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

机器学习在测光红移估算中的应用

陆君豪，罗智坚，陈建珍，曾璐瑾，束成钢

（上海师范大学 上海市星系与宇宙学半解析研究重点实验室，上海 200234）

摘要： 测光红移（photometric redshift，photo-z）是天文学中通过多波段测光数据估算星系及类星体红移的重要方法。随着观测数据呈指数级增长，传统光谱红移（spectroscopic redshift，spec-z）的测量效率已难以满足当前及下一代大规模巡天项目对海量天体红移信息的需求。目前绝大多数以成像为主的星系与类星体巡天项目都高度依赖测光红移，以获取数千万至数十亿天体的红移信息，从而支持宇宙大尺度结构、暗能量性质等前沿科学研究。在此背景下，机器学习（Machine Learning, ML）方法因其高效性和可扩展性，已成为获取高精度测光红移的主流工具。系统综述了当前机器学习在测光红移领域的最新研究进展，包括算法分类、模型优化策略及典型应用场景，并结合实际案例，分析了不同机器学习架构的技术特点与性能差异。

关 键 词： 测光红移；机器学习；数据分析

中图分类号： P157.2 **文献标识码：** A

1 引 言

当前星系宇宙学的重大科学问题，无论从星系的形成与演化路径、暗物质空间分布及物理属性，再到宇宙大尺度结构的形成演化规律，其突破都依赖于对大样本星系距离的精确测定。这种距离测量传统上主要通过观测由宇宙膨胀引起的光谱红移来实现。受实际观测条件限制，特别对于极暗弱天体或大规模星系样本，即使当前先进的多目标光纤和无缝光谱技术在观测中得到快速发展，光谱观测仍因其耗时费力而难以高效工作。这一局限性直接推动了测光红移这一替代性估计方法的诞生。该方法最早由 Baum[1] 提出理论雏形，经 Butchins[2] 体系化改进，Connolly 等人[3] 建立起了完整的理论框架。测光红移方法基于星系多波段测光数据所得的内禀能量分布与红移的统计关联，实现了无需光谱观测即可估算天体距离，为利用大规模多色巡天资料开展星系宇宙学研究提供了关键方法支持。

收稿日期：xxxx-xx-xx； **修回日期**：xxxx-xx-xx

资助项目：国家自然科学基金 (12141302, 12573009)；中国载人航天工程巡天空间望远镜专项科学研究 (CMS-CSST-2025-A05, CMS-CSST-2025-A07)

通讯作者：罗智坚，zjluo@shnu.edu.cn

在测光红移研究史上，真正具有里程碑意义的是斯隆数字巡天（Sloan Digital Sky Survey, SDSS）。作为首个实现多波段测光与光谱协同观测的大规模数字化巡天项目[4]，SDSS 不仅革新了天体数据采集模式，更为星系距离测量方法的多元化发展提供了前所未有的实验平台。这项开创性项目使得天文学家得以系统性地探索各类测光红移方法的有效性，其成功经验直接推动了后续一系列重大巡天项目的实施，包括宇宙演化巡天（Cosmic Evolution Survey, COSMOS）[5]、暗能量巡天（Dark Energy Survey, DES）[6]、千平方度巡天（Kilo Degree Survey, KiDS）[7]、超广角主焦相机巡天（Hyper Suprime-Cam Survey, HSC）[8]、詹姆斯 · 韦伯太空望远镜（James Webb Space Telescope, JWST）[9]、欧几里得太空望远镜（Euclid Space Telescope）[10]、薇拉 · 鲁宾天文台时空遗珍巡天（Legacy Survey of Space and Time, LSST）[11]，以及即将运行的南希 · 格雷斯 · 罗曼太空望远镜（Nancy Grace Roman Space Telescope, Roman）[12]、中国空间站巡天空间望远镜（Chinese Space Station Survey Telescope, CSST）[13] 等。在此背景下，天文学家通过整合海量星系样本的高精度多波段测光数据与先验光谱库，成功开展了精密宇宙学的研究并取得了令人瞩目的成就，标志着测光红移正式迈入大规模工程化应用阶段。

测光红移方法的原理在于星系能谱受宇宙学膨胀效应影响会向红端发生移动，导致在特定测光系统中，相同类型星系因不同距离其多波段星等（magnitude）与颜色（color）将表现出可区分的测光特征。典型的如 Lyman-α 星系（Lyα Emitters，LAEs）[14]，在高红移时因宇宙膨胀导致 Lyα 发射线红移到可见光或近红外波段而被观测到。然而，实际的应用远非如此简单，从测光参数空间到红移空间的映射函数具有高度非线性特征，其复杂性源于多重因素耦合作用，包括不同星系形态、大尺度结构环境、演化阶段等。

目前的测光红移估计方法体系可归纳为两大类：

1. 光谱能量分布（Spectral Energy Distribution, SED）模板拟合法（Template Fitting）[15, 16]：通过比对观测数据与理论模板或观测模板的匹配度推算红移值。理论模板[17] 基于恒星演化模型合成理论光谱，观测模板[18] 直接采用实测星系光谱构建模板库。该方法的优势在于物理可解释性强，但高度依赖模板库的完备性，且易受星际消光效应与仪器响应函数等不确定性的影响。
2. 机器学习（Machine Learning，ML）驱动型经验方法[19–22]：这类数据驱动型方法通过建立测光参数空间到红移空间的非线性映射实现预测。监督学习基于光谱验证样本训练回归模型，无监督学习利用自组织映射（Self-organizing Mapping，SOM）等方法在无标注数据中挖掘测光特征与红移的潜在关联，混合增强策略结合迁移学习与半监督学习缓解光谱数据稀缺问题。

相较于前者，机器学习方法在大数据处理效率与复杂模式捕捉能力方面表现突出，但面临物理可解释性弱、外推泛化能力受限等挑战。值得注意的是，最新研究趋势显示，两类方法正逐步走向融合，例如将 SED 模板、星系图像、星系形态、星系质量等作为物理约束项嵌入神经网络损失函数，形成物理信息机器学习（physics-informed ML）新型架构[23]。需要说明的是，SED 模板拟合方法同样存在着丰富多样的变体形式，多种混合方法如结合了贝

叶斯推断与嵌套采样技术[24]，进一步拓展了这类方法的应用维度。

精确测定星系的基本物理参数是大规模巡天的科学目标之一。理论上机器学习方法不仅能够估计红移，还可进一步推演恒星形成率（Star Formation Rate，SFR）和星系恒星质量[25]等关键物理量。例如，Bonjean 等人[26] 基于 SDSS Data Release 8 (DR8) 光谱数据训练随机森林模型，成功实现广域红外巡天望远镜 (Wide-field Infrared Survey Explorer, WISE) [27] 近红外源样本的 SFR 和星系恒星质量同步估算；Mucesh 等人[28] 将该方法应用于 DES 巡天数据，在测光参数受限情况下仍展现出良好性能；Euclid 团队[29] 探讨了利用模板拟合和机器学习方法（CatBoost、深度神经网络等）从模拟的 Euclid 卫星观测数据中恢复红移、星系恒星质量和 SFR 等物理参数的性能，并针对宽视场巡天 (Euclid Wide Survey, EWS) 和深场巡天 (Euclid Deep Fields, EDF) 的数据特性，比较了配对标签训练和混合标签训练策略的优劣，发现对于宽场巡天，利用模板拟合法训练深场数据得到的标签并结合宽场特征可得到最佳的结果[29–31]。这些成果充分表明，机器学习能够有效挖掘测光红移与星系物理属性之间的复杂关联，为 Euclid、CSST 等下一代大规模巡天项目的科学产出奠定关键基础。

本文聚焦于测光红移的机器学习估算方法，系统性地分析了各类算法的优势与局限性。通过对该领域关键问题的综合性论述，旨在为天文学家构建更为高效与精准的红移估算体系提供理论依据与实践指导，进而服务于 Euclid、CSST 等新一代巡天项目的科学探索，提升其科学价值。

2 机器学习方法

当前星系测光巡天的观测速度已远超光谱跟踪观测的承载能力，如 LSST 预计每年观测 3700 万星系，而光谱观测仅能覆盖约 0.1%，因此测光红移方法已成为河外天文领域不可或缺的研究工具。研究表明，机器学习方法在光谱训练集覆盖区间内展现出的预测精度可达 $|\Delta z|/(1+z_{\rm spec}) < 0.01$，其中 $\Delta z = z_{\rm phot} - z_{\rm spec}$ 为预测值与真实值之差；但该方法在外推至训练集未覆盖参数空间时（如 $z > 2.5$），预测性能准确性呈现系统性衰减[32]。尽管如此，机器学习方法在特殊天体研究中也取得了重大进展，例如在类星体红移估计中，随机森林算法成功将离群率降低至传统方法的 1/3[33]；广义相加模型则成功预测了 276 个长时伽马射电暴的红移[34]。这些创新方法推动了测光红移从单纯的“统计工具”向“物理诊断仪器”转变。当前该领域代表性方法和基本原理小结如下，首先是以经典机器学习为主的方法：

1. **自组织映射**[35–37]：SOM 通过无监督学习将高维测光数据投影到二维网格上，使相似测光特征的天体在网格中相邻分布，进而结合少量已知红移的标记节点或后续回归方法，推断未标记节点天体的红移。该方法尤其适用于大规模测光红移的批量校准，但对节点边界样本的敏感性较高。由于 SOM 可基于节点内已知红移样本进行局部校准（如均值或回归修正）[38]，并通过拓扑保持离散化映射，从而有效识别数据分布来修正偏差达到降低系统误差的目的。该方法适用于海量数据的可视化分析和测光红移初步估计。
2. **监督前馈神经网络**[22, 39–46]：监督式前馈神经网络 (Supervised Feed-forward Neural Net-

works）通过输入天体的多波段测光数据（如星等、颜色），利用带标签（已知光谱红移）的训练集，学习非线性映射关系，经隐藏层特征提取后输出红移预测值，再通过反向传播优化参数以最小化预测误差，实现从测光数据到红移的回归建模。

3. **自适应检测与去除异常测光或光谱数据方法**[47–50]：自适应神经网络（Self-adaptive Neural Networks）通过动态调整网络权重或结构，结合异常评分机制，如重构误差、预测偏差等，在线学习正常测光数据的分布特征，识别偏离模式或噪声干扰的异常点，并自适应优化剔除高于阈值的数据来实现高鲁棒性的异常值过滤。如 Cavuoti 等人[51] 在损失函数中引入红移-颜色先验关系，通过嵌入物理项约束的方式提高预测精度。
4. **支持向量机**[19, 52]：支持向量机（Support Vector Machine, SVM）通过在高维特征空间中寻找最优超平面，最大化不同红移类别间的间隔，从而实现对测光红移的分类或回归预测。其核函数（如高斯核）可非线性映射测光波段数据，降低红移与光度特征间的复杂简并关系，尤其适用于小样本高维数据。SVM 常结合模板拟合法或光谱特征来提升对低信噪比数据的鲁棒性。核函数的选择对该方法性能影响显著且计算复杂度随样本量呈超线性增长。
5. **基于树的方法**[21, 53–58]：包括随机森林（Random Forest, RF）、梯度提升决策树（Gradient Boosting Decison Tree，GBDT）、Catboost 等一系列变体，兼具可解释性与处理缺失数据能力。其中 RF[21, 59] 通过构建多棵决策树对天体的测光数据进行投票或平均预测，利用自助采样（bagging）降低方差并增强泛化能力。每棵树会随机选择特征和样本来减少过拟合效应，从而稳定处理高维、非线性的测光数据。其天然的特征重要性评估还能帮助筛选关键波段，提升预测精度；GBDT[57, 60] 则通过迭代训练多个弱决策树，每一棵树专注于修正前一棵树的预测残差，逐步优化测光数据与红移之间的非线性关系。其结合提升（boosting）策略和正则化技术（如学习率、树深度控制），有效降低偏差并抑制过拟合，从而在高噪声、多特征的测光数据中实现高精度红移预测，且特征重要性分析还可揭示关键波段对红移估计的贡献；CatBoost [54] 则通过对称决策树（oblivious trees）和有序提升（ordered boosting）处理天文学测光数据中的类别特征（如滤波器波段）和数值特征，有效减少过拟合。其内置的特征排列和自适应学习率机制能自动优化红移预测模型，特别适用于高维、含噪声的测光数据，同时通过目标变量统计高效编码分类变量，提升红移回归的精度和鲁棒性。
6. k **近邻算法**[20, 61–64]：k 近邻（k-Nearest Neighbours, kNN）通过计算目标天体与训练集中天体的测光数据之间的相似度（如欧氏距离），选取最邻近的 k 个样本，以它们的红移均值或加权值作为预测结果。该方法无需显式建模，直接依赖局部数据分布，适用于非线性关系，但对特征缩放和样本密度敏感，且计算复杂度随数据量增长显著。
7. **高斯过程**[65, 66]：高斯过程（Gaussian Process, GP）通过假设测光数据与红移之间的映射关系服从一个高斯随机场，利用协方差函数（如径向基核）刻画不同天体测光特征的相似性，从而预测红移及其不确定性。其非参数特性可灵活拟合复杂的非线性关系，并自然提供预测置信区间，尤其适用于小样本高精度红移估计，但对计算资源需求较高。
8. **混合密度网络**[67, 68]：混合密度网络（Mixture Density Networks，MDN）结合神经网络

和高斯混合模型，通过神经网络学习测光数据到红移的多模态条件概率分布而非单一预测值。其输出为多个高斯分布的参数（权重、均值、方差），可捕捉红移估计中的复杂非线性关系及不确定性，且提供全面的概率化预测结果，尤其适用于存在多解或噪声干扰的数据。

随着计算机性能的发展以及对红移精度要求越来越高的背景下，深度学习开始被广泛应用在测光红移的估算中，代表性方法有：

9. **联合图像的神经网络**[69–73]：适用于联合图像的神经网络有很多模型，包括深度神经网络（Deep Neural Networks，DNN）、卷积神经网络（Convolutional Neural Networks，CNN）、循环神经网络（Recurrent Neural Network，RNN）和长短期记忆网络（Long Short-Term Memory，LSTM）等。这些模型会直接端到端地学习天体多波段测光图像与红移之间的复杂映射，通过卷积层自动提取局部像素模式（如星系形态、颜色梯度）和全局分布特性。结合多任务学习或注意力机制，可同时建模图像与红移的非线性关系，显著提升预测精度，尤其适用于高维、结构化的测光图像数据，超越传统基于表格型特征的方法。
10. **混合应用机器学习方法**[50, 74–76]：混合机器学习方法（Hybrid methods）通过结合不同算法的优势（如 SVM 的核技巧、随机森林的特征选择、神经网络的非线性拟合），以级联、堆叠或加权融合的方式协同建模测光数据与红移的复杂关系，从而在精度、鲁棒性和不确定性量化上超越单一模型，尤其适用于多模态分布的天文数据，可用于对测光红移精度要求较高的宇宙学研究。如林秋帆等人[50] 利用有监督的对比学习 (Supervised Contrastive Learning, SCL) 和 kNN 算法来构建和校准原始红移概率密度（Probablity Distribution Function, PDF）估计。

以上机器学习方法可大致分为无监督学习（[1]）和有监督学习（[2] ∼ [10]）两种，其中有监督学习方法应用于测光红移预测的前提在于必须建立多波段测光数据与天体距离间的复杂映射关系模型。如果训练集的红移分布能连续覆盖目标区间，颜色-星等空间采样密度充分，特殊天体类型比例具有代表性，那么机器学习模型可展现出优越的红移预测性能[30, 32]。但如前文所述，这种高精度预测能力严格受限于训练集定义的特征空间边界，当预测对象超出训练样本的参数覆盖范围，如更高红移或更极端颜色特征，模型性能将呈现系统性衰减。值得注意的是，深度学习（[9]）相比于传统机器学习（[1] ∼ [8]）表面看都是解决问题的工具，但在内核上却是极大的跨越。数学上可以证明，包含至少一个隐藏层的神经网络可以拟合任意复杂函数，传统机器学习受限于浅层模型只能逼近简单函数，而深度学习通过增加层数拟合能力可获得指数级提升。两者在测光红移估计中的最根本区别在于对输入特征的处理方式，深度学习的优势在于其端到端的学习模式和“自动特征表示学习”能力，它能够通过多层非线性变换直接从更原始的数据形态中自动学习和提取最优的、用于预测红移的层次化特征，从而绕过或极大简化了手动特征工程的步骤，将主要任务从为算法设计最好的特征转变为设计最好的模型架构。

除了对训练集覆盖范围的依赖外，机器学习方法还面临另一个挑战，即对数据缺失问题

的处理。当训练集中存在大量缺失数据时，多数模型的结果往往会产生系统性偏差。这种偏差的产生机制源自于模型需要定义有效的“度量距离”，而该距离若要正确运作，必须基于具有相同维度的输入空间。该问题的复杂性在天文学领域尤为突出，因为“数据缺失”并非单一概念。它可分为两类：一是由于观测限制如探测器坏点、天体位于观测区域外造成的“数据空缺”，这类缺失不携带天体自身信息；二是观测星等达到了极限。后者虽然在数据表中常被标记为缺失值，但它明确指出天体在该波段的亮度低于某个已知阈值，这本身就是一种重要的物理约束信息。一个经典的例子是用于发现高红移星系的“莱曼断裂”或“dropout”现象。例如一个红移 $z \sim 4$ 的星系，其光谱中低于 Lyα 线（静止系 121.6nm）的辐射被中性氢完全吸收，导致其在蓝光波段（如 u 和 g 波段）“消失”或无法被探测到，但在红光波段（如 r 和 i 波段）依然可见。这种在特定波段的“非探测”恰恰是识别高红移天体的关键特征。因此，有效利用非探测信息，而非简单丢弃或填充，对于构建鲁棒的测光红移模型至关重要。当数据缺失比例较低时，常见的方案是剔除不完整数据或应用各类数据填补技术[77]；而对于高缺失率的数据集，更可靠的解决方案是采用对数据缺失问题鲁棒性更强的算法，例如概率随机森林（Probabilistic Random Forest）[49]、生成式对抗插补网络（Generative Adversarial Imputation Networks，GAIN）[78] 等。

在具体的模型训练过程中，样本集的划分亦是一个关键环节。关于样本随机划分训练集、验证集和测试集的操作，目前既无明确的相对比例标准，也没有统一的抽取机制规范，可以采用随机抽样、间隔抽取等方式。虽然可以通过试错法寻找最优划分策略与抽样方法，但根据经验法则，当数据量充足（至少达 10^3 数量级）时，采用 60%、20% 和 20% 的比例进行随机抽取的数据划分方式是一种常规选择。

为突破单一方法的瓶颈，研究者开始探索融合不同方法的创新方案。传统的 SED 拟合方法理论上虽无红移上限，但在数据充足的情况下，其预测精度通常不及监督式机器学习方法。反之，机器学习方法虽精度更高，却受限于训练样本的参数空间。正是这种鲜明的互补性，激发了近年来大量结合二者优势的融合方法研究，旨在突破各自的固有局限。比较直接的模型如辅助分类测光红移估计（Classification-aided Photometric redshift(z) estimation，CPz）方法[79]，通过结合模板拟合法与 RF 分类器，实现对恒星、普通星系、AGN、类星体（Quasar，QSO）等天体的识别及红移估计。该方法首先利用 LePhare[15] 进行 SED 模板拟合，再通过三个 RF 分类器分别实现恒星识别、最优红移模板库匹配和离群值检测，最后通过概率阈值整合结果。在 SDSS 多波段数据[80] 上的实验表明，CPz 方法对正常星系和 AGN 的测光红移精度达 $\sigma = 0.039$，离群比例仅 2.3%，且无需 X 射线数据即可有效区分 QSO 与恒星，为开展高精度宇宙学研究提供了新手段。另一方案则采用分层贝叶斯融合策略，该方案通过整合不同模型的 PDF（包括模板拟合法和机器学习法得到的 PDF），显著提升了单模型性能评估的准确性[81, 82]。另一种主流的融合方向则聚焦于优化 SED 模板本身。例如利用扰动算法根据实测的测光数据动态调整 SED 模板的形状，再将其应用于模板拟合从而获得精度更高的红移预测[83]。此方法在 CSST 模拟星表上已验证了其有效性，成功使离群比从 3.86% 降低至 2.55%[84]；此外，还有方案利用机器学习从测光数据中直接构建大规模、更具代表性的 SED 模板库，并结合多波段数据进行离群值分析，这不仅优化了现有的结合了模

板拟合与机器学习的测光红移软件如 Delight[85]，还揭示了窄带数据可能引入离群值的新现象[35, 65]。

近期，一项名为 HAYATE（Hybrid Algorithm for WI(Y)de-range photo-z estimation with Artificial neural networks and TEmplate fitting）的混合算法进一步展示了该领域的潜力。该方法创新地采用人工神经网络和模板拟合，通过利用低红移星系的 SED 生成覆盖更广红移区间的模拟数据，并同时优化红移估计与概率分布，实现了精度和鲁棒性的双重提升。测试表明，HAYATE 在低红移区间的误差优于传统模板拟合方法（EAZY），在高红移区间表现相当，且计算速度快了约 100 倍。尽管该方法为 JWST 等大规模巡天项目提供了高效、高精度的红移估计方案，但研究也指出，对于活动星系核等罕见天体，其模板库仍有待扩展[86]。

3　模型性能评估方法与指标

鉴于测光红移估算方法的多样性日益增长，建立统一评估框架以实现方法间的公平比较十分必要，大型巡天项目更催生了系列标准化评测。目前已有诸多研究尝试通过制定统一数据协议与评估标准系统比较不同方法的优劣[32, 87]。测光红移精度测试（PHoto-z Accuracy Testing，PHAT）[88, 89] 开创了此类评测先河，后续 Euclid 与 LSST 项目相继推出同类竞赛[30, 32]。这些测试均采用盲测机制，即将部分数据集作为独立测试样本，测试者在模型构建阶段无法接触该数据，从而确保评估的客观公正性[90]。评估工作通常基于标准化统计指标，包括标准差、偏差度、归一化中位绝对偏差、均方根误差及离群值比例等，开展从点估计角度出发衡量测光红移的估算质量，是学界公认为具有充分可信度的评估体系。

由于仅依赖点估计值表征测光红移质量存在系统性缺陷而可能导致显著偏差，当前学界逐步转向采用 PDF 构建更全面的置信区间评估体系，通过量化红移估计的不确定性分布来有效提升高精度应用场景的可靠性。从信息量的角度看，PDF 由于能提供点估计无法捕捉的丰富信息而成为更优越的评估载体，尤其是当参数空间存在简并性时，PDF 能反映出因简并导致的次级解。在实际应用中，PDF 不仅可以提升宇宙学和弱引力透镜测量的精度，还能充分分析从弱透镜层析（weak lensing tomography）到重子声学振荡（Baryon Acoustic Oscillations，BAO）等各种宇宙学参数不确定性[91]。目前如 KiDS、Euclid、LSST 等巡天项目已全面升级数据产品标准，其星表不仅包含点估计值，更提供了完整的 PDF 统计特征[30]。

尽管学界就采用 PDF 已达成共识，但在如何量化评估 PDF 本身的质量上却远未形成统一标准。这种分歧在 Euclid 与 LSST 两大巡天项目中体现得尤为显著：Euclid 项目采用的做法是首先根据源的点估计值，将源分配到不同的测光红移区间中，得到 PDF($z - z_{\rm spec}$)，最后将堆叠的 PDF 峰值附近 $\pm 0.05(1 + z_{\rm spec})$（记为 $f_{0.05}$）和 $\pm 0.15(1 + z_{\rm spec})$（记为 $f_{0.15}$）区间内 PDF 面积占总 PDF 面积之比作为优化参数[30]。而 Amaro 等人[92] 指出此类指标存在显著局限性，如设计的 PDF 为点估计附近的单峰分布则预测结果可以轻易达到优化标准进而导致评估结果失真。对基于 KiDS DR3 数据集（$z_{\rm spec} < 1$）的测试表明，当其采用虚

拟的 PDF 方案时，其 $f_{0.05}$ 和 $f_{0.15}$ 指标分别可达 93.1% 与 99.0%；而经验证成熟的算法（METAPHOR、ANNz2、BPZ）所得对应指标分别为 65.6%、76.9%、46.9%（$f_{0.05}$）和 91.0%、97.7%、92.6%（$f_{0.15}$）。这一发现表明将 $f_{0.05}$ 和 $f_{0.15}$ 作为测光红移 PDF 综合评估指标存在显著局限性。

LSST 则采用与 PDF 的累积分布函数（Cumulative Distribution Function，CDF）有关的常用指标，包括概率积分变换（Probability Integral Transform，PIT）

$$\mathrm{PIT}(\mathbf{x}_{\mathrm{val}}, \mathrm{y}_{\mathrm{val}}) = \int_{-\infty}^{y_{val}} \widehat{p}(\mathrm{y}|\mathbf{x}_{\mathrm{val}})\, dy, \tag{1}$$

和最高概率密度（Highest Probability Density，HPD）

$$\mathrm{HPD}(\mathbf{x}_{\mathrm{val}}, \mathbf{y}_{\mathrm{val}}) = \int_{\mathbf{y}:\widehat{p}(\mathbf{y}|\mathbf{x}_{\mathbf{val}}) \geq \widehat{p}(\mathbf{y}_{\mathbf{val}}|\mathbf{x}_{\mathbf{val}})} \widehat{p}(\mathbf{y}|\mathbf{x}_{\mathbf{val}})\, dy, \tag{2}$$

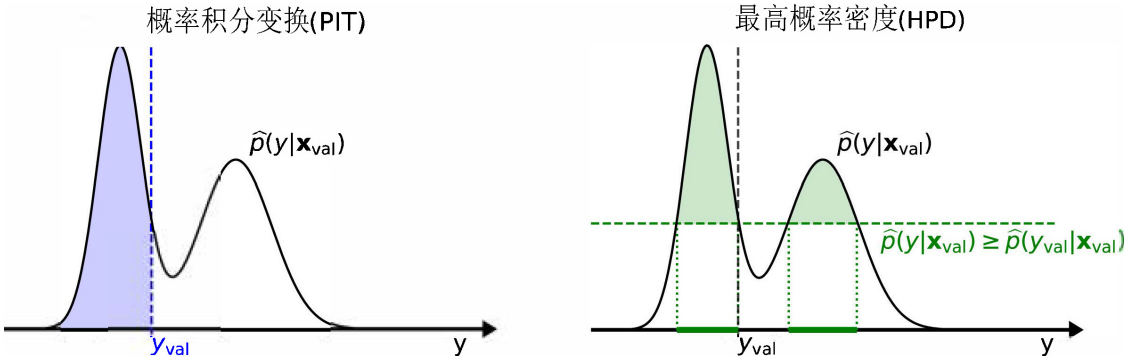
其中 $\mathbf{x}$ 为测光数据，$\hat{p}$ 表示 PDF，y 为红移，y_{val} 为用来验证的真实红移，两者计算的示意图如图1所示。PIT 与 HPD 指标与贝叶斯诊断的 p 值和 e 值具有直接关联。需要注意的是，虽然每个源的 PDF 都有 PIT/HPD 值，但模型的性能是通过由大量样本构成的 PIT/HPD 分布情况来讨论，一般会通过对整体的 PIT 和 HPD 分布的分位数-分位数（Quantile-Quantile，QQ）图可视化分析进行。Harrison 等人[93] 的研究表明，即使在多变量的情形下，若模型总体水平上校准良好，HPD 将和 PIT 一样服从均匀分布 $\sim U(0,1)$，QQ 图会将统计量（PIT/HPD）的分布与假设下的均匀分布进行对比，理想的 QQ 图中所有点应紧贴对角线，表明实际分布与理论分布完全一致，若 PIT/HPD 的概率分布在两端则表示模型预测过于自信，在中间区域则信心不足。还有其他一些指标也常被用来定量地确定 PIT/HPD 分布的均匀性，如相对熵（Kullback-Leibler divergence，KL），KS 检验（Kolmogorov-Smirnov，KS test），CvM 检验（Cramér-von Mises，CvM test）和 AD 统计量（Anderson–Darling，AD statistic）等。

虽然 PIT 和 HPD 指标应用广泛，但后续研究指出它们并非完美，在特定情况下仍可能给出误导性结果。Schmidt 等人[32] 指出，即使 PDF 未被准确估计，其值仍可能呈现均匀分布特征；在缺乏真实测光红移后验分布的情况下，目前仅存在条件密度估计损失（Conditional Density Estimation，CDE loss）可用于评估测光红移 PDF 性能，其定义为

$$\mathcal{L}(f, \hat{f}) = \iint \left[\hat{f}(z|\mathbf{X}) - f(z|\mathbf{X})\right]^2 dz dP(\mathbf{X}), \tag{3}$$

式中 $f(z|\mathbf{X})$ 代表未知的真实红移 PDF，$\hat{f}(z|\mathbf{X})$ 是基于测光数据 $\mathbf{X}$ 由模型给出的 PDF。该损失函数可视为标准回归中均方误差（Mean-Square Error，MSE）的对应指标。由于 CDE 损失依赖于真实的 PDF，其值虽无法直接计算，但可通过以下方式估计该损失

$$\hat{\mathcal{L}} = \mathbb{E}_X \left[\int \hat{f}(z|\mathbf{X})^2 dz\right] - 2\mathbb{E}_{x,z}\left[\hat{f}(z|\mathbf{X})\right] + K_f, \tag{4}$$



注：实线为假设的红移 PDF，y_{val} 表示真实的红移值。图中的阴影部分面积即表示对应的 PIT 与 HPD 值。

图 1　PIT（左）与 HPD（右）的构建原理[87]

式中第一项表示测光红移后验相对于测光数据 $\mathbf{X}$ 边缘分布的期望值，第二项表示相对于 $\mathbf{X}$ 与所有可能红移空间 z 联合分布的期望值，第三项 K_f 为仅取决于真实 PDF 的常数。CDE 损失函数最显著的优势在于即使在不掌握真实 PDF 分布的情况下，仍能估计出相对真实的损失函数，至多相差一个常数项（具体推导详见 2017 年 Izbicki 等人[94] 公式 7 及相关讨论）。这一特性使得在当前缺乏真实 PDF 数据的情况下，仍能对现有数据集的估计方法进行定量比较。

综上所述，无论是传统的点估计指标还是新兴的 PDF 评估函数，每种方法都有其适用场景与局限性，表1总结了目前主要的点估计及 PDF 优化的统计量。尽管着眼于优化其中某一项指标会导致生成无意义的 PDF，但综合考虑这些指标还是会得到相对可靠的判断标准，选择恰当的优化方法及度量 PDF 可靠性的指标目前仍属于悬而未决的开放研究课题。

需要说明的是，机器学习模型在实际应用中往往面临复杂多变的场景，这种表现的稳定性与可靠性仍需进一步探讨，如并非所有的测光数据都能适用于机器学习。下节将聚焦于描述模型在不同情境下可能产生的不确定性以了解模型的特性与应用的潜在影响。

4　机器学习估算测光红移的不确定性

现代星系宇宙学研究对降低各种误差要求极高。例如，在宇宙尺度的层析成像研究中，为控制暗能量估计中的噪声影响，需要保证每个红移区间内的偏差和弥散度在 ~ 0.003 或更低[95]。测光红移误差会模糊大尺度结构，影响聚类算法的测量精度，但该效应主要局限于较小尺度。在大面积测光巡天情况下，一定的误差也能提供有价值的结构信息，这对宇宙学研究意义重大[96]。不同类型的星系在测光红移方面表现各异，例如基于 SDSS 数据的机器学习研究表明，在亮红星系样本上获得的红移弥散度（$\sigma \sim 0.028$）相比蓝星系下降了约一半（$\sigma \sim 0.05$）[97]；对于红移 $z > 0.4$ 的星系，红移使星系的关键光谱特征和辐射峰值移至近红外波段，所以足够深的近红外图像尤为重要[98]。通过融合光学与红外波段数据，可显著提升

表 1 用于测光红移点估计及 PDF 优化的统计量[50]

平均值点估计	$\int_0^{z_{\max}} z \times p(z\|x)dz$
峰值点估计	$\{z\|max(p(z\|x))\}$
累积分布函数 $C(z\|x)$	$\int_0^{z'} p(z\|x)dz$
相对残差	$(z_{\rm phot} - z_{\rm spec})/(1 + z_{\rm spec})$
平均绝对离差（$\sigma_{\rm nMAD}$）	$1.4826 \times {\rm median}\left(\left\|\frac{\Delta z - {\rm median}(\Delta z)}{1+z_{\rm spec}}\right\|\right)$
概率积分变换[32]	$\int_0^{z_{\rm spec}} p(z\|x)dz$
Odds[68]	$\int_{z_{\rm phot}-\S_z}^{z_{\rm phot}+\S_z} p(z\|x)dz$
最高概率密度[87]	$\int_{\mathbf{y}:p(z\|x)\geq p(z_{\rm spec}\|x)} p(z\|x)dz$
一阶最佳传输距离	$\int_0^{z_{\rm spec}} C(z\|x)dz + \int_{z_{\rm spec}}^{z_{\max}} \|1 - C(z\|x)\|dz$
连续分级概率评分	$\int_0^{z_{\rm spec}} C(z\|x)^2dz + \int_{z_{\rm spec}}^{z_{\max}} (1 - C(z\|x))^2dz$
交叉熵	$-\log(p(z_{\rm spec}\|x))$
熵	$-\int_0^{z_{\max}} p(z\|x)\log(p(z\|x))dz$
标准差（σ）	$\sqrt{\int_0^{z_{\max}} (z - z_{\rm phot})^2 p(z\|x)dz}$
偏度	$\int_0^{z_{\max}} (z - z_{\rm phot})^3 p(z\|x)dz/\sigma^3$
峰度	$\int_0^{z_{\max}} (z - z_{\rm phot})^4 p(z\|x)dz/\sigma^4$
条件密度损失（$\hat{\mathcal{L}}$）[32]	$\mathbb{E}_x\left[\int p(z\|x)^2dz\right] - 2\mathbb{E}_{x,z}\left[p(z\|x)\right] + K_f$
$f_{0.05/0.15}$[30]	$\int_{z_{\rm peak}-0.05/0.15}^{z_{\rm peak}+0.05/0.15} p((z - z_{\rm spec})/(1 + z_{\rm spec})\|x)dz$

星系样本的测光红移精度。而要确保测光红移精度满足宇宙学测量要求，则需达到一定光学深度（$r \sim 24$）[96]。以下将简要例举影响测光红移估算的主要原因。

4.1 数据层面的不确定性

影响测光红移估算精度的最直接因素来源于输入数据本身，这类不确定性主要包括输入数据的误差、范围和数量，以及标签的可靠性。机器学习模型在很大程度上依赖于高质量的输入数据进行训练。如果输入测光数据本身存在较大测量误差，或者覆盖的波段范围不足，模型就难以捕捉到精确的星系特性。同样，训练样本的数量不足或其红移标签的准确性不高，都会直接限制模型的学习能力和泛化性能，导致最终估算的红移不确定性增加。

4.1.1 测光误差

测光红移的精度高度依赖于输入测光数据的质量，其中测光误差是关键因素。通常在模拟测光星表时，会假设测光误差服从高斯分布，然而实际测光误差分布呈现显著非高斯特性，这种“重尾分布”的成因是多方面的，除了理想化的热噪声和读出噪声外，还包括一系列系统效应，如未被完美扣除的天光背景起伏、宇宙线冲击、探测器坏点或饱和、以及在拥挤天区由天体重叠（blending）引起的测光污染。特别是在信噪比接近极限的暗弱天体中，这种非高斯性尤为突出。当样本中约 10% 的源处于尾部时，其对预测弥散的影响增加超过 2σ 以

上，甚至超过信噪比减半的影响[99]，这意味着需要尽量减少星等和颜色的尾部误差，尤其是对噪声较为敏感的 U 波段，因为该波段不仅仪器和大气透过率较低，而且是识别低红移星系巴尔末断裂或中等红移星系莱曼断裂的关键，不准确的 U 波段测光极易导致大比例的离群值。

在机器学习应用方面，处理这种非高斯误差常规的做法是将星表提供的测光误差作为输入特征的一部分与流量、颜色等信息一同送入模型进行训练[76]，这种方法的内在逻辑是期望模型能自动学习到噪声的复杂模式。然而对于标准的前馈网络或随机森林等模型，它们可能仅将误差视为另一个独立的数值特征而难以真正理解其作为“置信度”或“概率分布宽度”的物理内涵，尤其是在高维特征空间中误差特征所携带的信息可能被稀释。为了更深刻地利用误差信息，一种有效的方法是基于误差的权重化数据增强。该方案并非直接使用误差特征，而是将其作为标准差从高斯分布中抽取扰动量来生成一个“增强”后的模拟星表。模型可在这个经过合理扰动的“无限”数据集上进行训练从而学会在不同信噪比下的鲁棒性。在对 CSST 模拟星表的测试中该方法相比于直接输入误差特征的模型成功将红移离群比从 2.3% 降低至 2.0%[21]。这种方法本质上是让模型学习了在给定误差范围内数据可能呈现的各种形态从而迫使其学习星等与红移间更为本质的物理关联。更高级的概率化模型如贝叶斯神经网络和混合密度网络能够在其数学框架内自然地融合输入数据的不确定性[68]，该模型输入的一个特征对应一个概率分布而不是一个值，支持学习一对多的映射。这些模型不仅预测红移值还能直接从带有误差的输入中推导出预测结果的概率分布从而提供更完整的的不确定性量化，这对于描述因颜色简并性而可能存在多个红移解的情况至关重要。机器学习方法虽能自动学习噪声特征，但需要合适的参数空间和一致的数据集，通过在模型训练阶段有效控制和建模误差分布可以显著降低模型对完备且庞大的光谱训练样本的依赖[100]。在实践中为所有类型的、覆盖全部参数空间的暗弱天体获取光谱证认是不现实的，因此一个能够从带有真实噪声的测光数据中稳健学习的模型将是最大化这些巡天科学产出的关键。图2所示为 Blake 和 Bridle[96] 的研究，给出了在 $10,000\,\mathrm{deg}^2$ 巡天范围内测光红移的置信边界随误差阈值变化规律，为深度巡天设计提供了参考。

4.1.2　观测的波段数量及波段范围

测光红移的精度和可靠性与输入测光数据的波段数量和波长覆盖范围密切相关，通过在更宽广的波长范围内对天体的 SED 进行采样可以更精确地约束其 SED 形状，从而打破不同红移和天体类型之间的“简并性”。一个普遍被验证的结论是，更广的波段覆盖范围通常能带来更优的红移估计结果，这背后的物理原理因天体类型而异。对于正常星系而言，关键在于精确锁定光谱断裂特征。星系光谱中最重要的特征之一是 4000Å 断裂，它是由恒星大气吸收线密集叠加形成的。在静止坐标系下这个断裂位于光学蓝光部分，随着红移增加该特征会逐渐移向红端乃至近红外波段。傅莉萍等人[101] 的研究表明，近红外波段（Y,J,H,K_s）对测光红移预测准确性有积极贡献。他们对比了仅使用四个光学波段和添加近红外波段后的研究结果，发现添加近红外波段后能更准确地预测高红移星系红移且误判率降低 15%，显著提升了宇宙学参数 σ_8 与 Ω_M 的约束精度。这表明增加宽波段覆盖能有效缓解参数空间简并问题，尤其在红移 $z > 0.4$ 时，近红外数据对精度改善至关重要[98]。

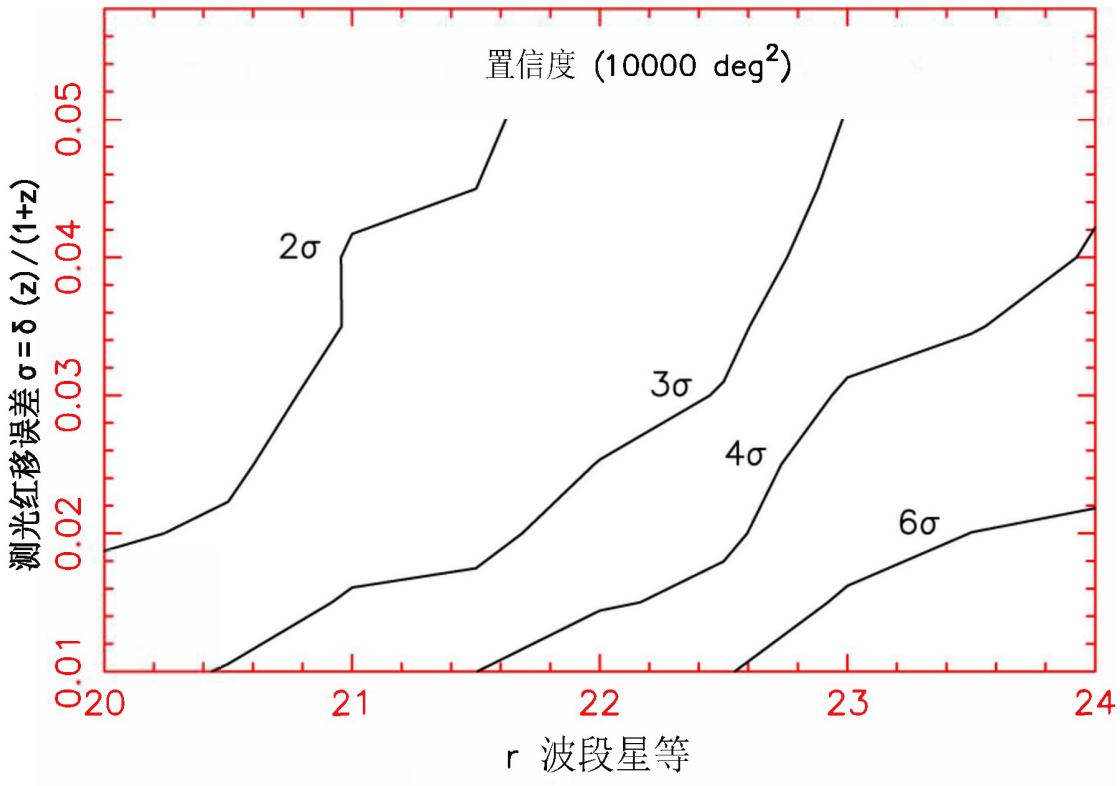


图 2 测光红移的置信度随误差及 r 波段星等的变化规律[96]

对于具有特殊 SED 的天体，宽波段覆盖更是不可或缺。以活动星系核（Active Galactic Nucleus，AGN）和类星体为例，它们的辐射并非来自单一的恒星族群，而是由中心超大质量黑洞的吸积盘（在紫外和光波段辐射）、以及周围的尘埃环（在红外波段再辐射）等多个部分组成的复杂连续谱。在 eROSITA（extended ROentgen Survey with an Imaging Telescop Array）巡天中[102]，Brescia 等人[39] 研究了 X 射线源对应体的测光对 AGN 源测光红移质量的贡献，结果表明，增加波段数量可改善测光红移估计效果。例如，仅使用光学波段会导致低红移源的红移值高估，添加 WISE 的中红外波段可减少这种影响，进一步添加 IRAC 波段能提高统计精度，降低异常值率。在 eROSITA 的观测深度下，机器学习模型和 SED 拟合方法在离群值百分比方面表现相近[39]。

需要注意的是，作为输入特征的波段并非越多越好，Carrasco 等人[59] 发现当使用 RF 去除最不重要的几个输入特征后预测精度反而提高了，这是因为不重要的波段特征可能包含大量噪声或测量误差。例如，某些波段的观测可能受仪器限制或环境因素（如大气吸收）影响较大，导致数据信噪比低。模型若试图从这些噪声中学习，反而会引入误差，尤其在数据量有限时，模型可能过度拟合噪声而非真实信号，影响泛化能力。另一方面高维空间中数据稀疏性增加，模型需要更多样本才能有效学习。当去除无关特征后，数据分布更紧凑，模型

在低维空间中更容易捕捉到红移与有效波段间的真实关系，提升计算效率和准确性。且如前文所述，不同波段间可能存在高度相关性，如相邻波段的测光数据相似，导致模型参数估计不稳定。去除冗余特征可减少共线性，增强模型鲁棒性，使权重分配更合理。

大规模巡天项目本身需优先选择能最小化参数简并的测光波段组合，并扩展多波长覆盖范围来实现科学目标。为了在“数量”和“范围”之间取得平衡，机器学习方法本身也提供了评估波段重要性的工具。以 RF 模型袋外（Out-Of-Bag，OOB）采样技术为例，在构建决策树时随机抽取未参与训练的数据样本可验证测光特征的信息贡献，识别并去除信息熵冗余或具有误导性的特征。需要注意的是，机器学习方法给出的重要性程度受模型本身及算法决定，且数据之间的相关性可能会误导判断结果，使重要性程度出现冗余信息。Euclid 团队[103]发现当复制了 VIS$-u$ 特征时，RF 给出的特征重要性程度发生了改变。陆君豪等人[21] 使用不同特征重要性计算方法，发现在模拟的 CSST 星表中，加入误差权重前后特征的重要性会出现变化，随后在对输入数据进行斯皮尔曼秩相关系数检验分类后，发现 z 波段和 y 波段有着极强的相关性。

4.1.3　光谱红移标签的可靠性

尽管目前已有众多测光红移估计方法，但至今没有一种能达到光谱红移测量的精度（$\sim 10^{-3}$）。如基于宽带测光获得的测光红移，其最佳精度的误差约为 $\sigma \sim 0.02$，与光谱红移相比精度仍有较大差距[104]。

通过监督学习获得的测光红移，很大程度上依赖于用作基准的光谱星表的完整性和质量。光谱样本的不完整性通常在测光暗弱部分更为突出，这可能引发选择效应而影响整个参数空间。同时光谱红移的残余误差也会影响机器学习模型训练验证指标的可靠性，进而直接影响测光红移的质量。一般光谱红移的可靠性在 95% $\sim$ 99% 之间，这意味着 1% $\sim$ 5% 的训练样本不可靠[35]。目前无法有效区分样本中光谱和测光不确定性的影响。大型巡天项目获得的星表数据量巨大，基于目视检查的手动分析方法难以实施，因此需要探索自动检索机器学习算法来区分数据中的不确定性来源。

Razim 等人[35] 提出了一种识别不可靠光谱红移样本的方法，利用深度成像多目标光谱仪（Deep Imaging Multi-Object Spectrograph，DEIMOS）星表以及 COSMOS2015 星表来改进测光红移估计：结合 SOM 与 MLPQNA 神经网络，通过引入光谱质量系数 $K_{\rm spec}$，可有效剔除 COSMOS 数据中不可靠的光谱红移样本，使测光红移预测的弥散度降低约一半；除了 $K_{\rm spec}$，该研究还使用 DEIMOS 光谱红移作为验证集，利用所谓的星系占据分布映射（galaxy occupation map）概念，使测光红移样本与验证集样本在 SOM 聚类网格中占据相同区域，确保测光数据与光谱数据之间的准确对应关系，则离群比将从约 11% 显著降低至约 2%。由此可见不可靠的光谱红移会让机器学习引入难以处理的偏差。未来大型巡天项目需兼顾光谱训练样本的数量与质量，并发展自动化验证技术，以充分释放测光红移在星系宇宙学研究中的潜力。

4.1.4 测光图像的影响

测光红移估计面临两大挑战是高红移时精度下降和测光颜色与光谱红移的简并性，即不同红移源可能具有相似的颜色特征，导致仅依赖星等和颜色的监督学习方法难以区分。传统测光数据如星等、颜色等仅利用了图像信息的一小部分，这主要是由于受点扩散函数、孔径选择等因素影响，无法充分捕捉星系的形态特征，这些形态特征与星系的恒星形成历史、恒星质量等物理参数密切相关，进而影响 SED 和测光红移估算。此外图像质量，包括信噪比、像素分辨率和背景噪声等决定了模型能否从图像中准确提取出有效的形态和颜色信息。在低信噪比的图像中星系的微弱特征可能被噪声淹没，导致模型无法有效区分不同类型的星系或估算其准确的红移，因此在利用机器学习方法进行测光红移估算时，对测光图像的质量和星系形态特征的深入分析同样至关重要。这里需要注意表面亮度涨落（Surface Brightness Fluctuations, SBF）[105] 测距法这一相关但不同的概念。SBF 是一种利用高信噪比图像测量邻近（$\sim$ 100Mpc）早型星系距离的经典方法。其原理是星系距离越远在探测器上成像就越“平滑”，单位像素内包含的恒星数量涨落就越小。SBF 方法只适用于年老的、气体和尘埃含量少的早型星系如椭圆星系、S0 星系以及旋涡星系的核球，因为它依赖于一个平滑、稳定的红巨星族群，对于充满年轻亮星、星团和尘埃带的旋涡星系旋臂该方法会变得不可用。尽管 SBF 也利用图像信息，但它是一种独立的、高精度的距离标定技术，属于宇宙距离阶梯的一部分，其物理原理和应用场景与本文讨论的、适用于大规模、遥远星系样本的测光红移估计方法完全不同，测光红移本质上是在估计 SED 的形状，而 SBF 提供的是一个直接的、物理的距离估计。在测光红移应用中，SBF 可作为一种用于校准和约束测光红移的强距离先验信息。对于更先进的模型如 CNN，SBF 信息可以被间接利用，CNN 能够自动识别图像的纹理信息，模型可以学会在看到一个“粗糙”的椭圆星系时，将其与较近的距离关联起来；看到一个“平滑”的椭圆星系时，与较远的距离关联。且多分支的神经网络可以同时处理多波段测光数据和高分辨率图像切片，提取其中的多层次特征，例如颜色梯度、盘状倾角、特殊形态等，避免了手动特征选择的偏差，最后将两者的信息融合，做出最终的红移判断。D'Isanto 和 Polsterer[100] 通过 CNN 与 MDN 结合，从 SDSS 图像中同时提取颜色和形态信息，使类星体测光红移的预测偏差控制在 0.004，标准差降至 0.069；Pasquet 等人[106] 使用 CNN 分析 SDSS 五个波段的图像，其预测偏差在 1σ 范围内仅为 ±0.02，显著低于 kNN 方法的 ±0.05，且未发现与红移直接相关的系统偏差；周兴晨等人[69] 使用 Hybrid transfer 算法将 CSST 模拟图像加入训练后，离群比例相比于只使用流量从 1.43% 降至了 0.9%。

尽管深度学习在图像特征挖掘上表现出潜力，但其在测光红移中的应用仍需验证。Pasquet 等人[106] 发现 CNN 预测对训练样本中密集红移区间存在 $\leq 5\%$ 的轻微偏差，需进一步结合测光数据进行交叉验证。未来大规模巡天项目有望通过深度学习直接从像素级图像中提取信息，避免传统测光参数选择的局限性，为解决高红移和简并问题提供更全面的解决方案。

4.2 训练样本与方法层面的不确定性

在解决了数据质量问题之后，如何利用这些数据进行有效训练，则引入了另一层面的不确定性。这主要体现在样本的代表性和对数据的处理方式。训练样本必须能充分代表整个巡天数据的分布，否则模型可能在未见过的红移或类型区间表现不佳。此外对数据的预处理方

法，如特征工程、归一化或异常值处理等，也直接影响模型的性能，不恰当的数据处理可能会丢失重要的物理信息，或引入新的偏差，从而加剧测光红移估算的不确定性。

4.2.1 训练集不充分及异常值

在测光红移预测中，识别和分析训练集中的异常值极为关键，因为这些异常值可能导致距离估计错误。它不仅能为后续光谱观测提供有效指导，优化训练数据集，补充欠采样区域数据，还能通过对红移质量进行统计预测，评估不同测光特征组合的有效性[107]。如监督学习的 RF 模型[59] 和非监督学习的 SOM[108] 均能实现参数空间优化，后者通过构建 Kohonen 映射实现无光谱先验的异常区域识别。

该问题在包含 AGN 的星系红移预测上显得尤为明显，因为 AGN 核区辐射占总体辐射的比例未知且因源而异。由于 X 射线选源在 $z < 1$ 时可靠性低，目前仅塞弗特星系（Seyfert galaxy）在有窄/中滤光片测光数据的情况下（如 COSMOS 巡天），其测光红移质量才可与正常星系相媲美[109]。混合方法（机器学习 +SED 拟合）虽取得进展[79]，但在进行 SED 拟合时，难以选择合适的模型[110]，其测光红移估计效率仍不及正常星系。同样，基于监督学习的机器学习模型受限于是否有足够大且完整的光谱训练样本，而大多数光谱样本通常从光学波段选源的星表中提取，这不可避免地导致 AGN 或其他特殊天体在样本中分布不均衡，这种偏差对未来射电巡天如平方公里阵列（Square Kilometre Array，SKA）的影响可参见 Norris 等人[111] 的研究。

为了应对 AGN 光谱训练样本稀缺且存在偏差的挑战，近期的研究开始利用新一代深度成像巡天，并结合更精细的物理特征提取方法来构建大规模、高质量的 AGN 专属训练集。该解决思路在于与其在以正常星系为主的样本中艰难处理 AGN 异常值，不如主动为 AGN 这类“异常值”建立专门的、具有代表性的训练基准。在最近的研究中，Saxena 等人[112] 利用暗能量光谱仪（Dark Energy Spectroscopic Instrument，DESI）成像遗珍巡天的第十次数据发布（Imaging Legacy Survey，LS10）来估算 X 射线检测到 AGN 的测光红移。该研究以 14,000 个 X 射线检测到的 AGN 作为训练样本，这种选源方式从根本上绕开了光学样本对 AGN 的偏差，减缓了此前 AGN 光谱样本规模不足的问题。针对 AGN 核区与宿主星系辐射贡献未知这一难题，该研究并未使用简单的单孔径星等，而是运用了精细的多孔径测光技术。这种方法旨在通过分析天体在不同孔径下的亮度变化来更好地区分和量化明亮核心与延展的宿主星系的光，为机器学习模型提供了更具物理意义的输入特征。基于上述高质量的训练集和特征，他们运用全连接神经网络算法 CIRCLEZ 进行训练。在对 2913 个 AGN 构成的独立测试集进行检验时，模型的预测弥散度达到了 0.067，离群值比例控制在 11.6%。这一结果甚至优于以往的研究，证明了通过构建专属、无偏的训练集并辅以精细化物量输入，完全可以实现对 AGN 这类特殊天体的高效红移估计。这项工作及其发布的 eROSITA/eFEDS 天区内 AGN 的更新版测光红移星表[113]，共同为解决 AGN 红移估计这一长期存在的难题提供了重要的方法。

4.2.2 输入特征的处理

对于有可靠的红移及测光数据的完备样本，不同输入特征处理方式会对结果产生影响。这种处理方式一般被称为特征工程（Feature Engineering），即将原始数据转换为更能代表问题本质的特征的过程，是连接原始数据与机器学习算法的桥梁。

基于已有数据构建新的特征也是提高预测精度的一种常用方式。在测光红移估算中，基础的特征工程就是利用不同波段的流量。虽然流量本身是直接的观测量，但由它们构建的“颜色”，即两个波段的星等差往往是更具物理意义的特征，这是因为颜色直接反映了天体 SED 的斜率，且 SED 的形状与天体的红移、年龄、金属丰度、尘埃含量等关键物理属性紧密相关。不同波段对红移预测的贡献，本质上取决于它们能否有效捕捉到随红移移动的关键光谱特征，下面将逐个分析常用测光波段在这一物理框架下的具体作用和贡献。

1. **紫外及蓝光波段（UV($\sim$ 200nm), u($\sim$ 350nm), g($\sim$ 480nm)）**：在低红移宇宙（$z < 0.5$）这几个波段的主要任务是精确定位 4000Å 断裂。研究表明，如果巡天项目中缺少 u 波段数据，其对 $z < 0.5$ 星系的红移估算精度会显著恶化，甚至无法满足 LSST 等项目对宇宙学科研的精度要求[114]。在高红移宇宙（$z > 2.5$），u 和 g 波段的作用则转变为利用莱曼断裂进行目标筛选。静止系波长为 1216Å 的 Lyα 线以及 912Å 的莱曼极限在星系际介质的中性氢吸收下会形成一个剧烈的光谱跌落。因此 UV 和蓝光波段对于准确识别高红移星系至关重要，缺少它们会导致高红移天体与低红移的尘埃红化星系产生混淆。
2. **光学中段波段（r($\sim$ 620nm), i($\sim$ 750nm), z($\sim$ 900nm)）**：这三个光学红端波段主要用于研究中等红移宇宙（$0.4 < z < 1.6$），接力 u 和 g 波段，继续追踪随着红移增加而不断向长波方向移动的 4000Å 断裂。
3. **近红外（near-Infrared, NIR）波段（Y($\sim$ 1μm), J($\sim$ 1.25μm), H($\sim$ 1.65μm), K($\sim$ 2.2μm)）**：当 4000Å 断裂这一关键特征被红移到 1μm 以外，超出了传统光学 CCD 的探测范围时，NIR 波段接管了对其进行探测的任务。NIR 波段对于打破年龄-尘埃-红移简并性至关重要，NIR 波段的观测可以有效地区分这两种情况，年老星系在 4000Å 断裂处表现为一个尖锐的跌落，而尘埃遮蔽星系的 SED 则通常是平滑的红化连续谱。除此以外，NIR 波段主要探测的是年老恒星族群如红巨星的辐射，因此它与星系的总恒星质量有很好的相关性，有助于更精确地确定恒星族群的性质。
4. **中红外波段（mid-Infrared, MIR）（W1($\sim$ 3.4μm), W2($\sim$ 4.6μm)）**：该波段主要探测的不是恒星光球层的直接辐射，而是由尘埃吸收恒星光后再辐射出的热辐射。例如，AGN 的中心被一个尘埃环包围，其热辐射在 MIR 波段形成独特的“红尾”，这成为将 AGN 与正常星系区分开来的决定性特征。若无 MIR 数据，许多 AGN 的红移会被严重误判[115]。

在测光红移估算中，早期研究多采用试错法或机器学习的组合搜索[100]，计算成本高且未必能获得最优解。近年来随机森林因其能有效评估特征重要性而被广泛应用，如 Brescia 等人[39] 开发的 ϕLAB（Parameter handling investigation LABoratory，PhiLAB）方法结合

随机森林和 L1 正则化，通过引入“影子特征（shadow features）”量化特征贡献，在多波段测光数据中识别出关键特征。他们发现仅使用星等作为输入特征时，K 波段因对应星系 SED 的拐点而最为重要；而在加入颜色信息后，颜色特征的总重要性占比超 60%[39]。在对 CSST 模拟星表利用 RF 判断特征重要性程度时发现，虽然 i 和 z 波段的流量的重要性程度极低（分别为 0.006 和 0.009），但两者构成的颜色 $i-z$ 重要性程度（0.398）却急剧增加，因为它直接对应了红移在 $z\sim1$ 附近的 4000Å 断裂位置[21]。因此在预处理阶段去除某些看似无用的特征时需十分谨慎。

本质上颜色不提供除了流量中已经包含的额外信息，传统的机器学习模型需要将颜色作为输入参数，表明模型结构无法捕获输入测光数据之间的相关性。罗智坚等人[116] 最近提出了一种基于 LSTM 的 RNN 新模型用于估计测光红移。该模型避免了大多数传统机器学习模型通常面临的输入特征选择问题，只需要按波长顺序组织每个波段的流量形成输入序列就可以自动有效地学习它们之间的依赖关系，捕获流量序列中的模式和相关性。由于 LSTM 模型固有的序列处理能力，该模型可以有效地捕获序列中各个数据点之间的相关性，从而减少对特征工程的需求，即该模型不再需要手动组合输入特征来生成颜色，也不需要复杂的特征选择工程，避免了人为误差或主观偏见。

4.3　巡天与红移区间层面的不确定性

除了数据和训练过程，红移区间和巡天策略的特性也对机器学习模型的性能构成挑战。在特定红移区间内，即使是优秀的模型也可能面临固有的困难。不同巡天项目采用的观测策略则决定了所能获得的测光数据质量。

4.3.1　不同红移区间下的影响

测光红移的最终精度和可靠性是多个因素综合作用的结果。除了前述的误差处理、波段选择和特征工程外，机器学习方法在不同的红移区间会有不一样的预测效果。低红移区间 ($z<0.5$) 是测光红移估算最成熟、最精确的区间。在此范围内天体通常更明亮，4000Å 断裂位于观测条件良好的光学波段，使得红移信号非常清晰。得益于 SDSS 等大规模光谱巡天，该区间的训练样本不仅数量庞大，而且具有极高的完备性和代表性。因此，无论是模板拟合法还是机器学习法，通常都能达到非常高的精度。主要的挑战在于区分光谱特征不明显的低红移星系与银河系内的恒星，以及处理少量因尘埃或 AGN 活动导致的异常值。如最近在 SDSS 数据上使用的混合输入 CNN 模型，在 $z<0.3$ 的范围内实现了极低的均方根误差，仅为 0.0007[117]。

中等红移区域（$0.5<z\lesssim1.5$）对于弱引力透镜和重子声学振荡等宇宙学研究至关重要。虽然各种测光红移方法的预测性能依然良好，但与低红移区域相比开始下降，离散度和离群率开始增加[118]。在此红移区间 4000Å 断裂逐渐移入红端光学波段（i,z）乃至近红外 Y 波段。此时红移精度高度依赖于这些红端波段的测光质量。颜色空间的简并性问题开始变得突出，不同类型或不同红移的星系可能呈现相似的颜色，导致模型误判和离群值比例上升。在这个红移段，光谱巡天的完备性逐渐下降，它们通常偏向于观测更亮或特定类型的亮红星系，这意味着训练集对于测光巡天所观测到的完整星系群体的代表性降低[119]。

高红移区间（$z > 1.5$）是研究早期宇宙和星系形成的关键窗口，也是大多数标准的、纯数据驱动的机器学习模型性能急剧下降的地方，离散度和离群率显著增加。因为在该红移范围可用的光谱样本稀疏，且严重偏向于最明亮、最稀有的天体如类星体，这些天体并不能代表典型的星系群体。这迫使机器学习模型在其训练数据范围之外进行远距离外推，而外推能力差正是机器学习方法众所周知的弱点。例如 ANNZ+ 算法在红移 $z \approx 1.3$ 以上就难以维持其准确率[46]，此时主要目标变成了识别高红移候选体，而非精确估计其红移值[56]；其次高红移天体内禀光度很低，导致测光测量的信噪比低，这种噪声进一步模糊了本已简并的颜色信息，增加了不确定性，尤其是在被称为“红移沙漠”的特殊红移区间内（$1.4 < z < 2.5$），莱曼断裂尚未进入光学波段，而 4000Å 断裂已完全移入近红外波段。对于缺乏紫外（u）和近红外（J, H, K）覆盖的巡天而言，星系在光学波段几乎没有强的光谱特征，导致红移解的不确定性极大，是灾难性离群值的高发区。

机器学习在高红移区域预测性能下降并非是某个算法“不好”的标志，而是监督学习模型在被迫外推到一个与训练数据（低红移、明亮、特定类型的星系）在根本上不同的数据域（高红移、暗淡、不同类型的星系）时的基本且可预期的行为。高红移星系不仅仅是更远，它们的内禀属性可能不同，例如具有不同的恒星形成历史、金属丰度等，并且它们被观测到的属性被星系际介质吸收效应影响，而这种效应在低红移处没有类似样本。因此要求一个标准机器学习模型基于 $z < 1.5$ 的星系训练数据去预测一个 $z = 4$ 星系的红移是一个向物理上完全不同领域的外推。这也解释了为何融合了物理知识或试图模拟这些缺失数据（如 HAYATE 等混合模型 [86]）的方法对于攻克高红移预测至关重要。

4.3.2 测光巡天策略的影响

测光红移的质量与所依赖的测光数据库的深度、面积和波段覆盖直接挂钩，不同的巡天项目服务于不同的科学目标，其数据库特性也决定了其测光红移的应用范围。SDSS 作为著名的浅层、超广域光学巡天，使用五个宽带滤光片（$ugriz$）对超过 14,000 平方度的北天进行了成像。虽然与现代巡天相比其深度较浅（$r \sim 22.2$），但其庞大的光谱观测为数百万星系提供了光谱红移，构建了多年来机器学习测光红移发展所依赖的主要训练和验证样本，成为了低红移宇宙（$z < 0.7$）测光红移研究的基石[120]。然而其较浅的测光深度和仅有的五个光学波段，使其几乎无法用于更高红移的精确研究。后续的 Pan-STARRS1 巡天（PS1）使用五个滤光片（$grizy$）对整个北天（约 30,000 平方度）进行了比 SDSS 更深的观测（$r \sim 23.2$），提供了更大样本、更深、更均匀的多波段数据。

另一种巡天方案则采用深度、多波长巡天的巡天方式，如 COSMOS、CANDELS 等，这类巡天面积极小（0.22 ∼ 2 平方度），但在此天区内汇集了从哈勃、斯皮策空间望远镜到各大地面台站的超深度、从 X 射线到射电的全波段数据。这些深度、全色数据集为暗弱星系提供了目前最可靠的测光红移，几乎所有最先进的高红移测光红移算法都在这些数据集上得到验证。

现代深度、广域巡天是前两者的折衷，如 DES、KiDS、HSC-SSP 等，这类巡天旨在实现宇宙学应用所需的大面积与高精度测光红移的统一。它们往往覆盖数千平方度的天区，深度远超 SDSS，波段也延伸至近红外 Y 波段。这些巡天是当前利用弱引力透镜等手段测量暗

能量性质的主力军，其测光红移精度要求在“宇宙学样本”（$0.2 < z < 1.3$）内达到前所未有的 1-2% 水平。表2总结了目前一些主要测光巡天项目特征及测光红移精度的对比。

除了巡天天区，测光红移的精度也受限于测光的光谱分辨率。宽带系统（如 SDSS 的 $ugriz$, DES 的 $grizY$）拥有 4-6 个滤光片，观测效率高，能捕获大量光子，从而实现大天区的深度巡天。然而，它们对 SED 的采样非常粗糙，从而将精度限制在 $\sigma_z/(1+z) \approx 0.03-0.1$ 的水平。窄带系统（如 S-PLUS 的 12 个滤光片，PAUS 的 40 个滤光片）则提供高得多的光谱分辨率（$R \approx 20-50$），这使得它们能够分辨 4000Å 断裂等特征，甚至探测到强的发射线，如同为每个天体配备了一台极低分辨率的光谱仪，这时精度将显著提升至 $\sigma_z/(1+z) \approx 0.003-0.02$。其代价是在固定的曝光时间内，窄带滤光片的信噪比要低得多，从而限制了巡天的深度[130, 131]。

而即使光谱分辨率足够高，星系重叠问题也影响着测光精度[132]。随着 HSC 和 LSST 等巡天达到前所未有的深度，天空中星系的投影密度急剧增加，以至于高达 50% 甚至更多的天体在二维图像上发生物理重叠。当星系发生重叠时标准的测光算法难以将其光线分离开来，导致对单个组分的流量、颜色和形状的测量出现错误，其测光结果可能是来自处于完全不同红移的多个天体的混合，这会导致一个完全错误的测光红移估算，并显著增加离群点的数量，这已被认为是 LSST 弱引力透镜宇宙学研究中的一个主要系统误差来源[133, 134]。图像质量会直接影响重叠问题，特别是决定了点扩展函数 PSF 大小的大气“视宁度”。更好的视宁度意味着星系看起来更锐利，重叠的可能性更小，使其测光结果更可靠。位于优良视宁度台址（如 HSC 的 0.6 角秒）的地面巡天项目，在缓解重叠问题方面比位于较差台址的巡天项目具有显著优势。然而即使是 HSC 卓越的视宁度也无法完全消除这个问题，这突显了来自 HST/CANDELS 的空间衍射极限成像在提供未重叠真实样本方面的优势。

这些影响测光红移性能的因素并非孤立存在，而是相互耦合且有时会放大彼此的负面效应，特别是观测深度，它引发了观测策略上的一个矛盾点。现代巡天的主要驱动力是追求更深的观测以增加星系数量从而提高宇宙学测量的统计效果。然而增加深度直接导致两个负面后果，一是更大比例的星系样本信噪比变低；二是星系的投影密度增加，导致图像重叠的比例急剧上升。图像重叠污染了测光数据，而低信噪比也意味着测光本身就充满噪声。因此追求更深观测以改进宇宙学测量的行为，本身却在同时降低新增天体的基础数据质量。这些被降质的测光数据最终导致更差的测光红移进而引入系统误差，可能抵消掉最初通过增加星系数量获得的统计增益。这个反馈循环是当前巡天的问题，但同时该问题也正在推动着新算法的发展，这些新算法试图同时对测光、重叠和测光红移进行建模。

目前没有任何一个单一的巡天能够提供精确宇宙学所需的所有数据，不同的巡天项目会被用来相互校准。例如利用像 COSMOS 这样的深度天区为广域巡天如 DES 中数百万个没有光谱红移的暗弱星系校准测光红移。该方法通常涉及 SOM，首先利用所有可用的滤光片，通过结合广域巡天和深度近红外数据定义一个多维“颜色空间”，然后使用 SOM 将高维颜色空间映射到一个二维网格上，颜色相似的星系被放置在同一个“单元格”中，利用来自 COSMOS 等天区的深度多波段测光数据来填充这个映射。由于这些天区拥有高可靠性的测光红移或光谱红移，每个单元格都与一个已知的红移分布相关联。接着将广域巡天中的

表 2 一些主要测光巡天项目特征及测光红移精度

巡天项目	巡天面积 [$\deg^2$]	滤光片系统	5σ 极限星等 [mag]	典型机器学习法测光红移精度 σ_{NMAD}
SDSS	$\sim 14,500$	$ugriz$	$r \sim 22.2$	~ 0.02 [121]
Pan-STARRS1	$\sim 30,000$	$grizy$	$r \sim 23.2$	~ 0.03 [122]
COSMOS	~ 2	>30 个波段（紫外到射电）	$i \sim 26+$	$\sim 0.007-0.015$ [123]
CANDELS	$\sim$ 0.22	近红外 + 光学	$H \sim 27$	~ 0.01–0.02 [124]
DES	$\sim 5,000$	$grizY$	$r \sim 24.4$	$\sim 0.03-0.04$ [125]
KiDS	$\sim 1,350$	$ugri$	$r \sim 24.8$	$\sim 0.019-0.022$（明亮星系） [90]
HSC-SSP（宽场巡天）	$\sim 1,400$	$grizy$	$r \sim 26.0$	$\sim 0.04-0.05$ [126]
DESI Legacy	$\sim 14,000$	grz (+WISE)	$r \sim 23.9$	$\sim 0.014-0.026$ [73, 127]
S-PLUS	$\sim$ 9,300	5 个宽带 + 7 个窄带	$r \sim 21$(宽带)	$\sim 0.019-0.023$ [128]
LSST	$\sim 18,000$	$ugrizy$	$r \sim 27.5$(叠加)	要求 <0.02
Euclid（宽场巡天）	$\sim 14,000$	1 宽带 (VIS) + 3 近红外 (Y,J,H) + 地面 $ugriz$	$I \sim 24.5$	要求 <0.05 [129]

一个星系根据其观测到的颜色放置到相应的单元格中，其红移分布就被假定为该单元格由 COSMOS 数据确定的。这个过程有效地将来自小面积深度天区的高质量红移信息“转移”到大面积广域巡天中，从而实现了对弱引力透镜层析成像所需整体红移分布的校准。然而这个过程并不完美，不同巡天之间的测光存在系统性差异，且深度天区本身也可能不完全具有代表性。这些效应必须被建模，通常通过向真实图像中注入模拟星系来精确测量选择效应和测光偏差，例如 DES 中的 Balrog 代码[135]。

另一方面，不能简单地在一个巡天的数据上训练一个机器学习模型，然后将其应用于另一个巡天。即使滤光片名义上相似，滤光片透过率曲线、测光定标和数据处理流程中的细微差别也会导致测量出的颜色存在系统性偏差。机器学习模型会学习到这些特定的系统误差，当应用于具有不同系统误差的数据时就会失败。这突显了在不同巡天之间进行极其仔细的交叉定标和采用一致的数据处理流程的必要性。为了使测光红移校准稳健，要么使用完全相同的流程处理所有数据，要么精确地对不同数据集之间的偏差传递函数进行建模[55]。

5 总结与展望

天文学处于数据密集型科学的前沿，即将到来的大型测光巡天项目预计将产生 PB 甚至 EB 级的数据，迫切需要机器学习方法来实现高效的数据处理与分析。当前，深度学习已展现出从校准图像的像素级数据中直接预测测光红移的潜力，这一策略不仅能规避传统测光参数选择的偏差，还能挖掘图像中隐含的形态信息如表面亮度、颜色梯度等，为解决高红移预测精度下降和参数简并问题提供新路径[106]。

数据驱动的机器学习方法的优势还体现在多方面，通过无监督模型可识别多维特征空间中的最优适用区域，辅助评估测光与光谱空间的数据分布平衡性；结合贝叶斯统计和混合模型，可实现红移、星系恒星质量和 SFR 的联合预测[26, 28]；利用 SOM 等技术，可有效检测训练数据中的异常值和不可靠光谱样本[35]。

通过结合经验模型、SED 模板拟合和贝叶斯推断的混合方法，未来的发展趋势将聚焦于跨方法融合与标准化评估，红移估计精度将得到提高并扩展至高红移区域。而大规模测光巡天通过统一数据和评估指标，推动了不同方法的公平比较，例如 LSST 中通过 PIT 和 QQ 图验证了深度学习模型在 PDF 估计中的有效性[32]。

计算科学与天文学的深度融合，将是应对未来海量天文数据的重要策略，其目标不仅是提升测光红移的精度，更在于为深入开展包括宇宙学参数测量、暗物质分布探测等前沿科学研究提供可靠的基础工具[91]。

参考文献：

[1] Baum W A. IAU Symposium, 1962, 15: 390

[2] Butchins S A. A&A, 1981, 97(2): 407–409
[3] Connolly A J, Csabai I, Szalay A S, et al. AJ, 1995, 110: 2655
[4] York D G, Adelman J, Anderson J, JOHN E., et al. AJ, 2000, 120(3): 1579–1587
[5] Scoville N, Aussel H, Brusa M, et al. ApJS, 2007, 172(1): 1–8
[6] Dark Energy Survey Collaboration, Abbott T, Abdalla F B, et al. MNRAS, 2016, 460(2): 1270–1299
[7] de Jong J T A, Kuijken K, Applegate D, et al. The Messenger, 2013, 154: 44–46
[8] Aihara H, Arimoto N, Armstrong R, et al. PASJ, 2018, 70: S4
[9] Kauffmann O B, Le Fèvre O, Ilbert O, et al. A&A, 2020, 640: A67
[10] Laureijs R, Amiaux J, Arduini S, et al. arXiv e-prints, 2011: arXiv:1110.3193. https://www.research.ed.ac.uk/en/publications/euclid-definition-study-report
[11] LSST Science Collaboration: Abell P A, Allison J, Anderson S F, et al. arXiv e-prints, 2009: arXiv:0912.0201. https://www.lsst.org/scientists/scibook
[12] Green J, Schechter P, Baltay C, et al. arXiv e-prints, 2012: arXiv:1208.4012. https://research.manchester.ac.uk/en/publications/wide-field-infrared-survey-telescope-wfirst-final-report
[13] Gong Y, Liu X, Cao Y, et al. ApJ, 2019, 883(2): 203
[14] Witten C, Laporte N, Martin-Alvarez S, et al. Nature Astronomy, 2024, 8: 384–396
[15] Ilbert O, Arnouts S, McCracken H J, et al. A&A, 2006, 457(3): 841–856
[16] Brammer G B, van Dokkum P G, Coppi P. ApJ, 2008, 686(2): 1503–1513
[17] Bruzual G, Charlot S. MNRAS, 2003, 344(4): 1000–1028
[18] Assef R J, Kochanek C S, Brodwin M, et al. ApJ, 2010, 713(2): 970–985
[19] Retana-Montenegro E, Röttgering H J A. A&A, 2020, 636: A12
[20] Holwerda B W, Hsu C C, Hathi N, et al. MNRAS, 2024, 529(2): 1067–1081
[21] Lu J, Luo Z, Chen Z, et al. MNRAS, 2024, 527(4): 12140–12153
[22] Dennis M T, Hu E M, Cowie L L. ApJ, 2025, 983(2): 173
[23] Dey B, Andrews B H, Newman J A, et al. MNRAS, 2022, 515(4): 5285–5305
[24] Feroz F, Hobson M P, Cameron E, et al. The Open Journal of Astrophysics, 2019, 2(1): 10
[25] de Diego J A, Nadolny J, Bongiovanni Á, et al. A&A, 2021, 655: A56
[26] Bonjean V, Aghanim N, Salomé P, et al. A&A, 2019, 622: A137
[27] Wright E L, Eisenhardt P R M, Mainzer A K, et al. AJ, 2010, 140(6): 1868–1881
[28] Mucesh S, Hartley W G, Palmese A, et al. MNRAS, 2021, 502(2): 2770–2786
[29] Euclid Collaboration, Enia A, Bolzonella M, et al. A&A, 2024, 691: A175
[30] Euclid Collaboration: Desprez G, Paltani S, Coupon J, et al. A&A, 2020, 644: A31
[31] Bisigello L, Kuchner U, Conselice C J, et al. MNRAS, 2020, 494(2): 2337–2354
[32] Schmidt S J, Malz A I, Soo J Y H, et al. MNRAS, 2020, 499(2): 1587–1606
[33] Baron D. arXiv e-prints, 2019: arXiv:1904.07248
[34] Narendra A, Dainotti M G, Sarkar M, et al. A&A, 2025, 698: A92
[35] Razim O, Cavuoti S, Brescia M, et al. MNRAS, 2021, 507(4): 5034–5052
[36] Stölzner B, Joachimi B, Korn A, et al. MNRAS, 2023, 519(2): 2438–2450
[37] Jalan P, Bilicki M, Hellwing W A, et al. A&A, 2024, 692: A177
[38] Wright A H, Hildebrandt H, van den Busch J L, et al. A&A, 2020, 637: A100
[39] Brescia M, Salvato M, Cavuoti S, et al. MNRAS, 2019, 489(1): 663–680
[40] Broussard A, Gawiser E. ApJ, 2021, 922(2): 153
[41] Beck R, Dodds S C, Szapudi I. MNRAS, 2022, 515(4): 4711–4721
[42] Sen S, Singh K P, Chakraborty P. New Astronomy, 2023, 99: 101959
[43] Qu H, Sako M. ApJ, 2023, 954(2): 201
[44] Crenshaw J F, Kalmbach J B, Gagliano A, et al. AJ, 2024, 168(2): 80
[45] Zhou X, Gong Y, Zhang X, et al. ApJ, 2024, 977(1): 69
[46] Pathi I M, Soo J Y H, Wee M J, et al. JCAP, 2025, 2025(1): 097
[47] Hoyle B, Rau M M, Zitlau R, et al. MNRAS, 2015, 449(2): 1275–1283

[48] Baron D, Poznanski D. MNRAS, 2017, 465(4): 4530–4555
[49] Reis I, Baron D, Shahaf S. AJ, 2019, 157(1): 16
[50] Lin Q, Ruan H, Fouchez D, et al. A&A, 2024, 691: A331
[51] Cavuoti S, Brescia M, D'Abrusco R, et al. MNRAS, 2014, 437(1): 968–975
[52] Jones E, Singal J. A&A, 2017, 600: A113
[53] Humphrey A, Cunha P A C, Paulino-Afonso A, et al. MNRAS, 2023, 520(1): 305–313
[54] Li C, Zhang Y, Cui C, et al. AJ, 2024, 168(6): 233
[55] Janiurek L, Hendry M A, Speirits F C. MNRAS, 2024, 533(3): 2786–2800
[56] Ye G, Zhang H, Wu Q. ApJS, 2024, 275(1): 19
[57] Huang J, Luo B, Brandt W N, et al. ApJ, 2025, 979(2): 107
[58] Kim T, Sohn J, Hwang H S, et al. ApJS, 2025, 277(2): 41
[59] Carrasco Kind M, Brunner R J. MNRAS, 2013, 432(2): 1483–1501
[60] Reza M. New Astronomy, 2025, 115: 102316
[61] Han B, Qiao L N, Chen J L, et al. Research in Astronomy and Astrophysics, 2021, 21(1): 017
[62] Curran S J, Moss J P, Perrott Y C. MNRAS, 2021, 503(2): 2639–2650
[63] Luken K J, Norris R P, Park L A F, et al. Astronomy and Computing, 2022, 39: 100557
[64] Luken K J, Norris R P, Wang X R, et al. Publications of the Astronomical Society of Australia, 2023, 40: e039
[65] Soo J Y H, Joachimi B, Eriksen M, et al. MNRAS, 2021, 503(3): 4118–4135
[66] Naidoo K, Johnston H, Joachimi B, et al. A&A, 2023, 670: A149
[67] Ansari Z, Agnello A, Gall C. A&A, 2021, 650: A90
[68] Teixeira G, Bom C R, Santana-Silva L, et al. Astronomy and Computing, 2024, 49: 100886
[69] Zhou X, Gong Y, Meng X M, et al. MNRAS, 2022, 512(3): 4593–4603
[70] Yao L, Qiu B, Luo A L, et al. MNRAS, 2023, 523(4): 5799–5811
[71] Ait Ouahmed R, Arnouts S, Pasquet J, et al. A&A, 2024, 683: A26
[72] Jones E, Do T, Li Y Q, et al. ApJ, 2024, 974(2): 159
[73] Zhou X, Li N, Zou H, et al. MNRAS, 2025, 536(3): 2260–2276
[74] Mu Y H, Qiu B, Zhang J N, et al. Research in Astronomy and Astrophysics, 2020, 20(6): 089
[75] Li M, Gao Z, Qiu B, et al. MNRAS, 2021, 506(4): 5923–5934
[76] Zhou X, Gong Y, Meng X M, et al. Research in Astronomy and Astrophysics, 2022, 22(11): 115017
[77] Ejaz Awan S, Bennamoun M, Sohel F, et al. Neurocomputing, 2021, 453: 164–171
[78] Luo Z, Tang Z, Chen Z, et al. MNRAS, 2024, 531(3): 3539–3550
[79] Fotopoulou S, Paltani S. A&A, 2018, 619: A14
[80] Alam S, Albareti F D, Allende Prieto C, et al. ApJS, 2015, 219(1): 12
[81] Carrasco Kind M, Brunner R J. MNRAS, 2014, 442(4): 3380–3399
[82] Duncan K J, Brown M J I, Williams W L, et al. MNRAS, 2018, 473(2): 2655–2672
[83] Crenshaw J F, Connolly A J. AJ, 2020, 160(4): 191
[84] Li Y, Fu L, Chen Z, et al. Research in Astronomy and Astrophysics, 2025, 25(5): 055021
[85] Leistedt B, Hogg D W. ApJ, 2017, 838(1): 5
[86] Tanigawa S, Glazebrook K, Jacobs C, et al. MNRAS, 2024, 530(2): 2012–2038
[87] Dalmasso N, Pospisil T, Lee A B, et al. Astronomy and Computing, 2020, 30: 100362
[88] Hildebrandt H, Arnouts S, Capak P, et al. A&A, 2010, 523: A31
[89] Cavuoti S, Brescia M, Longo G, et al. A&A, 2012, 546: A13
[90] Bilicki M, Hoekstra H, Brown M J I, et al. A&A, 2018, 616: A69
[91] Mandelbaum R. ARA&A, 2018, 56: 393–433
[92] Amaro V, Cavuoti S, Brescia M, et al. MNRAS, 2019, 482(3): 3116–3134
[93] Harrison D, Sutton D, Carvalho P, et al. MNRAS, 2015, 451(3): 2610–2624
[94] Izbicki R, Lee A B. Electronic Journal of Statistics, 2017, 11(2): 2800 – 2831
[95] Ma Z, Hu W, Huterer D. ApJ, 2006, 636(1): 21–29
[96] Blake C, Bridle S. MNRAS, 2005, 363(4): 1329–1348

[97] Csabai I, Budavári T, Connolly A J, et al. AJ, 2003, 125(2): 580–592
[98] Bolzonella M, Miralles J M, Pelló R. A&A, 2000, 363: 476–492
[99] Wittman D, Riechers P, Margoniner V E. ApJ, 2007, 671(2): L109–L112
[100] D'Isanto A, Polsterer K L. A&A, 2018, 609: A111
[101] Fu L, Liu D, Radovich M, et al. MNRAS, 2018, 479(3): 3858–3872
[102] Cappelluti N, Predehl P, Böhringer H, et al. Memorie della Societa Astronomica Italiana Supplementi, 2011, 17: 159
[103] Euclid Collaboration: Humphrey A, Bisigello L, Cunha P A C, et al. A&A, 2023, 671: A99
[104] Salvato M, Ilbert O, Hoyle B. Nature Astronomy, 2019, 3: 212–222
[105] Cantiello M, Blakeslee J P. arXiv e-prints, 2023: arXiv:2307.03116
[106] Pasquet J, Bertin E, Treyer M, et al. A&A, 2019, 621: A26
[107] Breiman L. Machine learning, 2001, 45(1): 5–32
[108] Carrasco Kind M, Brunner R J. MNRAS, 2014, 438(4): 3409–3421
[109] Salvato M, Hasinger G, Ilbert O, et al. ApJ, 2009, 690(2): 1250–1263
[110] Ananna T T, Salvato M, LaMassa S, et al. ApJ, 2017, 850(1): 66
[111] Norris R P, Salvato M, Longo G, et al. PASP, 2019, 131(1004): 108004
[112] Saxena A, Salvato M, Roster W, et al. A&A, 2024, 690: A365
[113] Salvato M, Wolf J, Saxena A, et al. EAS2024, European Astronomical Society Annual Meeting, 2024: 2540
[114] Crenshaw J F, Leistedt B, Graham M L, et al. arXiv e-prints, 2025: arXiv:2503.06016
[115] Kunsági-Máté S, Beck R, Szapudi I, et al. MNRAS, 2022, 516(2): 2662–2670
[116] Luo Z, Li Y, Lu J, et al. MNRAS, 2024, 535(2): 1844–1855
[117] Henghes B, Thiyagalingam J, Pettitt C, et al. MNRAS, 2022, 512(2): 1696–1709
[118] Beck R, Lin C A, Ishida E E O, et al. MNRAS, 2017, 468(4): 4323–4339
[119] Sánchez C, Alarcon A, Bernstein G M, et al. MNRAS, 2023, 525(3): 3896–3922
[120] Soo J Y H, Shuaili I Y K A, Pathi I M. American Institute of Physics Conference Series, 2023, 2756: 040001
[121] Beck R, Dobos L, Budavári T, et al. MNRAS, 2016, 460(2): 1371–1381
[122] Tarrío P, Zarattini S. A&A, 2020, 642: A102
[123] Eriksen M, Alarcon A, Cabayol L, et al. MNRAS, 2020, 497(4): 4565–4579
[124] Kodra D, Andrews B H, Newman J A, et al. ApJ, 2023, 942(1): 36
[125] Toribio San Cipriano L, De Vicente J, Sevilla-Noarbe I, et al. A&A, 2024, 686: A38
[126] Tanaka M, Coupon J, Hsieh B C, et al. PASJ, 2018, 70: S9
[127] Li C, Zhang Y, Cui C, et al. MNRAS, 2023, 518(1): 513–525
[128] Lima E V R, Sodré L, Bom C R, et al. Astronomy and Computing, 2022, 38: 100510
[129] Stanford S A, Masters D, Darvish B, et al. ApJS, 2021, 256(1): 9
[130] Laur J, Tempel E, Tamm A, et al. A&A, 2022, 668: A8
[131] Navarro-Gironés D, Gaztañaga E, Crocce M, et al. MNRAS, 2024, 534(2): 1504–1527
[132] Melchior P, Joseph R, Sanchez J, et al. Nature Reviews Physics, 2021, 3(10): 712–718
[133] Boucaud A, Huertas-Company M, Heneka C, et al. MNRAS, 2020, 491(2): 2481–2495
[134] Nourbakhsh E, Tyson J A, Schmidt S J, et al. MNRAS, 2022, 514(4): 5905–5926
[135] Everett S, Yanny B, Kuropatkin N, et al. ApJS, 2022, 258(1): 15

Machine Learning for Photometric Redshift Estimation

LU Jun-hao, LUO Zhi-jian, CHEN Jian-zhen, ZENG Lu-jin, SHU Cheng-gang

(Shanghai Key Lab for Astrophysics, Shanghai Normal University, Shanghai 200234, China)

Abstract: Photometric redshift (photo-z) is a vital method for estimating the redshifts of galaxies and quasars by multi-wavelength photometric data. With the exponential growth in observational data, the efficiency of traditional spectroscopic redshift (spec-z) measurements can no longer meet the demand for massive redshift information required by current and next-generation large-scale sky surveys. Consequently, most imaging-based galaxy and quasar surveys rely heavily on photometric redshifts to provide redshift information for tens of millions to billions of celestial objects, enabling forefront research on the large-scale structure of the universe and the nature of dark energy. Against this background, machine learning (ML) methods have emerged as mainstream tools for obtaining high-precision photometric redshifts due to their efficiency and scalability. In the present paper, a comprehensive review of recent advances in the application of ML in photo-z estimation, including algorithmic classifications, model optimization strategies, and typical application scenarios, together with examples to show the technical characteristics and performance differences of various ML architectures, are summarized.

Key words: photometric redshift; machine learning; data analysis