Predicting lysine methylation sites using a convolutional neural network 论文复现

摘要

赖氨酸甲基化是一种重要的翻译后修饰,与多种疾病密切相关。本文复现并分析了 CNN-Meth 模型，该模型通过卷积神经网络结合进化特征、结构特征、物理化学特征和二进制编码，实现了赖氨酸甲基化位点的自动化预测。复现过程中，本文首先明确了数据处理流程，验证了 KNN 下采样在数据平衡中的有效性，并探讨了不同模型架构对性能的影响。最后的实验结果显示，复现模型在几项性能指标上接近原文，验证了 CNN-Meth 方法的有效性，并为 PTM 预测提供了新的思路和方法。

关键词： Methylation；Automated Feature Extraction；Convolutional Neural Network

1 引言

翻译后修饰 (Post translational modifications, PTMs) 是指某些蛋白质合成后，氨基酸侧链发生的酶促变化，常见的有磷酸化、泛素化、甲基化和乙酰化等等。PTMs 是调节蛋白质功能的一种机器重要的机制，一些 PTMs 异常已知与癌症、阿尔茨海默病和心血管疾病等人类疾病有关 [3]。

甲基化 (Methylation) 是一种可逆的 PTM，在真核生物中赖氨酸甲基化较为常见，其过程由赖氨酸甲基化酶和赖氨酸去甲基化酶分别催化甲基的添加和去除，如图 1 所示。当甲基化功能失调时，可能引发癌症、精神健康障碍以及发育异常等疾病。目前，已有基于甲基化可逆性的药物治疗方法用于应对相关疾病，而寻找甲基化位点则是其中关键的一环 [4]。

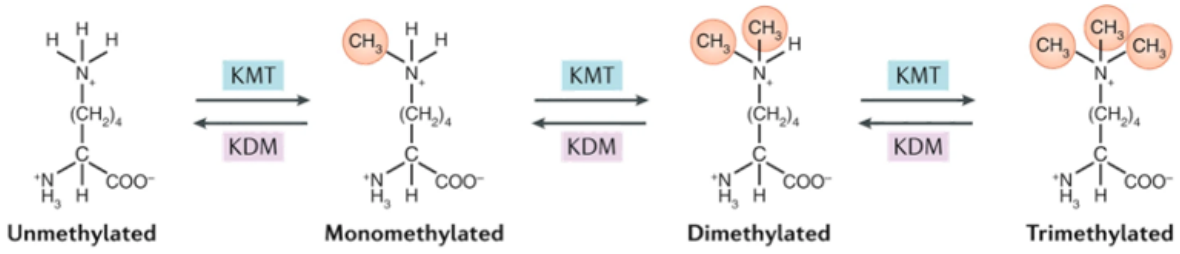

图 1. 赖氨酸甲基化相关过程

在过去，通过生物实验的方法发现甲基化点位的方法费时又费力 [5]。而近些年机器学习方法的成熟和识别 PTM 的计算方法的成功实践，显著提升了识别的效率与精确度 [15]，给甲

基化点位识别研究带来了新的思路与突破。Austin Spadaro 等人提出的 CNN-Meth 方法首次将卷积神经网络和包含进化、结构、物理化学特性以及二进制编码的组合特征结合，创新性地应用在了赖氨酸甲基化点位预测任务上，在准确度、灵敏度和特异性等指标上都优于先前的研究 [17]。本文将围绕 CNN-Meth 模型展开复现与分析，探讨其方法中的关键环节，并进一步评估数据平衡策略和模型改进对预测性能的影响。

2 相关工作

随着生物信息学和计算方法的不断发展，基于机器学习的工具在 PTMs 识别中取得了显著进展。针对赖氨酸甲基化位点预测，研究着们提出了多种方法尝试从不同特征视角优化预测模型的性能。

2.1 传统机器学习方法

传统机器学习方法在赖氨酸甲基化位点预测中占据着重要地位，这些方法通常依赖于手工特征提取技术，从序列或结构中提取有意义的信息。Shi 等人提出的 PLMLA 使用了从 SPIDER 2 提取的二级结构特征以及组权重和氨基酸组成特征结合支持向量机进行赖氨酸甲基化点位预测 [16]，而 Ju 等人则在 iLM-2L 中使用了 k-gap 氨基酸对和伪氨基酸组成的特征 [10]。之后，在 Lee 等人提出的 MethK 模型中使用了基于序列的特征和源自 PSSM 矩阵的进化特征的组合训练，实现了更优的预测结果 [12]。

2.2 多 PTM 标签预测方法

多 PTM 标签的预测方法可以同时预测赖氨酸上多种 PTM 修饰类型，提供了比单一 PTM 预测方法更为全面的蛋白质修饰点位预测。Hasan 等人提出的 mLysPTMpred 使用经过伪氨基酸组成特征训练的 SVM 同时预测赖氨酸是否是潜在的乙酰化、巴豆酰化、甲基化或琥珀酰化位点，但是没有说明仅预测甲基化位点的准确性 [8]。而 Ahmed 等人提出的 predML-Site 模型是用基于序列的特征进行训练，优化了特征表示方法和数据平衡策略，进一步提高了多标签预测的精确度 [1]。

以上的研究主要都是依靠手工提取氨基酸一种或两种特性来进行特征提取，通常依赖于研究人的先验知识，可能会忽略一些有用的信息，无法全面捕捉的潜在的有助于预测的重要模式。包含 CNN 在内的自动化特征提取方法可以从原始数据中自动学习到更全面、更精确的特征，但从未被用于赖氨酸甲基化点位修饰预测的任务。

3 本文方法

3.1 本文方法概述

本文提出了一种名为 CNN-Meth 的深度学习方法，用于预测蛋白质赖氨酸甲基化位点。该方法以 CNN 为核心，结合进化特征、结构特征、物理化学特征以及二进制编码，将目标赖氨酸及其周围 31 个氨基酸的特征表示为二维矩阵输入到模型中，通过 CNN 自动提取特征，避免传统手工特征提取可能导致的信息损失，从而实现对赖氨酸甲基化位点的精准识别。

3.2 数据准备与处理

本文使用了 Protein Lysine Modification Database (PLMD) 数据集 [13,14,18]，该数据集包含多种蛋白质赖氨酸修饰的实验验证点位数据，是目前最大的赖氨酸甲基化点位资源。数据集中包含 6,322 个甲基化位点和 133,293 个非甲基化位点，为构建预测模型提供了充足的样本基础。为减少数据冗余，研究中使用了 CD-HIT 工具 [9] 去除数据集中序列相似度高于 40% 的重复序列，经过去冗余处理后的数据集保留了 4,913 个甲基化位点和 108,958 个非甲基化位点，保证了数据的独立性和多样性。

为进一步统一数据格式，本文对目标赖氨酸附近的氨基酸序列进行了长度标准化处理。具体而言，选取目标赖氨酸上下游各 15 个氨基酸作为特征窗口。对于不足 15 个上下游氨基酸的序列，通过“镜像扩展”方法填补空缺，即复制现有的上下游序列并反转后补充到缺失区域，从而保证所有序列长度一致。这一处理有效避免了因序列长度不一致对模型特征提取带来的干扰。

此外，由于数据集中甲基化位点和非甲基化位点数量分布极不平衡，为避免模型在训练过程中对多数类别的偏倚，本文采用 K 近邻 (K-Nearest Neighbor, KNN) 方法对非甲基化位点进行了下采样，将正负样本比例调整为 1:3，通过最大化样本间的欧氏距离筛选非甲基化位点，使负样本具有更高的代表性。KNN 使用的欧几里得距离公式如下所示：

$$\text{Euclidean Distance} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}, \tag{1}$$

经过平衡处理后，最终数据集包含 4,913 个甲基化位点和 14,739 个非甲基化位点，确保了数据分布的均衡性并保留了尽可能多的非冗余信息。

3.3 特征获取

本文构建了一个二维特征矩阵用于表示目标赖氨酸及其邻域氨基酸序列的特征，并作为 CNN 网络的特征输入。特征矩阵整合了进化特征、结构特征、物理化学特征和二进制编码，构建了一个 31×103 的特征矩阵，从不同层面全面表征序列信息。

3.3.1 进化特征

进化特征通过位置特异性评分矩阵 (Position-Specific Scoring Matrix, PSSM) 提取。PSSM 可由基于迭代比对的工具 PSI-BLAST [2] 计算得出的，其通过查询序列与数据库中相似序列的多轮比对，捕捉目标序列中每个位置的保守性和进化特性。使用 PSI-BLAST 工具生成的 PSSM 是一个 $L \times 20$ 的矩阵，其中 L 表示序列长度，本文 L 为窗口长度 31，每一列表示 20 种氨基酸的替代概率。通过综合分析目标赖氨酸及其上下游序列的替代模式，PSSM 能有效捕捉序列中赖氨酸位点的进化保守性信息，为模型提供关键的进化特征支持。

3.3.2 结构特征

结构特征是蛋白质序列的重要表征，能够捕捉目标赖氨酸及其邻域氨基酸的空间构象信息和局部结构特性。在本文中，结构特征由 SPIDER2 工具 [19] 计算，包含可达表面积 (Accessible Surface Area, ASA)、局部主链/二面角（ϕ, ψ, θ, τ）和二级结构信息（螺旋、片层和

环）共 8 个子特征。其中 ASA 表示氨基酸与溶剂的接触面积，反映目标赖氨酸在蛋白质中的溶剂可及性；局部主链/二面角以四个角度值描述了氨基酸的几何形态和空间定位；二级结构信息通过三种状态描述氨基酸的局部折叠模式。最终整合后的结构特征维度为 31×8，全面刻画了目标赖氨酸位点的空间结构和局部环境，为模型提供了多层次的结构信息支持。

3.3.3 物理化学特征

物理化学特征使用了 AAIndex 数据库中经过筛选的 55 项有效属性 [6, 7, 11]，包括氨基酸的物理、化学及物理化学性质，例如亲水性、极性、体积等。以窗口长度为 31 的序列为单位，每个样本生成一个 31×55 的矩阵，描述目标位点及其邻域的物理化学特性。

3.3.4 二进制编码

二进制编码将每个氨基酸类型转换为一个 1×20 的二进制向量，其中仅对应目标氨基酸的元素为 1，其余为 0。通过连接窗口内的 31 个氨基酸的编码，形成一个 31×20 的特征矩阵，如图 2 所示，二进制编码可以简单高效地表示氨基酸序列信息。

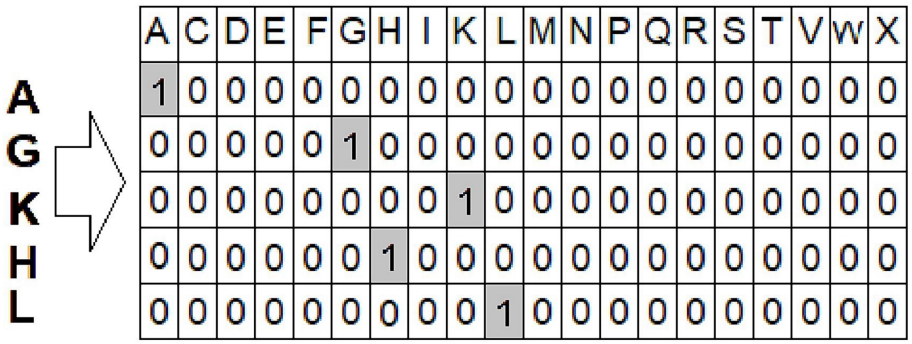

图 2. 氨基酸序列二进制编码

3.4 模型架构

CNN 架构如图 3 所示。CNN-meth 由 2 层卷积组成，均包含 64 个卷积核，每个卷积核的窗口大小为 [3,3]。每个卷积层还与使用窗口大小 [2,2] 的最大池化层配对。然后，这些层的输出展平处理后输入到一系列完全连接的隐藏层中。其中包括 5 个含有 50 个神经元的全链接层，2 个丢弃率为 0.2 的丢弃层，以及一个具有单个神经元的输出层。

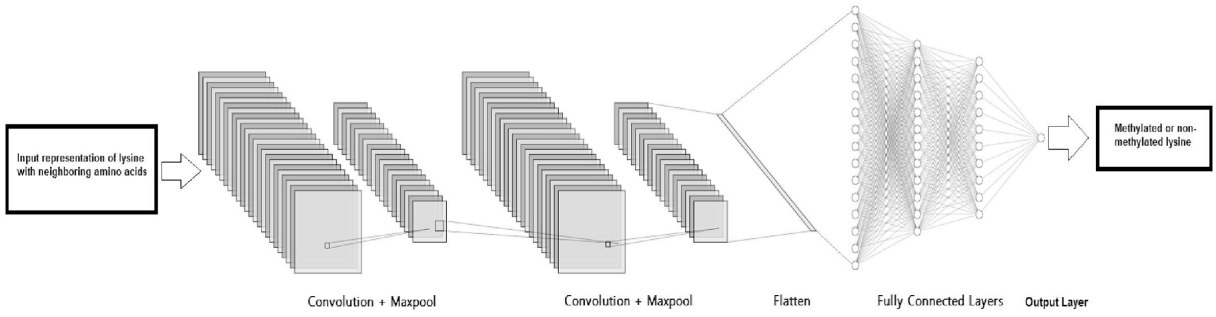

图 3. CNN 模型架构

3.5 评价指标

本文采用多种评价指标对模型的预测性能进行评估，包括准确率 (Accuracy)、灵敏度 (Sensitivity)、特异性 (Specificity) 和马修斯相关系数 (MCC)。其中，准确率用于衡量模型对所有样本的整体预测准确程度；灵敏度反映模型对正类样本（赖氨酸甲基化位点）的识别能力，定义为真正例占所有正类样本的比例；特异性用于衡量模型对负类样本（非甲基化位点）的识别能力，定义为真负例占所有负类样本的比例；MCC 是一种综合评价指标，结合了正类和负类的预测结果，能够更好地反映模型在不平衡数据集上的性能，取值范围为 -1 到 1，其中 1 表示完全预测正确，0 表示随机预测，-1 表示完全预测错误。这些指标共同构成了模型性能的全面评价体系，可验证模型在实际应用中的有效性和可靠性。四个指标的具体计算公式如下：

$$\text{sensitivity} = \frac{PS_+}{PS_+ + PS_-} \tag{2}$$

$$\text{specificity} = \frac{NS_+}{NS_+ + NS_-} \tag{3}$$

$$\text{accuracy} = \frac{PS_+ + NS_+}{PS_+ + PS_- + NS_+ + NS_-} \tag{4}$$

$$\text{MCC} = \frac{(NS_+ \times PS_+) - (NS_- \times PS_-)}{\sqrt{(PS_+ + PS_-)(PS_+ + NS_-)(NS_+ + PS_-)(NS_+ + NS_-)}} \tag{5}$$

以上公式中，PS+ 代表真正例，NS+ 代表真负例，PS 代表假正例，NS 代表假负例。

4 复现细节

4.1 与已有开源代码对比

原作者仅在 Github 仓库https://github.com/MLBC-lab/CNN-Meth中提供了少量独立数据集、用于获取物理化学特征和二进制编码的代码、组合形成特征矩阵的代码、测试代码以及 tensorflow 导出的模型文件，未提供训练数据集以及模型代码。因此，复现工作包括了数据特征获取和模型复现两大部分。其中数据特征获取包含下载元数据、正负样本分割、去冗余、KNN、部署比对数据库、生成进化特征与结构特征等步骤及相关代码的撰写；模型复现对作者提供的模型与原文描述的模型进行了复现，对两个模型之间的矛盾进行了比较与分析；最后对不同的数据平衡优化算法进行了尝试与实验分析。

4.2 实验环境搭建

硬件环境	CPU	Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz
	GPU	NVIDIA GeForce RTX 3090
	内存	377G
操作系统	OS	Ubuntu 22.04.4 LTS
开发环境	编程语言	Python 3.8.5
	深度学习框架	Pytorch
	开发工具	PyCharm 2024.2.3 (Community Edition)

4.3 赖氨酸甲基化点位数据获取

本文使用的 PLMD 数据集现已整合进 Compendium of Protein Lysine Modifacations 4.0 (CPLM 4.0) 数据集 [20] 中，可从链接https://cplm.biocuckoo.cn/进入，可选择下载指定物种或指定的修饰类型的数据。下载的数据格式主要有序列 ID、修饰类型、修饰点位和原始序列。同一序列上存在多个可修饰点位，同一可修饰点位上可能发生多种不同的修饰类型。

4.4 数据处理流程

在本文并没有对去冗余、数据平衡、特征获取和序列裁切的处理顺序做说明，通过对各个操作环节分析得到完整正确的处理流程应如图 4 所示。对于从 CPLM 上获取的原始数据，应首先基于完整的蛋白质序列是用 CD-Hit 工具去冗余。实际在复现过程中仅选用智人的赖氨酸甲基化点位数据，而 CD-Hit 工具会导致稀少的正样本变得更少，因此最后没有采用 CD-Hit 去冗余步骤。

物理化学特征和二进制编码只依赖与单个氨基酸极其领域，属于局部的固有属性，可以在获取前先对序列裁切从而提高处理效率。而 PSI-BLAST 需要通过查询序列与完整数据库中的蛋白质比对来捕获远缘同源序列中的进化保守性，SPIDER2 预测结构特征需要依赖于蛋白质的全局三维构象，都基于完整序列构建特征矩阵，因此进化特征和结构特征都需要对完整蛋白质序列处理后再进行裁切。

随后，将基于局部窗口序列的特征矩阵根据修饰类型划分为正样本和负样本，再使用 KNN 算法对负样本进行下采样，从而平衡正负样本。最后将处理完的样本特征输入到 CNN 网络中进行训练。然而由于用于比对的 nr 数据量高达几百 GB，未能成功部署和用于生成进化特征和结构特征，在模型复现实验中将使用作者提供的独立数据集。

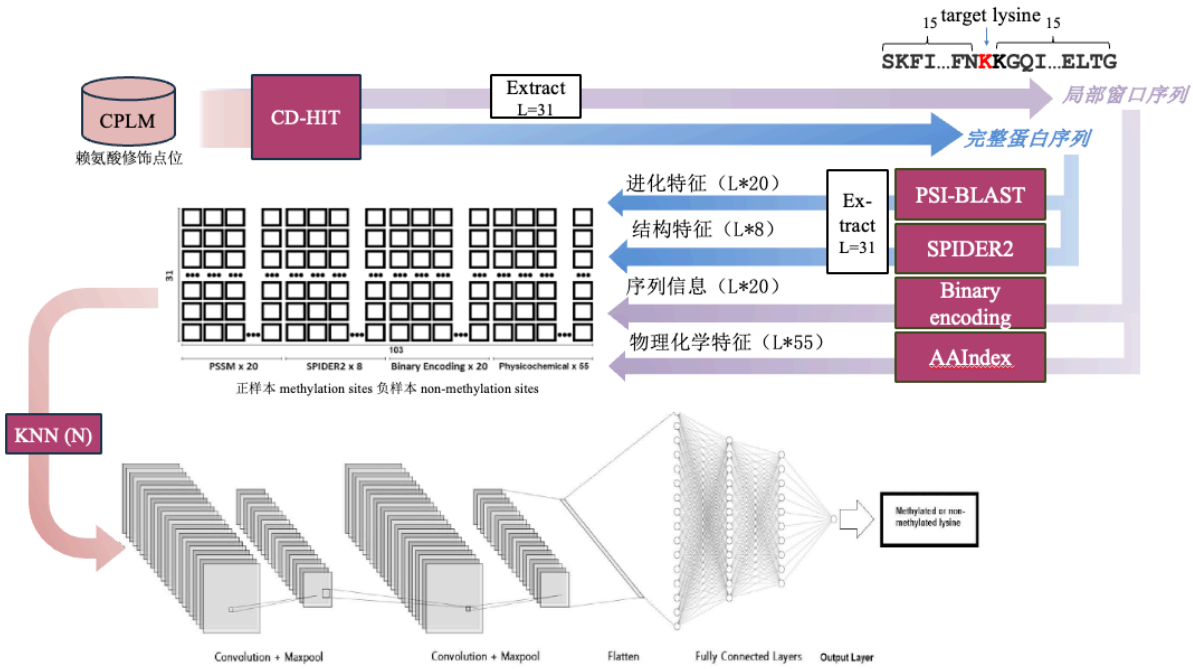

图 4. CNN-Meth 整体处理流程

4.5 模型复现

通过对作者提供的模型结构进行导出分析，CNN-Meth 包含两个卷积层，每层由 64 个大小为 [3,3] 的卷积核组成，展平后连接 4 个全连接层和 2 个丢弃层。然而，该模型与论文中的描述存在不一致之处，缺少卷积层后的两个池化层和一个额外的全连接层。因此，在后续实验中对这两种模型结构分别进行了对比与分析。此外，模型训练使用了二元交叉熵作为损失函数，训练参数设置为 20 个 epoch 和批量大小为 32。

4.6 创新点

本文首次将四种经过验证的特征组合与 CNN 分类器相结合应用于 PTM 预测，并通过 CNN 自动提取 PSSM 特征，避免了传统手工特征提取可能导致的信息丢失。在复现过程中，进一步探讨了通过上采样方法处理样本不平衡问题的可行性。

5 实验结果分析

CNN-Meth 仓库中提供的独立测试集含有，983 个正样本和 20237 个负样本，使用本文提供的已训练好的模型测试的结果在表 1 最后一行，可见与本文中的结果十分相近。

Model	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
PLMLA	82.6	83.5	83.0	0.66
iLM-2L	85.3	91.9	76.5	0.69
MethK (histone)	81.4	80.3	85.6	0.56
MethK (non-histone)	87.2	88.7	69.1	0.44
mLysPTMpred	83.7	–	–	–
predML-Site	84.2	–	–	–
CNN-Meth	96.0	96.4	85.1	0.65
CNN-Meth*	**95.7**	**96.2**	**85.1**	**0.64**

表 1. 将 CNN-Meth 与独立测试集文献中发现的先前研究进行比较。

使用独立数据集训练的复现模型分为两种：根据提供的模型结构复现的模型记为 M，按照论文描述复现的模型记为 P。两种模型在进行 KNN 下采样平衡正负样本与未进行处理的对比实验结果如表 2 所示。不进行样本平衡处理时，由于负样本数量远多于正样本，尽管准确率和特异性与原文结果接近，但敏感度和 MCC 均显著降低。采用 KNN 下采样能够有效缓解数据不平衡对敏感度的影响，但由于样本量减少，整体模型性能仍有一定局限。此外，从表 2 中可以看出，在数据较为平衡的情况下，按照论文描述复现的模型 P 的预测结果略胜于 M 模型。在验证了 KNN 下采样的有效性后，进一步对平衡后的复现模型与原文结果进行了 5 折和 10 折交叉验证，结果如表 3 所示。

在最后，使用 SMOTE 对训练集样本过采样至正负样本数量为 1:3 后对模型进行训练和测试，与 NKK 下采样的比较结果如表 4 所示。实验结果表明 SMOTE 上采样的表现不如

方法	样本数量	模型	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
原文	4913 : 14739	CNN-Meth	96.0	96.4	85.1	0.65
不处理	983 : 20237	CNN-Meth_M	95.5	98.6	32.0	0.39
		CNN-Meth_P	95.2	98.7	24.4	0.32
KNN 下采样	983 : 2949	CNN-Meth_M	85.8	93.2	63.4	0.60
		CNN-Meth_P	86.4	93.6	65.0	0.62

表 2. 不同复现的模型下 KNN 消融实验结果

方法	交叉验证	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
原文	5-Fold	88.4	92.6	78.7	0.69
	10-Fold	90.1	93.0	81.7	0.74
复现	5-Fold	85.2	90.6	68.9	0.61
	10-Fold	86.2	92.4	67.5	0.62

表 3. 复现模型和原模型交叉验证实验结果比较

KNN 下采样，推测原因在于 SMOTE 采用插值法生成新的正样本，虽然能够使得数据平衡，但同时引入了噪声和无意义的样本，导致召回率依旧较低。

方法	Accuracy (%)	Specificity (%)	Sensitivity (%)	MCC
KNN 下采样	85.8	93.2	63.4	0.60
SMOTE 过采样	95.5	98.6	30.5	0.38

表 4. 不同平衡数据样本的方法比较

6 总结与展望

本文复现并分析了 CNN-Meth 模型在赖氨酸甲基化位点预测任务中的表现，从数据处理、特征提取到模型设计进行了全面探讨。通过整合多维特征并引入卷积神经网络，CNN-Meth 在敏感度和整体性能上显著优于传统方法。复现工作中发现数据平衡策略对模型性能的影响尤为关键，KNN 下采样有效提升了模型的敏感度，而 SMOTE 过采样则因噪声引入表现不佳，有待寻找能够充分利用数据并有效的数据平衡方法。

在此基础上，未来研究可以进一步探讨使用比 CNN 更复杂、能够更全面捕捉全局序列信息的深度学习模型是否能够提升预测性能。此外，PTMs 涉及多种修饰类型，如乙酰化、磷酸化等，本文提出的特征组合和模型方法是否同样适用于其他 PTMs 的预测值得进一步研究。这些探索不仅有助于深入理解 PTMs 的生物学机制，还为蛋白质功能预测和疾病相关研究提供了更广泛的应用前景。

参考文献

[1] Sabit Ahmed, Afrida Rahman, Md Al Mehedi Hasan, Julia Rahman, Md Khaled Ben Islam, and Shamim Ahmad. predml-site: predicting multiple lysine ptm sites with optimal feature representation and data imbalance minimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(6):3624–3634, 2021.

[2] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[3] Bing Bai, David Vanderwall, Yuxin Li, Xusheng Wang, Suresh Poudel, Hong Wang, Kaushik Kumar Dey, Ping-Chung Chen, Ka Yang, and Junmin Peng. Proteomic landscape of alzheimer' s disease: novel insights into pathogenesis and biomarker discovery. *Molecular neurodegeneration*, 16(1):55, 2021.

[4] Kamakoti P Bhat, H Ümit Kaniskan, Jian Jin, and Or Gozani. Epigenetics and beyond: targeting writers of protein lysine methylation to treat disease. *Nature Reviews Drug Discovery*, 20(4):265–286, 2021.

[5] Michael Bremang, Alessandro Cuomo, Anna Maria Agresta, Magdalena Stugiewicz, Valeria Spadotto, and Tiziana Bonaldi. Mass spectrometry-based identification and characterisation of lysine and arginine methylation in the human proteome. *Molecular bioSystems*, 9(9):2231–2247, 2013.

[6] Abdollah Dehzangi, Kuldip Paliwal, Alok Sharma, Omid Dehzangi, and Abdul Sattar. A combination of feature extraction methods with an ensemble of different classifiers for protein structural class prediction problem. *IEEE/ACM transactions on computational biology and bioinformatics*, 10(3):564–575, 2013.

[7] Abdollah Dehzangi, Alok Sharma, James Lyons, Kuldip K Paliwal, and Abdul Sattar. A mixture of physicochemical and evolutionary–based feature extraction approaches for protein fold recognition. *International journal of data mining and bioinformatics*, 11(1):115–138, 2015.

[8] Md Al Mehedi Hasan and Shamim Ahmad. mlysptmpred: Multiple lysine ptm site prediction using combination of svm with resolving data imbalance issue. *Natural Science*, 10(9):370–384, 2018.

[9] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[10] Zhe Ju, Jun-Zhe Cao, and Hong Gu. ilm-2l: A two-level predictor for identifying protein lysine methylation sites and their methylation degrees by incorporating k-gap amino acid pairs into chou s general pseaac. *Journal of Theoretical Biology*, 385:50–57, 2015.

[11] Shuichi Kawashima and Minoru Kanehisa. Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374, 2000.

[12] Tzong-Yi Lee, Cheng-Wei Chang, Cheng-Tzung Lu, Tzu-Hsiu Cheng, and Tzu-Hao Chang. Identification and characterization of lysine-methylated sites on histones and non-histone proteins. *Computational biology and chemistry*, 50:11–18, 2014.

[13] Zexian Liu, Jun Cao, Xinjiao Gao, Yanhong Zhou, Longping Wen, Xiangjiao Yang, Xuebiao Yao, Jian Ren, and Yu Xue. Cpla 1.0: an integrated database of protein lysine acetylation. *Nucleic acids research*, 39(suppl_1):D1029–D1034, 2011.

[14] Zexian Liu, Yongbo Wang, Tianshun Gao, Zhicheng Pan, Han Cheng, Qing Yang, Zhongyi Cheng, Anyuan Guo, Jian Ren, and Yu Xue. Cplm: a database of protein lysine modifications. *Nucleic acids research*, 42(D1):D531–D536, 2014.

[15] Shahin Ramazi and Javad Zahiri. Post-translational modifications in proteins: resources, tools and prediction methods. *Database*, 2021:baab012, 2021.

[16] Shao-Ping Shi, Jian-Ding Qiu, Xing-Yu Sun, Sheng-Bao Suo, Shu-Yun Huang, and Ru-Ping Liang. Plmla: prediction of lysine methylation and lysine acetylation by combining multiple features. *Molecular BioSystems*, 8(5):1520–1527, 2012.

[17] Austin Spadaro, Alok Sharma, and Iman Dehzangi. Predicting lysine methylation sites using a convolutional neural network. *Methods*, 226:127–132, 2024.

[18] Haodong Xu, Jiaqi Zhou, Shaofeng Lin, Wankun Deng, Ying Zhang, and Yu Xue. Plmd: an updated data resource of protein lysine modifications. *Journal of Genetics and Genomics*, 44(5):243–250, 2017.

[19] Yuedong Yang, Rhys Heffernan, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, and Yaoqi Zhou. Spider2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. *Prediction of protein secondary structure*, pages 55–63, 2017.

[20] Weizhi Zhang, Xiaodan Tan, Shaofeng Lin, Yujie Gou, Cheng Han, Chi Zhang, Wanshan Ning, Chenwei Wang, and Yu Xue. Cplm 4.0: an updated database with rich annotations for protein lysine modifications. *Nucleic Acids Research*, 50(D1):D451–D459, 2022.