Modeling and Simulation 建模与仿真, 2023, 12(1), 202-211
Published Online January 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.121020

Hans汉斯

抗乳腺癌候选药物的ADMET性质的预测研究

钱祖建

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2022年11月28日；录用日期：2023年1月12日；发布日期：2023年1月19日

摘　要

近年来，乳腺癌已经成为全世界范围内女性患病率和死亡率非常高的恶性肿瘤，研究与制作抗乳腺癌药物已经迫在眉睫。在此背景下，本文主要研究了能够拮抗ERα活性的抗乳腺癌候选药物的ADMET (吸收Absorption、分布Distribution、代谢Metabolism、排泄Excretion和毒性Toxicity)性质的预测模型，对临床试验得到的1974个化合物的ADMET数据进行预处理和相关分析。运用BP神经网络和XGBoost回归两种方法建立并研究了两种对化合物ADMET性质的定量预测模型。实验研究结果表示，对比于BP神经网络方法，XGBoost分类预测模型对于该任务误差最低、效果最好。

关键词

乳腺癌药物，BP神经网络，XGBoot分类预测模型

Predictive Study of the ADMET Properties of Anti-Breast Cancer Drug Candidates

Zujian Qian

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Nov. 28th, 2022; accepted: Jan. 12th, 2023; published: Jan. 19th, 2023

Abstract

In recent years, breast cancer has become a malignancy with a very high prevalence and mortality rate in women worldwide, and the research and production of anti-breast cancer drugs has become urgent. In this context, this paper focuses on the prediction model of ADMET (Absorption Absorption, Distribution Distribution, Metabolism Metabolism, Excretion Excretion and Toxicity) properties of anti-breast cancer drug candidates capable of antagonizing ERα activity, for 1974 compounds obtained from clinical trials. The ADMET data were preprocessed and correlated. Two quantitative

文章引用: 钱祖建. 抗乳腺癌候选药物的 ADMET 性质的预测研究[J]. 建模与仿真, 2023, 12(1): 202-211.
DOI: 10.12677/mos.2023.121020

prediction models for the ADMET properties of the compounds were developed and investigated using both BP neural network and XGBoost regression methods. The results of the experimental study indicated that the XGBoost classification prediction model had the lowest error and the best results for this task compared to the BP neural network approach.

Keywords

Breast Cancer Drugs, BP Neural Network, XGBoot Classification Prediction Model

1. 引言

21 世纪以来，乳腺癌已经成为全世界范围内女性患病率和死亡率比较高的恶性肿瘤。乳腺癌是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。乳腺癌的发生发展与雌激素水平密切相关。约 70%的乳腺癌是雌激素受体 α (Estrogen receptors alpha, ERα)阳性乳腺癌[1]。ERα 的异常表达会促使乳腺癌的发生及进展。目前，治疗乳腺癌的方法有很多，比如手术治疗、物理治疗以及药物治疗等。针对 ERα 表达的乳腺癌患者的治疗，抗激素药物往往能够起到一定的效果，也是对于 ERα 表达的乳腺癌患者治疗的一种常用的手段。所以，那些可以拮抗 ERα 活性的药物就有可能用来治疗乳腺癌。

在抗乳腺癌药物[2]的研究中，抗乳腺癌候选药物应该具有很好的生物活性才能够更好地抑制 ERα。而一个化合物想要成为候选药物，不仅需要具备良好的生物活性(此处指抗乳腺癌活性)，还需要在人体内具备良好的药代动力学性质和安全性，也就是本文研究的 ADMET。具体来说，不良的药物吸收(Absorption)、分布(Distribution)、代谢(Metabolism)、排泄(Excretion)性质和毒性(ADMET)是导致药物开发失败的主要原因之一[3] [4]。一个化合物的活性再好，如果其 ADMET 性质不佳，比如很难被人体吸收，或者体内代谢速度太快，或者具有某种毒性，那么其仍然难以成为药物。所以，良好的 ADMET 特性是一个化合物能够成为乳腺癌候选药物的必要条件之一。基于此，我们对抗乳腺癌候选药物筛选过程中的 ADMET 性质的分类预测模型进行了研究，希望通过数据挖掘的处理技术来解决药物筛选建模的问题，并且可以通过数学建模的过程之中实现对候选模型的筛选过程中预测模型[5]的优化，这对于寻找更合适的抗乳腺癌药物、治疗 ERα 表达的乳腺癌患者具有重要意义。

本文利用临床试验所提供的 1974 个化合物的 ADMET 数据，分别构建化合物的 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型。利用“1”和“0”来分别表示 ADMET 性质的好坏。用小肠上皮细胞渗透性(Caco-2)为例，‘1’代表该化合物的小肠上皮细胞渗透性比较好，‘0’代表该化合物的小肠上皮细胞渗透性比较差。其余的四组 ADMET 性质的分类方法类似。本文分别运用 BP 神经网络[6]和 XGBoost 回归[7] [8]两种方法建立并研究了对化合物 ADMET 性质的定量预测模型。结果表示，对比于 BP 神经网络方法，XGBoost 分类预测模型对于该任务误差最低、效果最好。

2. 模型设计

2.1. 双隐藏层神经网络模型

本文采用含有一个隐藏层两层的多输入单输出结构 BP 神经网络模型来进行分类预测。采用该模型

分别对 ADMET 数据集中 Caco-2、CYP3A4、hERG、HOB、MN 五种性质的数据进行预测。以 Caco-2 的模型构建为例，其余的四组，改变样本输入的数据即可。

第一步：使用 Pandas [9]将 1974 个样本数据进行整理与分类，先进行标准化处理，将每一行作为一个样本，批量地进行训练。将 1974 个样本数据整理成数据集，并且按照 8:2 的比例分为训练集数据和测试集数据。

第二步：输入选取的是全部的分子描述符信息变量，Caco-2 的属性分类预测值作为网络的输出。所以输入层的神经元个数 $n = 20$，输出层的神经元的个数只有 1 个即 $m = 1$。输入数据选择的为全部分子描述符变量，需要学习的数据特征较多，增加了一层的隐藏层，更加抽象稳定地学习数据特征，得出样本中蕴含的规律，从而展现出数据的更为抽象的特性，这些特性能够用来更好的线性地划分。本方案考虑到在验证集上准确度，采用凑试法，经过反复的测试，初步确定双层隐含层的节点数为 128 和 10，有利于逐渐收敛网络的宽度和特征的高层次提取。

设计的双隐藏层神经网络结构示意图如图 1 所示。

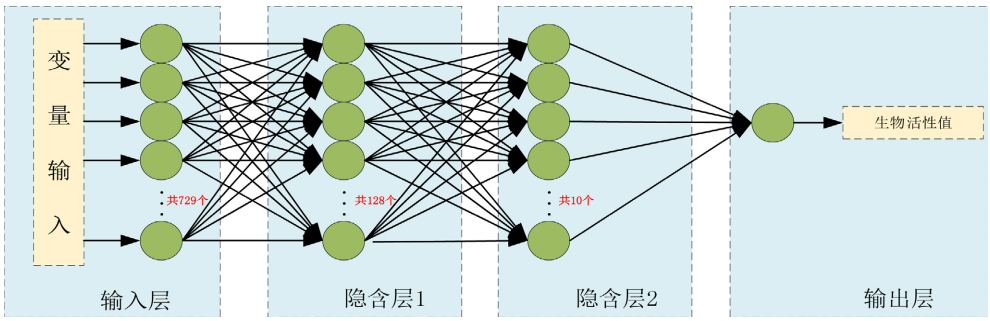


Figure 1. Double hidden layer neural network structure diagram
图 1. 双隐藏层神经网络结构图

2.2. XGBoot 模型

XGBoost 是一种改进的 Boosting 集成算法[10]。通过不断拟合上一颗树残差来不断产生新树，将树模型组合成为一个正确率最高，是泛化能力最强的分类器。XGBoost 由多棵决策树组成，若所有决策树的累加结果就是最终解[11]。

2.2.1. CART 回归树

XGBoost 所使用的树为 CART 回归树[12]，假设树为二叉树，通过不断将特征进行分裂形成回归树。比如当前树节点是基于第 j 个特征值进行分裂的，设该特征值小于 s 的样本划分为左子树，大于 s 的样本划分为右子树，有

$$R_1(j,s) = \left\{x \mid x^{(j)} \le s\right\} \tag{1}$$

$$R_2(j,s) = \left\{x \mid x^{(j)} > s\right\} \tag{2}$$

CART 回归树实质上就是在该特征维度对样本空间进行划分，而这种空间划分方式的优化是一个 NP (Non-Deterministic Polynomial)问题，因此在决策树模型[13]中是使用启发式方法解决的。经典 CART 回归树的目标函数为：

$$\sum_{x_i \in R_m} \left(y_i - f\left(x_i\right)\right)^2 \tag{3}$$

因此，求解此问题的最优的切分特征 j 和切分点 s 就转化为了公式所示的目标优化问题：

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}\left(y_i-c_1\right)+\min_{c_2}\sum_{x_i\in R_2(j,s)}\left(y_i-c_2\right)\right] \tag{4}$$

那么只要遍历所有特征的所有切分点，就能找到最优的切分特征 j 和切分点 s，最终找到一颗回归树。

2.2.2. XGBoost 算法思想

XGBoost 的算法思想就是不断添加树，然后再不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。当训练完成得到 k 棵树时，就要预测一个样本的分数，简单而言就是根据这个样本的特征，在每棵树中会落到对应的一个叶子节点，每个叶子节点就对应一个分数，最后只需要将每棵树对应的分数加起来就是该样本的预测值，即

$$\hat{y}_i=\Phi\left(x_i\right)=\sum_{k=1}^{K}f_k\left(x_i\right) \tag{5}$$

F 为所有 CART 树的函数空间，每个函数 f_k 函数值为样本点所在样本点的得分。 $\omega_{q(x)}$ 为叶子节点 q 的分数，T 表示叶子节点的个数，$f(x)$为其中一颗回归树。

2.2.3. XGBoost 目标函数

XGBoost 的一般目标函数被定义为：

$$Obj=\sum_{i=1}^{n}l\left(y_i-\hat{y}_i\right)+\sum_{k=1}^{K}\Omega\left(f_k\right) \tag{6}$$

$$\Omega\left(f_k\right)=\gamma T+\frac{1}{2}\lambda\left\|\omega\right\|^2 \tag{7}$$

上式中的 γ、λ 均为惩罚系数，经过 t 轮迭代后的目标函数为：

$$L^{(t)}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{(t-1)}+f_t\left(x_i\right)\right)+\Omega\left(f_t\right) \tag{8}$$

为了求最优化问题，让损失函数 $l\left(y_i,\Phi\right)$ 在 $\Phi=\hat{y}_i^{(t-1)}$ 处的二阶泰勒展开，得到最优解目标函数为：

$$\hat{L}^{(t)}=\sum_{i=1}^{n}\left[g_i\omega_q\left(x_i\right)+\frac{1}{2}h_i\omega_q^2\left(x_i\right)+\gamma T\right] \tag{9}$$

则最小损失的目标函数为：

$$\tilde{L}_{\min}=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{10}$$

其中 $G_j=\sum_{i\in I_j}g_j$，$H_j=\sum_{i\in I_j}h_i$。

信息增益(Information Gain)为：

$$\text{Gain}=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{G_L+G_R}{H_L+H_R+\lambda}\right]-T \tag{11}$$

XGBoost 算法在每轮测试时都会求出一个最优的信息增益，然后根据这个最优的信息增益来指导生成决策树，通过不断迭代找到最优树模型。

2.2.4. XGBoost 模型建立

根据已经有的 1974 个已知样本，每个样本包括 729 个特征(分子描述符)，将所有样本按照 8:2 的比例划分为训练集和验证集。训练中的参数：学习率为0.30 (learning rate=0.30)，最大深度为2 (max_depth=2)，最小叶子结点权重为 3 (min_child_weight=3)，迭代次数为 100 (num_boost_round=100)。通过 XGBoost 模型，得到预测结果。

3. 模型仿真与分析

3.1. BP 神经网络模型仿真验证

Python 拥有各种可调用的类库与框架，本文分别使用了 Python 中 Pandas 进行文件的数据预处理操作，使用 Tensorflow 框架[14]进行 BP 神经网络预测模型的建立、使用 Matplotlib 库进行数据可视化绘制。其中 BP 神经网络预测模型参数设置为：729 个输入节点、第一层网络神经元个数为 128，第二层网络神经元个数为 10，输出节点为 1。如图 2 为数据集的划分及其作用是示意图。

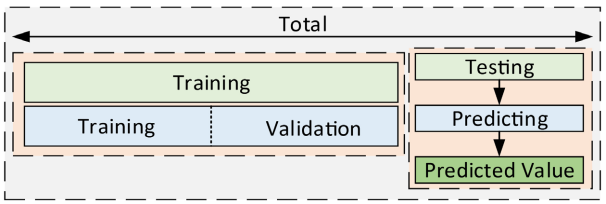
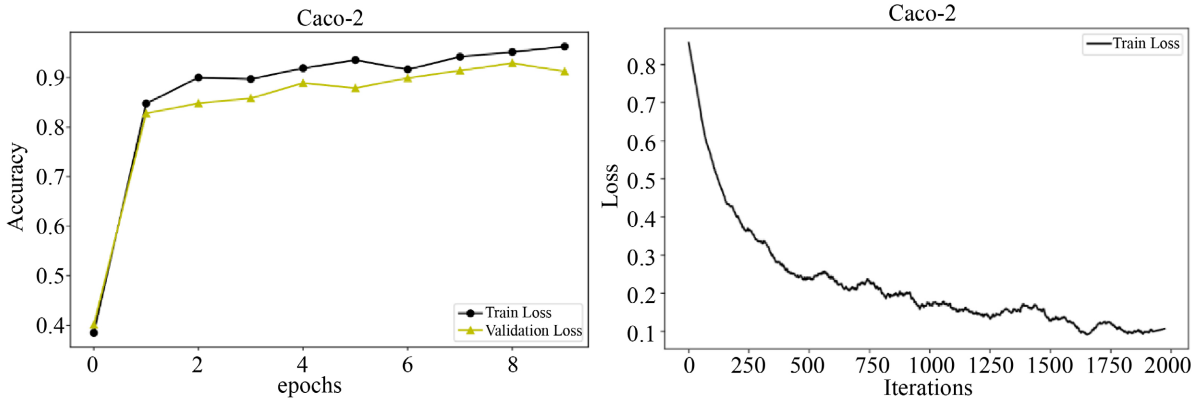


Figure 2. Data set division and role
图 2. 数据集划分及作用

再经过将样本数据集的 1974 个样本数据进行整理，按照 8:2 的比例拆分为训练集数据和验证集数据。其中验证集数据用于调试神经网络参数。

在数据集中随机抽取 1579 (80%)个样本用作训练，395 (20%)个样本用验证预测，将数据进行归一化并完成建模后，用 ADMET 数据集中 Caco-2、CYP3A4、ERG、HOB、MN 五种属性的分别独立的模型训练，得到如图 3 所示的结果。图 3 展现的是五种属性的准确度和目标损失随着迭代次数的变化规律。

如图 3，在左侧一列的五个图中，每个图中含有两个准确度，即训练集的准确度和验证集的准确度。右侧一列为对应的训练损失函数趋势图。对于 ADMET 的五种属性的训练结果图中发现，在相同网络模型结构下二分类情况训练结果趋势大致相同。



(a) (b)

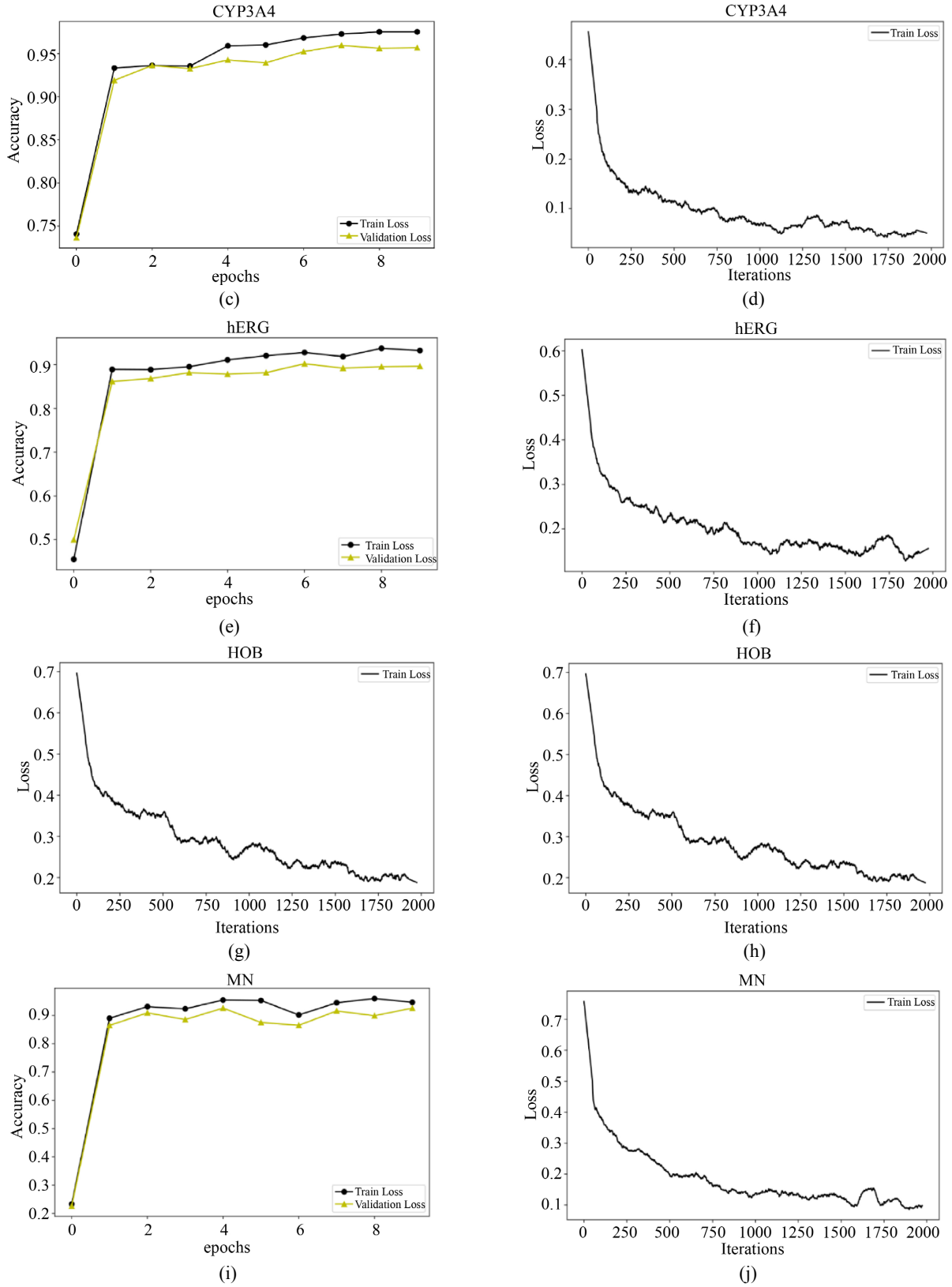


Figure 3. Accuracy and target loss curve variation of five ADMET characteristics
图 3. 五种 ADMET 特性的准确度与目标损失曲线变化图

双层隐藏网络在训练集上训练时，将每一次更新后的神经元权重参数对整个训练集的数据进行预测，随着迭代次数的增加，整个神经网络模型的两个准确度基本在完整的两个 epoch 后开始逐渐在一个区间内进行小幅波动。取三次后(含第三次)的两个准确度的平均值训练集上的平均准确率为 0.91，在验证集上的平均准确率为 0.865，见表 1。这二者说明此双层隐藏层模型的精度较高。此时，相对于真实目标分类进行分类的每个数据特征基本学习完成，但是 ADMET 的五种属性的损失预测仍存在约为 10%~20%的误差。

Table 1. Training set prediction rate and validation set prediction rate
表 1. 训练集预测率和验证集预测率

Attribute	Train Validation Accuracy	Average Validation Accuracy
Caco-2	0.928	0.891
CYP3A4	0.879	0.822
hERG	0.921	0.882
HOB	0.883	0.833
MN	0.934	0.896

3.2. XGBoot 模型仿真验证

表 2 表示为 XGBoost 模型对 ADMET 性质在验证集上正确率分析。

Table 2. XGBoost model accuracy rate
表 2. XGBoost 模型正确率

ADMET 属性	正确率
Caco-2	89.87%
CYP3A4	94.68%
hERG	91.64%
HOB	85.82%
MN	97.21%

图 4~8 分别为五个 ADMET 性质的 XGBoost 模型训练 loss 曲线和验证 loss 曲线。

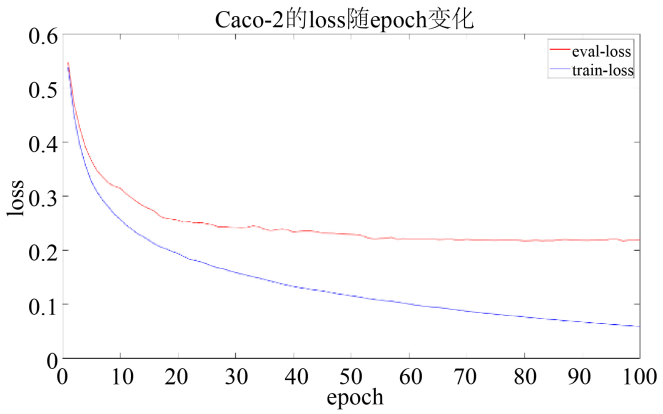


Figure 4. Loss curve of XGBoost model of Caco-2
图 4. Caco-2 的 XGBoost 模型 loss 曲线

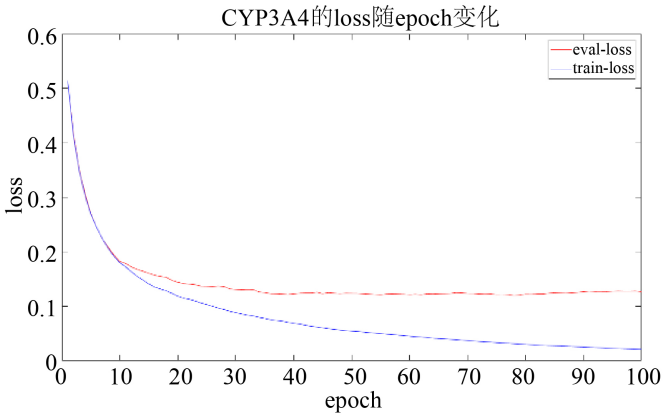


Figure 5. Loss curve of XGBoost model of CYP3A4
图 5. CYP3A4 的 XGBoost 模型 loss 曲线

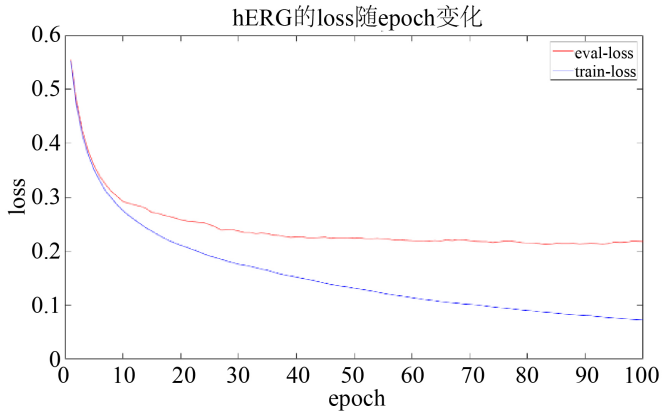


Figure 6. Loss curve of XGBoost model of hERG
图 6. hERG 的 XGBoost 模型 loss 曲线

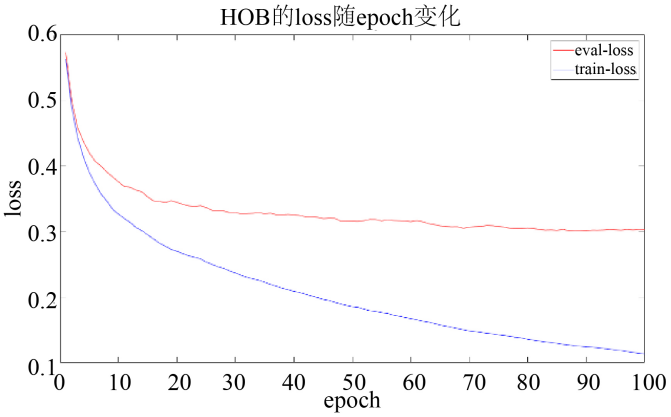


Figure 7. Loss curve of XGBoost model of HOB
图 7. HOB 的 XGBoost 模型 loss 曲线

图 4~8，利用 XGBoost 模型对 ADMET 的五种属性进行二分类预测，得到可视化结果。每种 ADMET 属性的目标损失随着 epoch 增加在该模型下的趋势变化情况，且每个属性的训练损失和评估(验证)损失都随着迭代次数的增加在逐渐减小，说明 XGBoost 模型在每轮测试时找到的最优树模型对输入的整体数据

特征的学习度在慢慢增加，对数据的拟合逐渐接近，验证了 XGBoost 模型对 ADMET 五种属性的二分类预测是适合的；这也体现在最终的模型验证正确率上，如表 2。ADMET 的五种属性平均验证正确率为 91.84%；在 MN 属性上拟合验证的效果最优，其验证正确率是 97.21%

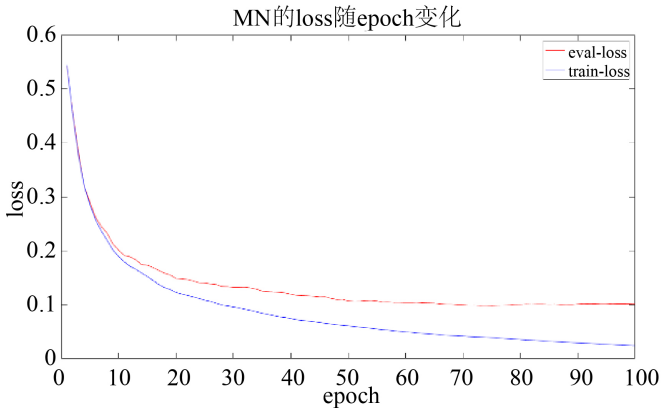


Figure 8. Loss curve of XGBoost model of MN
图 8. MN 的 XGBoost 模型 loss 曲线

3.3. 模型对于与分析

表 3 给出了分别使用 BP 神经网络和 XGBoost 模型预测五种 ADNET 性质的数值对比。

Table 3. Comparison of BP and XGBoost predicted values (only part shown)
表 3. BP 和 XGBoost 预测值对比(仅展示部分)

SMILES	BP					XGBoost				
	Caco-2	CYP3A4	HERG	HOB	MN	Caco-2	CYP3A4	HERG	HOB	MN
1	0	1	1	0	1	0	1	1	0	1
2	0	0	0	1	1	0	1	0	0	1
3	0	1	0	1	1	0	1	1	0	1
4	1	1	1	1	1	0	1	1	0	1
5	0	1	1	1	0	0	1	1	0	1
6	1	1	1	1	0	0	1	1	0	1
7	0	1	0	1	0	0	1	1	0	0
8	0	1	1	1	0	0	1	1	0	1
9	0	1	1	0	0	0	1	1	0	1
10	1	1	1	1	1	0	1	1	0	1

由表 3 所示，对比于 BP 神经网络方法，XGBoost 分类预测模型对于该任务误差最低、效果最好。

4. 总结

综上所述，本文针对抗乳腺癌候选药物的 ADMET 性质的数据预处理和相关分析。本文尝试利用两种不同的模型分别为BP神经网络模型和XGBoot模型来建立定量的预测模型，对比两种模型的预测结果，

可以发现 XGBoot 模型的效果更好，这可以更好的解决候选药物的 ADMET 性质的预测问题。将该模型进行推广，为候选药物的选择提供新思路，具有较好的借鉴意义。

参考文献

[1] 高源. 雌激素受体 α 抑制乳腺癌转移的作用及其机制研究[D]: [博士学位论文]. 西安: 中国人民解放军空军军医大学, 2018.

[2] 贺萍, 伍雁琦, 罗婷. 抗体药物偶联物在乳腺癌中的治疗现状及进展[J]. 临床肿瘤学杂志, 2022, 27(3): 255-264. https://kns.cnki.net/kcms/detail/detail.aspx?FileName=LCZL202203011&DbName=CJFQ2022

[3] Hou, T. and Wang, J. (2008) Structure-ADME Relationship: Still a Long Way to Go? *Expert Opinion on Drug Metabolism & Toxicology*, **4**, 759-770. https://doi.org/10.1517/17425255.4.6.759

[4] Van de Waterbeemd, H. and Gifford, E. (2003) ADMET in Silico Modelling: Towards Prediction Paradise? *Nature Reviews Drug Discovery*, **2**, 192-204. https://doi.org/10.1038/nrd1032

[5] 耿旭东. 基于机器学习的股票指数预测研究[D]: [硕士学位论文]. 开封: 河南大学, 2019.

[6] 孟建军, 潘彦龙, 陈晓强, 等. 基于灰色 BP 神经网络的高速列车轴箱轴承温度预测方法[J]. 轴承, 2022(4): 77-82. https://doi.org/10.19533/j.issn1000-3762.2022.04.013

[7] 李想. 基于 XGBoost 算法的多因子量化选股方案策划[D]: [硕士学位论文]. 上海: 上海师范大学, 2017. https://doi.org/10.7666/d.Y3258284

[8] 蒋晋文, 刘伟光. XGBoost 算法在制造业质量预测中的应用[J]. 智能计算机与应用, 2017, 7(6): 58-60. https://doi.org/10.3969/j.issn.2095-2163.2017.06.017

[9] 黄必栋. 基于 PySpark 和 Pandas 融合的大数据时序分析方法[J]. 电子技术与软件工程, 2022(1): 201-204.

[10] 韩亚鲁, 李绍稳, 郑文瑞, 等. 基于集成提升算法的土壤速效氮近红外光谱回归预测[J]. 激光与光电子学进展, 2021, 58(16): 547-557. https://doi.org/10.3788/LOP202158.1630005

[11] 瑞溢. 利用 XGBoost 和 SVR 算法的地铁站客流量模型研究[J]. 三明学院学报, 2019, 36(6): 56-64.

[12] 张兵, 夏时雨, 赵庆华, 等. 基于 CART 回归树模型的深基坑施工安全事故分析与预测[J]. 土木工程与管理学报, 2021, 38(3): 32-38+44. https://doi.org/10.3969/j.issn.2095-0985.2021.03.006

[13] 高虹雷, 门昌骞, 王文剑. 多核贝叶斯优化的模型决策树算法[J]. 国防科技大学学报, 2022, 44(3): 67-76. https://doi.org/10.11887/j.cn.202203009

[14] 江世雄, 黄鸿标, 陈苏芳, 等. 基于 GA 改进 LSTM-BP 神经网络的智慧楼宇用能行为预测方法[J]. 沈阳工业大学学报, 2022, 44(4): 366-371. https://doi.org/10.7688/j.issn.1000-1646.2022.04.02