基于数据宽度处理的药物性质分类预测神经网络模型

李 梦[1,2],应 豪[1,2]
1. 重庆工商大学 数学与统计学院,重庆 400067
2. 经济社会应用统计重庆市重点实验室,重庆 400067

摘 要:目的 针对常规数据处理导致分类预测精度不高等问题,提出 Optuna-MLP-LightGBM 组合模型用于抗癌候选药物的性质分类预测。**方法** 针对收集的 1 974 种化合物(每个化合物各 729 个分子描述符),首先利用多层感知机(MLP)对高维数据进行聚合处理,再采用跳转连接实现数据的宽度处理,将输出数据与输入数据合并组成宽度数据集,以此提高数据的特征识别,同时避免有用信息的缺失从而提高信息的流通;然后,用 LightGBM 替换 MLP 神经网络中的分类层,可以更好地进行分类处理及避免过拟合问题,最后构建基于 Optuna 优化的 MLP-LightGBM 分类预测模型,用于候选药物的小肠上皮细胞渗透性(Caco-2)的分类预测。**结果** 模型准确率、AUC 值和 F_1 值分别达到 91.03%、97.31 %和 90.48 %,由消融实验可以发现,通过 MLP-LightGBM 实现数据宽度处理以及分类后,模型分类效果相比 MLP 模型得到提升,3 种指标分别提升了 0.51%、1.22%和 0.7%;与逻辑回归(LR)、Attentive FP、MLP 等传统模型相比该模型能更好整合数据信息,其中与基模型相比平均增长幅度分别达到 5.94%、5.65%和 6.56%。**结论** 由于跳接处理使 MLP 网络可以达到特征的有效提取和扩充数据集的目的,同时引入机器学习可以更好地提高分类精度,因此在药物高通量筛选中可以成为重要的辅助工具。

关键词:MLP 神经网络;LightGBM;Optuna 自动化调参;抗癌候选药物;分类预测
中图分类号:TP389.1 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0006.012

Neural Network Model for Classification Prediction of Drug Properties Based on Data Width Processing

LI Meng[1,2], YING Hao[1,2]
1. School of Mathematics and Statistics, Chongqing Technology and Business University, Chongqing 400067, China
2. Chongqing Key Laboratory of Social Economic and Applied Statistics, Chongqing 400067, China

Abstract: Objective Aiming at problems such as low accuracy in classification prediction by conventional data processing, an Optuna-MLP-LightGBM combination model for predicting the properties of anticancer candidate drugs was proposed. **Methods** A total of 1 974 compounds (729 molecular descriptors for each compound) were collected. Firstly, a multi-layer perceptron (MLP) was used to aggregate high-dimensional data. A jump connection was used to realize the width processing of the data. The output data and input data were merged to form a width data set. This enhanced feature recognition and prevented the loss of useful information, thereby improving information flow. Then, LightGBM replaced the classification layer in the MLP neural network for better classification and to avoid overfitting issues. Finally, the

收稿日期:2023-03-05 **修回日期:**2023-08-26 **文章编号:**1672-058X(2025)06-0086-11
基金项目:重庆市自然科学基金面上项目(CSTC2020JCYJ-MSXMX0162)资助.
作者简介:李梦(1973—),女,四川开江人,副教授,硕导,博士,从事大数据分析及应用研究.
通信作者:应豪(1997—),男,重庆万州人,硕士研究生,从事大数据分析及应用研究. Email:1294434941@qq.com.
引用格式:李梦,应豪. 基于数据宽度处理的药物性质分类预测神经网络模型[J]. 重庆工商大学学报(自然科学版),2025,42(6):86-96.
LI Meng, YING Hao. Neural network model for classification prediction of drug properties based on data width processing[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(6): 86-96.

MLP-LightGBM classification prediction model based on Optuna optimization was constructed to predict the classification of the permeability of the small intestinal epithelial cells of the candidate drug (Caco-2). **Results** The accuracy, AUC, and F_1 values of the model reached 91.03%, 97.31 %, and 90.48 %, respectively. Through ablation experiments, it was found that the model' s classification performance has been improved compared with the MLP model after implementing data width processing and classification with MLP-LightGBM, with increases of 0.51%, 1.22%, and 0.7% in the three metrics, respectively. Compared with traditional models such as Logistic Regression (LR), Attentive FP, and MLP, this model can better integrate data information, with average growth rates compared with the base model of 5.94%, 5.65%, and 6.56%, respectively. **Conclusion** The jump-join processing enables the MLP network to effectively extract features and expand datasets. Introducing machine learning can further improve classification accuracy. Therefore, it can become an important auxiliary tool in high-throughput drug screening.

Keywords: MLP neural network; LightGBM; Optuna automatic parameter tuning; anticancer candidate drugs; classification prediction

1 引　言

乳腺癌是乳腺常见的恶性病变,表现为乳腺上皮细胞在多种致癌因子的作用下发生增殖失控的现象。2018 年国际癌症研究机构(IARC)最新的调查数据显示,乳腺癌在全球女性癌症当中发病率为 24.2%,位居女性癌症的首位,也常被称为“粉红杀手”。而乳腺癌患者有 52.9%发生在发展中国家,中国每年大约新增乳腺癌患者 42 万人,而且近年来年发病率每年递增 3%到 4%,因此抗乳腺癌的药物研发对乳腺癌的药物治理有着重要意义。

而在研发中的各个药物如何从众多药物当中脱颖而出成为候选药物,除了需要考虑其必须具备的良好生物活性以外,还需要此类药物在进入人体后能够具备良好的药物代谢动力学性质,该性质包括吸收(Absorption)、分布(Distribution)、代谢(Metabolism)和排泄(Excretion)以及毒性(Toxicity),合称为 ADMET 性质。这些性质由于不可预见,所以在药物研发过程中可能会出现研发失败的情况从而造成资源的浪费,在一般的情况下,药物研发失败的成本损失大约占总成本的 60%,因此在药物研发过程中,药物的 ADMET 性质分析与预测就显得尤为重要[1-3]。为此本文研究基于分类预测模型对药物 ADMET 性质中吸收性质(Absorption)进行分类预测。

近年来,传统药物筛选尽管衍生出了定向筛选、对特定样品的筛选、比较筛选以及随机筛选等方法,但仍有一定的局限性。例如对特定样品的筛选在已有信息条件下对特定的样品范围内进行筛选,这使其具有较高成功率,但其筛选的范围受到限制,忽略了广泛的资源,有时可能会造成低效药品的高投入研究。而近代科学技术的发展也带来了化学组合技术的新发展,新的候选药物呈几何级数的增长,药物发现数量已超过以往几十年的总和,当前方法或者模型难以满足高通量筛选的需要,新药的发现需要投入更多的成本,且在现实研究中面对的数据极度不规律。

由于机器学习以及深度学习方法在非线性的条件下能够有效捕捉到数据的内在联系或者规律,被用于各类数据集的预测[4-5],如:ZHANG RUIZE 等[6]提出一种基于多层感知器(MLP)的网络,根据卷积模块提取医学图像的局部特征和并行多个 MLP 来提取和分析特征图的全局特征信息,使图像分类过程更加高效,而且有效地减少了所需的数据量。CHEN CHENG 等[7]针对实验方法鉴定蛋白质-蛋白质相互作用(PPI)PPI 既耗时又昂贵,提出 LightGBM-PPI,即使用弹性网络选择最优特征子集并消除冗余特征,以 LightGBM 作为分类器预测 PPI,建立 LightGBM-PPI 模型,实验表明:使用机器学习方法预测 PPI 对于改善细胞生长和发育以及进一步了解细胞的生命活动非常有效。SHI TINGTING 等[8]提出分子二维图像的卷积神经网络(CNN)模型用于 ADMET 特性的预测,建立的 CNN 模型预测能力与基于人工结构描述和特征选择的现有机器学习模型相当,可以有效提取与分子 ADMET 特性相关的关键图像特征。Yang 等[9]融合了遗传算法(GA)的特征选择与共扼梯度(CG)的参数优化,所构建模型在四个数据集上的预测精度均显著高于单一 SVM 模型。虽然机器学习和深度学习能够提取非线性的特征,但相应的参数设置较为繁琐,特别是深度学习参数量过多时,不能更好地提取

特征。除此之外,基于卷积神经网络的特征提取虽然可以提取到不同层次的特征,但是其中的池化处理会涉及原始数据信息的丢失,可能会导致药物分子之间的区别不是特别明显。

除了传统数据集以外,部分学者还从分子化合物特有的分子图和分子指纹[10-12]的角度进行分析。XIONG ZHAOPING 等[13]提出了基于图形的神经网络模型(Attentive FP),Attentive FP 通过将节点从附近的节点传播到更远处的节点来表征原子的局部特征信息,同时还应用图注意机制来考虑分子内的非局部效应,在充分考虑分子特有结构的前提下,能够有效地捕获任何节点之间隐藏的关键链接信息,进而能够更加有效的根据分子信息进行预测研究。Pires 等[14]提出 pkCSM 方法,该方法采用分子描述符(亲脂性、分子量等)和药效团指纹(疏水性、族类、供体等)作为分子特征,用于预测新型多样化分子的各种 ADMET 性质,实验表明 pkCSM 的性能与目前可用的类似方法一样好或更好。顾耀文等[15]提出基于图注意力网络构建的 ADMET 预测模型,通过图注意力神经网络提取分子的结构特征,并将其与分子指纹合并作为数据集,用于药物分子的 AMEDT 性质预测。Iqbal 等[16]提出了基于图的深度学习模型。该模型将 MPNN 与类加权损失函数集成在一起,有效解决了 H2V 数据集中的类不平衡问题。Naveed 等[17]提出了一种深度学习模型,该模型通过卷积神经网络(CNN)、长短期记忆网络(LSTM)和变压器架构来提取特征并预测耐药性,与现有的耐药性预测模型相比所提模型有更好的一个预测效果。Lv 等[18]构建了一个三模态学习模型,该模型通过融合 SMILES 向量、ECFP 指纹和分子图信息,在精度和可靠性上实现了显著提升。分子指纹和分子图的特征虽然在一定层面上实现了数据集的扩充,但是由于其数据形式,在神经网络中并不能有效进行分类预测,与此同时图神经网络对于不同的数据集,需要构建相应的邻接矩阵数据信息,降低了其通用性;此外,与传统神经网络一样,其在处理小样本图数据时容易过拟合,特别是在图中节点数量较少的情况下。

虽然传统的深度学习和机器学习在特征提取方面和分类预测方面中取得显著效果。但采用单一深度学习或机器学习算法用于分类或预测研究,会出现数据清洗不完全和算法参数优化时迭代繁琐等问题,无法满足快速运算的同时找到最优解,因此将机器学习算法和深度学习算法融合并对参数进行优化对分类研究很重要。

为进一步提升在药物性质分类预测中的准确率以及相应指标数值,本文提出基于 Optuna 优化的 MLP-LightGBM 组合模型,将 MLP 与传统的分类模型 LightGBM 结合起来,在抗癌候选药物筛选工作中,建立此模型对候选药物的小肠上皮细胞渗透性(Caco-2)进行分类预测,辅助减少药物研发中的时间与资金投入。本文主要贡献如下:

(1) 设计了 Optuna-MLP-LightGBM 组合模型,用于候选药物的小肠上皮细胞渗透性(Caco-2)的分类预测;

(2) 在 MLP 网络的基础上引入跳转连接实现数据的宽度处理,即通过 MLP 实现对数据聚合处理,在此基础上采用 Connection 将聚合数据和原始数据合并到一起作为宽度数据集,保证特征的可重用性和解释性,避免有用信息的缺失从而提高信息的流通;

(3) 将宽度处理的 MLP 神经网络在输出层之前截断,将每个神经元输出数据整合成后续分类器需要的形式,用 LightGBM 模型替换 MLP 神经网络中分类层,实现小肠上皮细胞渗透性的分类预测,通过 MLP 和 LightGBM 的结合使用来提高数据的特征识别;

(4) 不同神经网络的超参数调优一直都受到超参数数量过多而资源消耗过多的问题,本文并未直接对 MLP 神经网络进行超参数调优,而是对 LightGBM 框架运用超参数调优 Optuna,以此实现在不消耗过多资源前提下,寻找最佳分类效果,实现结果显示了本文模型的有效性。

2 本文方法

本节将分为两个小节介绍提出的分类框架。首先展示总体框架,以展示其工作原理,如图 1 所示。

本文将 MLP 与 LightGBM 两种模型相结合,并采用 Optuna 对其进行相关参数优化。模型分为 3 个模块,分别为数据处理模块、分类模块和优化模块。在数据处理模块中,对数据集中 1 949 种化合物的 729 个分子描述符信息进行描述性统计分析与预处理;在分类模块中,将其分为两步,如图 1 所示,通过 MLP 模型对数据进一步处理,提升对容易信息的获取,同时采用跳转连接进行数据的宽度处理,在 MLP 处理得到的数据基础上采用 Connection 的方式引入原始数据信息,保证数据的可解释性以及可重用性,在第二步中,将宽度数据集输入到 LightGBM 模型中进行分类,即将 MLP 网络中的分类层更改为 LightGBM,发挥梯度提升树分解问题的优势;在优化模块中引入超参数调优框架 Optuna 为 LightGBM 模型寻找最优分类结果。下面的小节给出了分类模块和优化模块的详细描述。

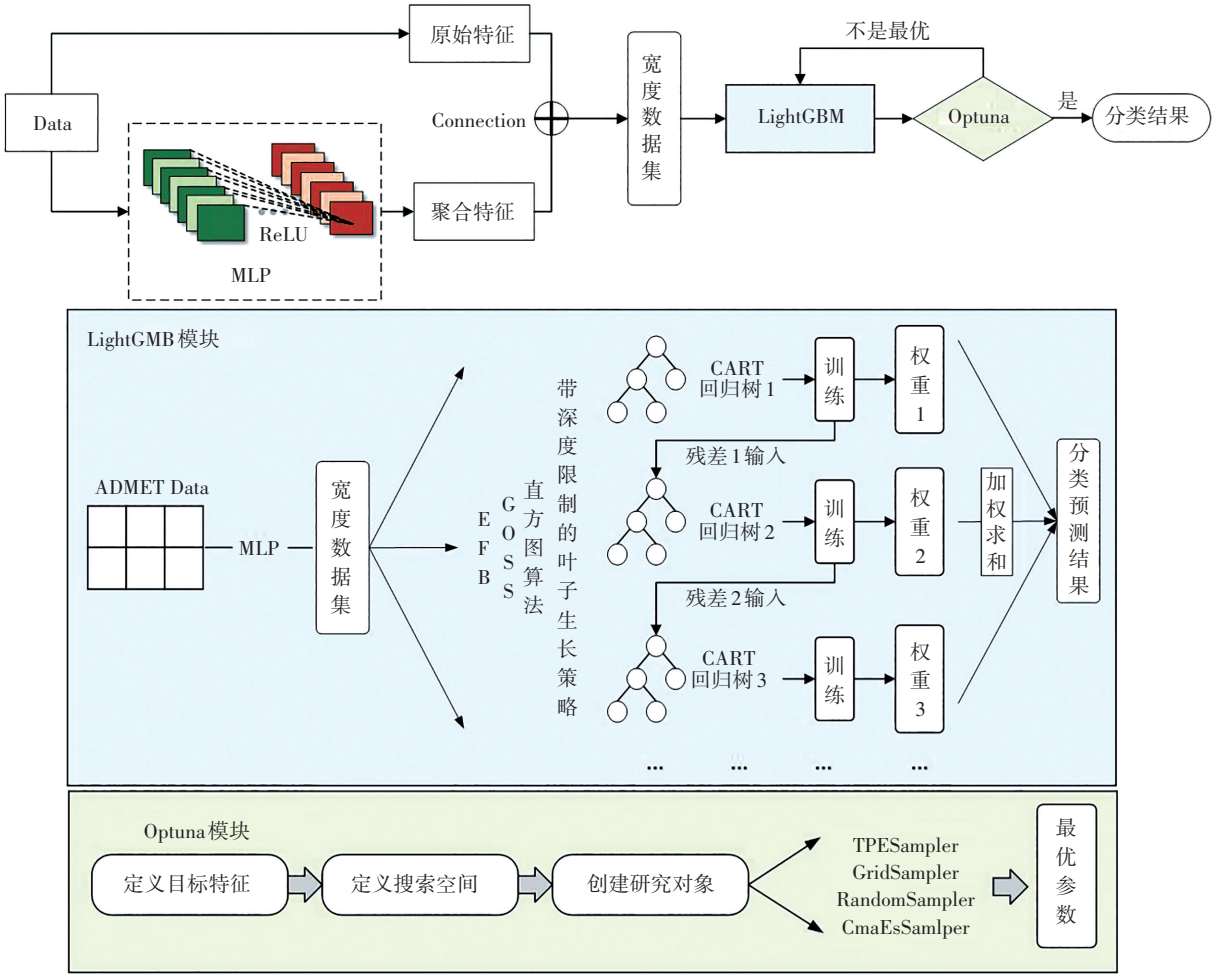

图 1　基于宽度数据处理的神经网络模型结构示意图

Fig. 1　A neural network model for width data processing

2.1　分类模块

该模块分为两步进行,如图 2 所示,首先将预处理过后的 ADMET 数据输入到多层感知机(MLP)中,通过 MLP 对数据做进一步整合后得到相应的输出结果,其次把输出结果作为样本特征加载到监督学习器 LightGBM,通过 LightGBM 对数据做分类处理。

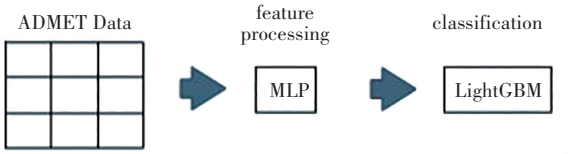

图 2　分类模块流程

Fig. 2　Process of the classification module

2.1.1　基于跳转的数据宽度处理

为了提升对抗乳腺癌数据特征的识别度,本文采用多层感知机(Multi-Layer Perceptron,MLP)对标准化的特征数数据做聚合处理。MLP 由感知机(Perceptron Learning Algorithm,PLA)推广而来,其主要的特点便是拥有多个神经元层,也被称为人工神经网络(Artificial Neural Network,ANN)。典型的 MLP 结构包括 3 层:输入层、隐藏层以及输出层,考虑到隐藏层个数过多虽然可以降低误差,提高精度,但是也使网络复杂化,从而增加模型训练时间和出现“过拟合”的倾向,本文设置两层隐藏层,通过修改 MLP 中神经元个数,实现对数据的进一步处理。

多层感知机层与层之间是全连接的(即上一层的某一个神经元与下一层所有神经元都有连接)。其中输入层(input)的节点表示数据的输入,其他层的节点则通过将输入层输入的数据与层上节点的权重 W 和偏置 b 进行线性组合后在对其应用激活函数,采用的激活函数有

(1) Sigmoid 函数

$$\sigma(x)=\frac{1}{1+e^{-x}}$$

(2) ReLU 函数

$$f(x)=\max(0,x)$$

当得到输出层的输出结果后,对于参数 W 还需要进行反馈传播进行优化,这与 BP 神经网络优化一致,通过计算得到损失值,再通过链式法则求得各个权重的导数后去更新权重 ,其中需要注意的便是求导同时需要对运用的激活函数进行求导,最终得到优化过后的模型:

$$h=(b^{(1)}+\boldsymbol{W}^{(1)}x)$$

$$o=b^{(2)}+\boldsymbol{W}^{(2)}(s(b^{(1)}+\boldsymbol{W}^{(1)}x))$$

$$f(x)=g(b^{(2)}+\boldsymbol{W}^{(2)}(s(b^{(1)}+\boldsymbol{W}^{(1)}x)))$$

其中,x 为输入数据,$\boldsymbol{W}^{(1)}$ 为输入层到隐藏层的权重矩阵, $b^{(1)}$ 为相应的偏置,s 为所选取的激活函数,相应的 $\boldsymbol{W}^{(2)}$ 为隐藏层到输出层的权重矩阵,$b^{(2)}$ 为相应的偏置,o 为输出层输出, g 为 Softmax 回归,$f(x)$ 即为最后的输出值。

本文通过设置隐藏层中的神经元个数,以此达到对数据的进一步处理,聚合处理数据如表 1 所示。

表 1 多层感知机(MLP)部分输出值

Table 1 Partial output values of MLP

特征 1	特征 2	特征 3	…	特征 63	特征 64
54.47	0	0	…	38.91	0
55.54	0	0.78	…	39.08	0
62.92	0	0	…	34.59	12.59
42.34	0	11.36	…	39.51	5.32
…	…	…	…	…	…

在聚合数据的基础上,引入跳转连接实现数据的宽度处理。

2.1.2 LightGBM

考虑到监督学习器 LightGBM 是轻量级(Light)、快速的、高性能的梯度提升机器(GBM),本文采用 LightGBM 对抗乳腺癌数据做分类研究。在第一层工作基础上用 LightGBM 替换 MLP 神经网络中的分类层,发挥传统机器学习分解问题的优势。

传统的 boosting 算法(例如 GBDT)在常规数据集中有较好效果,但对于大样本以及高维数据,由于每一个特征都需要进行相应的计算,以此来找到最好的切分节点,传统的 boosting 算法在处理该类数据时其能力明显变弱,主要体现在模型内存消耗增加导致其训练速度变慢。

事实上候选药物的性质研究包含药物数量和药物的各种化学信息,导致其数据具有大样本和高维度等特点,为此本文采用 LightGBM 对数据进行分类研究,具体工作流程如图 1 所示。

LightGBM 是 GBDT 模型的进化版本,它延续 XGBoost 的集成学习方式,但相对于 XGBoost,其通过直方图算法(Histogram)解决候选分类点数量过多的问题,可通过单边梯度抽样算法(GOSS)解决样本数量过多的问题,通过互斥特征捆绑算法(EFB)解决特征数量过多的问题,因此采用 LightGBM 做抗乳腺癌大数据分类研究,可支持类别特征做训练以及特征的并行计算,可提高模型训练速度、节省内存消耗。

因此将其代替 MLP 神经网络中传统 Softmax 分类层,可以明显提高抗乳腺癌候选药物数据分类精度。

2.2 优化模块

在训练期间的机器学习或深度学习模型会学习到很多参数,不同参数值对模型性能的影响程度不同,因此对模型参数进行训练有助于训练出更好的模型。传统神经网络优化存在训练时间过长等问题,而 Optuna 是一个超参数的优化工具,它能够设置条件超参数,例如许多超参数只有在与其他超参数组合使用时才达到预期效果,但使用 Optuna 可以对单个超参数进行优化,即单独改变某个参数时也会产生预期效果。Optuna 具体工作流程如图 1 所示。

同时相比传统网格搜索(Grid Search),Optuna 可以快速搜索大空间并更快地修剪没有希望的试验以获得更好和更快的结果。它默认使用被称为 TPE Sampler(Tree-structured Parzen Estimator)的方法,这种方法依靠贝叶斯概率来确定哪些超参数选择是最有希望的并迭代调整搜索。在参数采样方式中,它还可以使用网格搜索(Grid Sampler)将每个超参数的搜索空间离散化,对每个超参数分别进行学习,并在最后选择最佳组合,使用随机搜索(Random Sampler)对搜索空间进行随机采样,直到满足停止条件为止,使用进化算法(CmaEs Sampler),即通过适应度函数值来寻找最优超参数。此外,还拥有轻松并行化和超参数搜索等优点。

因此本文采用 Optuna 对监督学习器 LightGBM 参数进行优化以提升分类精度。

3 仿真实验与结果分析

3.1 数据预处理

3.1.1 数据来源

虽然乳腺癌治疗中最重要的目标基因是 ERα,但是化合物相应的 ADMET 性质也是不可忽略的。本文数据

选自 2021 年华为杯数学建模竞赛(网址为 https://cpipc. acge. org. cn//cw/detail/4/2c9080147c73b890017c7779e57e07d2),数据集包含 1 949 个化合物及每个分子化合物的 729 个分子描述符信息,部分数据如表 2 所示,其中第一列是 1 949 个化合物各自的表达式 SMILES(Simplified Molecular Input Line Entry System),最后一列为每个化合物相应的小肠上皮细胞渗透性(Caco-2),数值由二分类法生成的相应数值,其中“1”表示该化合物在小肠上皮细胞上渗透性较好,“0”表示该化合物在小肠上皮细胞上渗透性较差。

表 2　数据集

Table 2　Dataset

SMILES	nAcid	ALogP	…	Zagreb	Caco-2
Oc1c…cc4	0	-0. 286	…	166	0
Oc1ccc2O…cc4	0	-0. 862	…	174	0
Oc1ccc…cc4	0	0. 729 6	…	176	0
Oc1ccc2O…cc4	0	-0. 318 4	…	174	0
…	…	…	…	…	…

在分类问题当中样本数据是否均衡是首要关注点。抗癌候选药物分子结构相似时可能存在相同性质,也可能存在不同性质,进而出现多个不同分子是同一类别的情况,因此需要对数据进行均衡分析。一般当样本比例是 1∶2 至 1∶10 时将其称为轻微不均衡,可以不进行样本不均衡处理,而比例若是超过 1∶10,就必须进行样本不均衡处理,以避免因为极度不均衡样本导致预测出错。

本文选取的类别样本数据占比如图 3 所示,可以发现两个类别样本占比为 4∶6,为轻微不均衡,因此不需要进行样本不均衡处理。

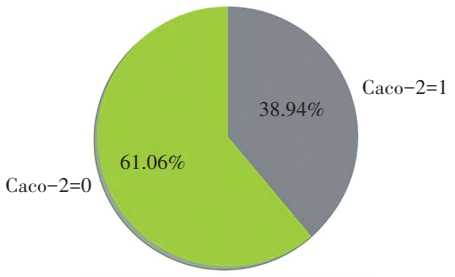

图 3　类别样本占比

Fig. 3　Percentage of category samples

3. 1. 2　数据统计分析

对找到的数据需要对其进行统计分析来描述数据各个特征的具体情况,其中包括数据的频数分析,即在预处理部分通过频数分析检验异常值;数据的集中趋势分析,即通过平均数、中位数、众数等指标来反映数据的一般水平。描述性统计分析部分数值如表 3 所示。

表 3 说明不同特征的分子描述符有较大的统计差距,因此在进行分析预测之前,需要对数据进行归一化处理,为此采用 min-max 标准化对整个数据集进行归一化处理,一方面提升模型的收敛速度,另一方面提高模型的精确度。

表 3　描述性统计分析数值

Table 3　Descriptive statistical analysis values

SMILES	nAcid	ALogP	…	XLogP	Zagreb
count	1 949. 00	1 949. 00	…	1 949. 00	1 949. 00
mean	0. 11	1. 11	…	2. 96	149. 73
std	0. 35	1. 43	…	1. 62	40. 74
min	0. 00	-23. 11	…	-3. 59	62. 00
25%	0. 00	0. 35	…	1. 93	116. 00
50%	0. 00	1. 15	…	2. 82	146. 00
75%	0. 00	1. 93	…	3. 67	180. 00
max	4. 00	5. 18	…	14. 28	748. 00

3. 1. 3　归一化处理

在正式进行建模分析之前,对于收集到的数据,由于各个指标之间差异过大影响模型性能,因此对数据进行归一化处理,归一化的方法选择常用的标准化,它是对原始数据进行相应的线性变换后使其落在[0,1]的区间当中,转换函数如下:

$$\hat{x}=\frac{x-\min}{\max-\min}$$

其中,max 为某一指标的最大值, min 则为这一指标的最小值。

3. 2　实验环境

在本节,首先简要描述本文中使用的数据集,然后,基于数据集评估所提出的框架的性能,并将其分别与传统模型进行比较,以证明所提出方法的优势。此外,还将所提出的方法应用于另外两个数据集中,以验证方法的泛用性。所提出组合模型实验是在一个环境中实现的,CPU 为 Intel(R) Core(TM) i5-9300H CPU @2. 40GHz,内存为 16 GB,代码使用基于 Python 平台的 Python3. 8 软件编写。

3.3 评价指标

模型的优劣需要相应评价指标来衡量,而对于二分类模型的评价指标,一般除了模型自带的预测得分(score)以外,还有 ROC 曲线等等,因此模型选取准确率(Accuracy)、AUC(Area Under Curve)值、召回率(Recall)、精确率(Precision)和 F1 为判定标准。

对于一个二分类问题,分别将正样本归为 T,负样本归为 F,而样本的真实类别和预测类别划分为 4 类,并构建如表 4 所示的混淆矩阵。

表 4 混淆矩阵

Table 4 Confusion matrix

	预测为正	预测为负
实际为正	True Positive(TP)	False Negative(FN)
实际为负	False Positive(FP)	True Negative(TN)

其中,真正例(True Positive,TP):也即真实的样本类别为正样本,而预测的类别也为正样本的情况;假正例(False Positive,FP):当真实的样本类别为负例,而预测的类别为正例的情况;假负例(False Negative,FN):当真实的样本类别为正例,而预测的类别为负例的情况;真负例(True Negative,TN):当真实的类别为负例,而预测的类别为负例的情况。

基于上面 4 个指标,就可以得到构建 ROC 曲线所需的指标 True Positive Rate(R_{TP})和 False Positive Rate(R_{FP}),计算公式如下:

$$R_{TP}=\frac{N_{TP}}{(N_{TP}+N_{FN})}$$

$$R_{FP}=\frac{N_{FP}}{(N_{FP}+N_{FN})}$$

ROC(Receiver Operating Characteristic)曲线,又被称为受试者工作特征曲线,其经常被用于二分类问题的模型度量当中,其原因在于它对于正负样本比例不敏感,即无视样本不平衡,ROC 曲线有两个指标构成,分别为 TPR(真正例率)与 FPR(假正例率),其中纵坐标为 TPR,横坐标为 FPR,ROC 曲线也是通过遍历所有设置的阈值来绘制整条曲线的,即 ROC 曲线上的每一个点,都对应着特定预测阈值下的一对 FPR 和 TPR。

准确率(f_{Acc}),描述分类器的分类准确率,相应公式如下:

$$f_{Acc}=\frac{(N_{TP}+N_{TN})}{(N_{TP}+N_{TN}+N_{FP}+N_{FN})}$$

S_{AUC} 为 ROC 曲线与 X 轴所围成的面积,由 ROC 曲线的构成,可以得知 S_{AUC} 数值越接近 1 表示分类器越好。

召回率(f_{Rec}),它描述的是当实际为正的样本被预测为正样本的概率,相应公式如下:

$$f_{Rec}=\frac{N_{TP}}{(N_{TP}+N_{FN})}$$

精确率(f_{Pre}),描述得是预测为正的样本中有多少是真正的正样本,相应公式如下:

$$f_{Pre}=\frac{N_{TP}}{(N_{TP}+N_{FP})}$$

F_1 评价指标由精确率(Precision)和召回率(Recall)两个指标构成,它能够反映模型的两部分性能,即模型的准确性和完整性,相应公式如下:

$$F_1=2*\frac{(f_{Pre}*f_{Rec})}{(f_{Pre}+f_{Rec})}$$

3.4 仿真实验结果与分析

3.4.1 消融实验比较

在多层感知机中,将数据分别输入两次,一次将 1 949 个原始数据按照 8∶2 的比例划分为训练集和测试集,其中训练集用于训练模型以及相关参数,测试集用于测试模型效果以此判断模型优劣。而损失函数设定为分类问题中传统的二元交叉熵(Binary cross entropy),其计算公式如下:

$$L_{Loss}=-\frac{1}{N}\sum_{i=1}^{N}[y_i\cdot\log p(y_i)+(1-y_i)\cdot\log(1-p(y_i))]$$

其中,y 是二元标签 0 或者 1,$p(y_i)$ 是输出属于 y_i 标签的概率。

为了更清楚的研究各个模块对于分类性能的影响,做了如下的消融实验:

(1) 通过多层感知机(MLP)直接对数据进行分类;

(2) 通过 LightGBM 直接对数据集进行分类;

(3) 在第 1 个基础上,通过多层感知机(MLP)对数据进行宽度处理后,再由 LightGBM 对数据进行分类;

(4) 在第 3 个基础上,使用 Optuna 调优框架对分类层的 LightGBM 参数进行优化;

相应结果如图 4 和表 5 所示。

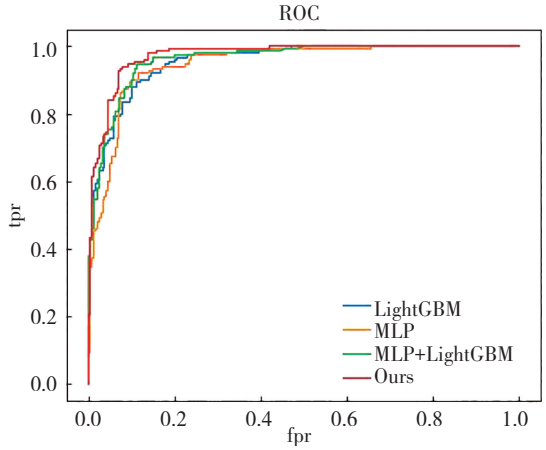

图 4　不同模块消融分析结果可视化

Fig. 4　Visualization of ablation analysis results of different modules

表 5　不同模块消融分析结果

Table 5　Results of ablation analysis of different modules　%

	f_{Acc}	S_{AUC}	f_{Rec}	f_{Pre}	F_1
MLP	87.95	94.96	98.00	87.73	87.76
LightGBM	88.46	95.59	82.67	88.09	87.70
MLP+ LightGBM	88.46	96.18	80.67	88.43	87.60
Ours	91.03	97.31	87.33	90.65	90.48

如图 4 和表 5 发现,如果 MLP 的输出层替换为监督学习器 LightGBM,则相关结果会得到改善。AUC 的变化较为明显,MLP 的分类效果 AUC 值仅为 94.96%,替换为 LightGBM 过后 AUC 值提高到 96.18%(提升了 1.22%)。其余 4 个指标也有相应变化,Accuracy 从原本的 87.95%(MLP 模型)提高到 88.46%(MLP + LightGBM 模型),但是对于分类问题,指标仍然以 AUC 指标与 F_1 指标主,Recall 从原本的 98.00%(MLP 模型)降低到 80.67%(降低了 17.33%);Precision 从原本的 87.73%提高到 88.43%(提高了 0.7%),F_1 从原本的 87.76%降低到 87.60%(降低了 0.16%)。3 个指标中 Accutacy 和 AUC 值都有一定提升。同时相比 LightGBM 的分类效果,AUC 值的提升幅度(0.59%)明显大于 F1 的降低幅度(0.1%)。

而在此基础上加入优化模块(Optuna)后模型性能得到了更明显的提升,AUC 指标达到了 97.31%,相比之前的 96.18%,提升了 1.13%;Accuracy 指标相比(MLP + LightGBM 模型)提高了 2.57%;Recall 相比(MLP+LightGBM 模型)提高了 6.66%,Precision 相比(MLP+LightGBM 模型)提高了 2.22%;F_1 相比(MLP+LightGBM 模型)提高了 2.88%。

使用 MLP 对数据进行特征聚合的同时采取宽度处理能够保证不丢失数据的各种特征信息,因此上述结果表明:本文提出的模型在数据处理方面提高分类效果方面具有一定优势。

3.4.2　与传统模型的对比

在传统的分类问题中,对于高维数据的处理一般都偏向于删去其中不重要的特征后再进行分类,但是本文模型选择对数据进行合并,将模型与其他传统模型进行了比较,其中包括 LR 模型、GBDT 模型、XGBoost 模型、LightgGBM 模型、MLP 模型、SVM 模型、RF 模型以及 Attentive FP 模型。

上述模型分类效果如图 5 和表 6 所示。可以观察到,在 Accuracy、AUC 值、Recall、Precision 和 F_1 这 5 个指标中,当传统的机器学习算法准确率在 84.00% ~ 88.50%范围内波动时,本文模型的准确率达到了 91.03%,在另外 4 个指标中,本文模型也分别达到了 97.31%、87.33%、90.65%和 90.48%。

与此同时以传统的 LR 模型作为基础模型,找到了每个模型相应的拟合效果增长幅度,当基础模型 LR 的分类效果 Accuracy、AUC 值和 F_1 值分别为 84.10%、91.52% 以及 83.03% 时,相比于 GBDT、XGBoost、LightGBM、MLP、SVM、RF、Attentive FP 的提升效果,本文模型提升效果较为明显,在 Accuracy 指标中的提升效果为 6.93%;在 AUC 指标中的提升效果为 5.79%;在 Recall 指标中的提升效果为 10.00%;在 Precision 指标中的提升效果为 7.28%;在 F_1 指标中的提升效果为 7.40%,模型的提升效果十分明显。

从每一个指标数值中可以看出,本文模型充分识别了数据集中的信息,不管是模型效果还是指标数值的增长情况,模型都有很好的效果,这进一步证明了模型的优越能力。

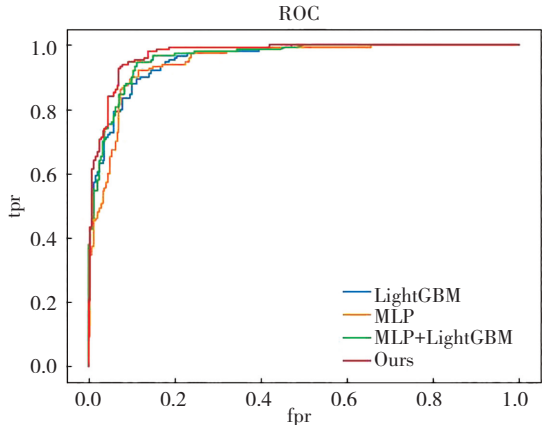

图 5　各模型指标对比分析可视图

Fig. 5　Visual diagram of comparative analysis of each model index

表 6 各模型指标具体数值

Table 6 Specific values for each model indicator %

	f_{Acc}	S_{AUC}	f_{Rec}	f_{Pre}	F_1
LR	84.10	91.52	77.33	83.37	83.08
GBDT	85.13	92.03	80.67	84.29	84.29
XGBoost	86.41	93.99	81.33	85.74	85.59
SVM	82.31	87.63	87.33	81.64	81.89
RF	87.44	94.55	85.79	87.43	86.45
LightGBM	88.46	95.59	82.67	88.09	87.70
MLP	87.95	94.96	98.00	87.73	87.76
AttentiveFP	87.18	95.14	78.67	87.08	86.20
MLP+ LightGBM	88.46	96.18	80.67	88.43	87.60
Ours	91.03	97.31	87.33	90.65	90.48

3.4.3 泛用性分析

为了探究方法是否有一定局限性,选择另外两组数据集进行分类研究,选取的仍然是 2021 年华为杯数学建模竞赛(网址为 https://cpipc.acge.org.cn//cw/detail/4/2c9080147c73b890017c7779e57e07d2)的数据集,数据集具体信息如表 7 所示。实验设置与之前一致,相应消融分析结果和各模型对分分析结果如表 8 和表 9 所示,相应 ROC 曲线如图 6 和图 7 所示。

表 7 数据集信息

Table 7 Dataset information

数据类型	ADMET 属性	数据描述	样本量	阳性样本量	阴性样本量
代谢	CYP3A4	细胞色素 P450 酶(化合物的代谢稳定性)	1 974	1 461	513
毒性	hERG	化合物心脏毒性	1 974	875	1 099

表 8 不同模块消融分析结果

Table 8 Results of ablation analysis with different modules %

	CYP3A4					*hERG*				
	f_{Acc}	S_{AUC}	f_{Rec}	f_{Pre}	F_1	f_{Acc}	S_{AUC}	f_{Rec}	f_{Pre}	F_1
MLP	91.39	97.17	94.15	89.04	89.04	85.82	95.31	96.57	87.93	84.61
LightGBM	92.66	97.67	96.19	91.40	90.45	90.38	96.01	92.27	90.12	90.04
MLP+ LightGBM	92.66	96.45	95.50	90.92	90.57	90.63	96.16	92.27	90.35	90.31
Ours	94.18	97.94	96.19	92.69	92.56	91.65	97.22	92.70	91.34	91.37

表 9 各模型指标具体数值

Table 9 Specific values for each model indicator %

	CYP3A4					*hERG*				
	f_{Acc}	S_{AUC}	f_{Rec}	f_{Pre}	F_1	f_{Acc}	S_{AUC}	f_{Rec}	f_{Pre}	F_1
LR	90.13	95.69	93.08	87.36	87.47	84.81	88.30	88.84	84.50	84.18
GBDT	89.62	94.79	96.19	88.81	85.94	87.59	94.58	93.56	88.01	86.92
XGBoost	92.15	96.63	94.46	89.92	90.04	89.37	95.81	92.70	89.29	88.92
SVM	85.57	92.35	92.04	82.23	80.98	82.03	89.00	81.55	81.43	81.66
RF	90.38	94.56	93.08	87.60	87.82	88.35	95.83	89.27	87.88	88.01
LightGBM	92.66	97.67	96.19	91.40	90.45	90.38	96.01	92.27	90.12	90.04
MLP	91.39	97.17	94.15	89.04	89.04	85.82	95.31	96.57	87.93	84.61
AttentiveFP	89.62	97.23	88.24	85.88	87.70	86.58	95.02	92.70	86.91	85.85
MLP+LightGBM	92.66	96.45	95.50	90.92	90.57	90.63	96.16	92.27	90.35	90.31
Ours	94.18	97.94	96.19	92.69	92.56	91.65	97.22	92.70	91.34	91.37

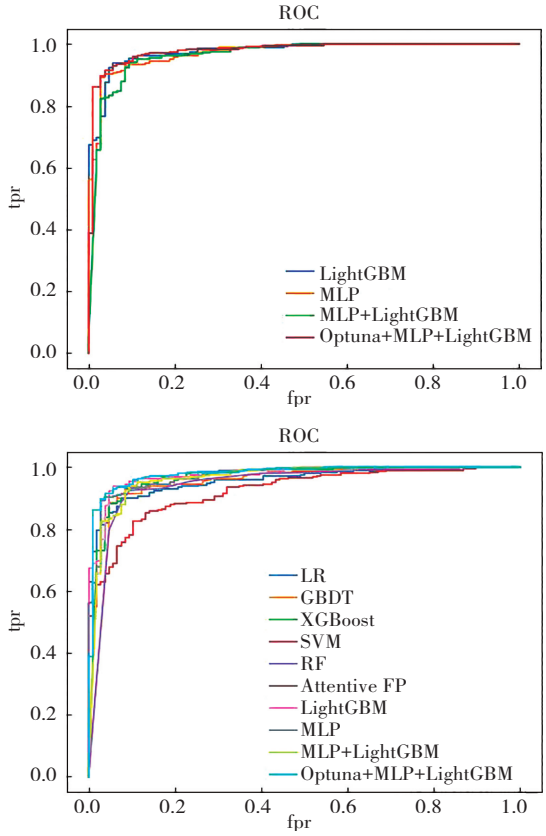

图 6　CYP3A4 数据集 ROC 曲线

Fig. 6　ROC curves for the CYP3A4 dataset

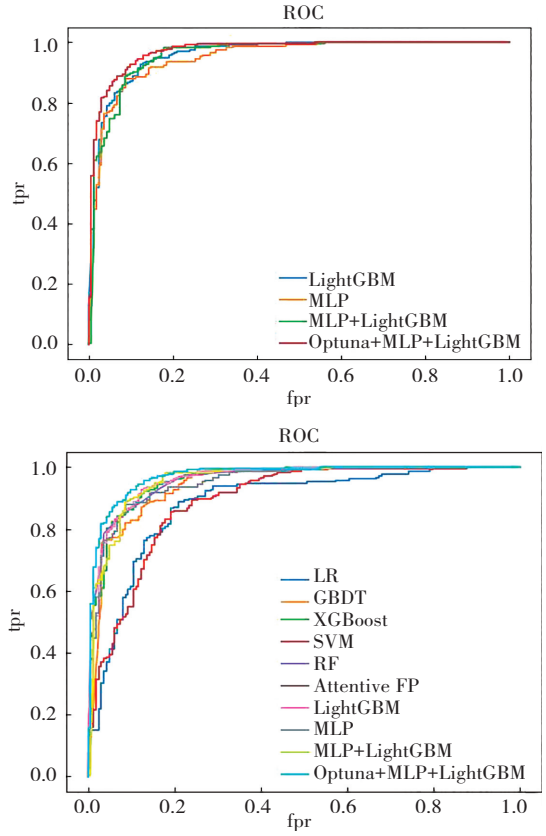

图 7　hERG 数据集 ROC 曲线

Fig. 7　ROC curves for the hERG dataset

在 CYP3A4 数据集中,跳接处理以及机器学习器的引入使得模型分类效果有进一步提升,Accuracy 从原本的 91.39%(MLP 模型)提高到 92.66%(MLP + LightGBM 模型);AUC 值从原本的 97.17%降低到 96.45%;F_1 从原本的 89.04%提高到 90.57%(3 个指标中 Accutacy 和 F_1 值都得一定提升)。模型整体性能仍然是有一定提高。而在此基础上加入优化模块(Optuna)后模型性能得到了更明显的提升。AUC 值达到了 97.94%(提高了 1.49%),Accuracy 和 F_1 值则分别提升了 1.52%和 1.99%。此外,相比基模型(LR),Accuracy、AUC 值、F_1 值分别提升了 4.05%、2.25%和 5.09%。

在 hERG 数据集中,Accuracy 从原本的 85.82%(MLP 模型)提高到 90.63%(MLP+LightGBM 模型);AUC 从原本的 95.31%提高到 96.16%;F_1 从原本的 84.61%提高到 90.31%(3 个指标都得一定提升)。而在此基础上加入优化模块(Optuna)后模型性能得到了更明显的提升。AUC 值达到了 97.22%(提高了 1.06%),Accuracy 和 F_1 值则分别提升了 1.02%和 1.06%。此外,相比基模型(LR),Accuracy、AUC 值、F_1 值分别提升了 6.84%、8.92%和 7.19%。

在另外两种数据集的实验中,可以发现模型的分类效果仍然是最好的,表明模型具有一定的泛用性,可以在其余数据集发挥很好的分类性能。

4　结　论

本文以抗乳腺癌候选药物的分子描述符为数据集,研究抗乳腺癌候选药物的性质,提出新的基于 Optuna 优化的 MLP-LightGBM 分类预测模型。该模型利用 MLP 实现特征信息地有效提取,同时采用宽度处理保证特征信息的可重用性和解释性,利用 MLP-LightGBM 框架对数据进行分类处理,发挥深度学习和传统机器学习的各自优势,最后引入 Optuna 框架寻找最优分类效果。与传统的 LR 模型、GBDT 模型、XGBoost 模型、LightgGBM 模型、MLP 模型、SVM 模型、RF 模型和 Attentive FP 模型相比,本文模型更好整合到数据中的关键信息。实验表明:模型在所搜集数据集中有更好的分类效果,Accuracy 相比基模型平均提高了 5.94%,AUC 值平均提高了 5.65%,F_1 指标平均提高了 6.56%,提升效果显著。因此,有理由相信集成模型可以为自动化虚拟筛选和药物设计提供一种

更有效的方法。

在分类以及预测研究中,仅仅采用分子描述符作为分子的特征描述还有一定的局限性,主要是因为人工提取的分子描述符可能存在信息丢失等情况。到目前为止,已有一些文章开始关注分子结构的其他特征,如分子指纹、分子图等特征,并有一些文章对存在"活性悬崖"时的预测问题以及神经网络"黑匣子"属性可视化方面展开研究。因此,未来研究将集中在化合物分子多属性特征提取、筛选、融合以及相应解释上。

参考文献(References):

[1] WICHMANN K, DIEDENHOFEN M, KLAMT A. Prediction of blood-β rain partitioning and human serum albumin binding based on COSMO-RS σ-moments[J]. Journal of Chemical Information and Modeling, 2007, 47(1): 228-233.

[2] DAVIS A M, RILEY R J. Predictive ADMET studies, the challenges and the opportunities[J]. Current Opinion in Chemical Biology, 2004, 8(4): 378-386.

[3] VAN D E WATERBEEMD H, GIFFORD E. ADMET in silico modelling: Towards prediction paradise?[J]. Nature Reviews Drug Discovery, 2003, 2(3): 192-204.

[4] FRÖHLICH H, WEGNER J , SIEKER F, et al. Kernel functions for attributed molecular graphs-A new similarity-based approach to ADME prediction in classification and regression[J]. QSAR & Combinatorial Science, 2006, 25(4): 317-326.

[5] CUCOS A M, IANTOVICS L B. Comparative study of random forest, gradient boosted trees, feedforward neural networks and convolutional neural networks using fingerprints and molecular descriptors for adverse drug reaction prediction[J]. Procedia Computer Science, 2024, 246: 1895-1904.

[6] ZHANG R, WANG L, CHENG S, et al. MLP-based classification of COVID-19 and skin diseases[J]. Expert Systems with Applications, 2023, 228: 120389.

[7] CHEN C, ZHANG Q, MA Q, et al. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion[J]. Chemometrics and Intelligent Laboratory Systems, 2019, 191: 54-64.

[8] SHI T, YANG Y, HUANG S, et al. Molecular image-based convolutional neural network for the prediction of ADMET properties[J]. Chemometrics and Intelligent Laboratory Systems, 2019, 194: 103853.

[9] YANG S Y, HUANG Q, LI L L, et al. An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs[J]. Artificial Intelligence in Medicine, 2009, 46(2): 155-163.

[10] WANG J, ZHANG L, SUN J, et al. Predicting drug-induced liver injury using graph attention mechanism and molecular fingerprints[J]. Methods, 2024, 221: 18-26.

[11] CERETO-MASSAGUÉ A, OJEDA M J, VALLS C, et al. Molecular fingerprint similarity search in virtual screening[J]. Methods, 2015, 71: 58-63.

[12] FENG H, ZHANG L, LI S, et al. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints[J]. Toxicology Letters, 2021, 340: 4-14.

[13] XIONG Z, WANG D, LIU X, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism[J]. Journal of Medicinal Chemistry, 2020, 63(16): 8749-8760.

[14] PIRES D E V, BLUNDELL T L, ASCHER D B. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures[J]. Journal of Medicinal Chemistry, 2015, 58(9): 4066-4072.

[15] 顾耀文, 张博文, 郑思, 等. 基于图注意力网络的药物ADMET分类预测模型构建方法[J]. 数据分析与知识发现, 2021, 5(8): 76-85.
GU Yao-wen, ZHANG Bo-wen, ZHENG Si, et al. Predicting drug ADMET properties based on graph attention network[J]. Data Analysis and Knowledge Discovery, 2021, 5(8): 76-85.

[16] BASIT IQBAL A, ASSAD A, BHAT B, et al. MPNN-CWExplainer: an enhanced deep learning framework for HIV drug bioactivity prediction with class-weighted loss and explainability[J]. Life Sciences, 2025, 378: 123835.

[17] NAVEED S, HUSNAIN M, ALSUBAIE N. HybridDLDR: a hybrid deep learning-based drug resistance prediction system of Glioblastoma (GBM) using molecular descriptors and gene expression data[J]. Computer Methods and Programs in Biomedicine, 2025, 270: 108913.

[18] LU X, XIE L, XU L, et al. Multimodal fused deep learning for drug property prediction: Integrating chemical language and molecular graph[J]. Computational and Structural Biotechnology Journal, 2024, 23: 1666-1679.

责任编辑:陈 芳