarf stars are still unclear. The huge differences of chemical abundance among hot subdwarfs with different spectral types indicate their different formation origins. With the new dataset release from photometric and spectroscopic suveys, the number of newly discovered hot subdwarf stars will increase fast. However, traditional spectral classification of hot subdwarfs depends on manual insepection that it is not suitable for the analysis of a large numer of stars. In this study, we employed a machine learning model, which integrates a convolutional neural network (CNN) and a random forest (RF), to classify hot subdwarf stars with different spectral types automatically. The CNN+RF model have a total accuracy of 90.8% for all the types of hot subdwarf stars in the validation dataset of LAMOST spectra. We also used the trained model to classify hot subdwarfs with SDSS spectra, and got a total accuracy of 94%. These resluts indicate a good generalization ability of the model. The CNN+RF deep learning model could be used to efficiently classify newly discoverd hot subdwarf stars in a large size of sample.

Key words: Hot subdwarfs; spectral classification; machine learning