Modeling and Simulation 建模与仿真, 2023, 12(4), 3930-3942
Published Online July 2023 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2023.124359

Hans汉斯

基于秃鹰搜索的抗乳腺癌候选药物优化建模

龙楷潮，袁学枫，张　利*

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2023年6月7日；录用日期：2023年7月17日；发布日期：2023年7月24日

摘　要

乳腺癌在全球范围内已取代肺癌成为最常见的癌症，并且其死亡率居高不下。因此，利用机器学习和智能优化算法等技术筛选乳腺癌药物对于推动乳腺癌治疗药物的发展至关重要。本文提出了一种基于改进的随机森林算法构建ERa活性预测模型的方法，并筛选出对生物活性最具影响力的前20个分子描述符。然后，使用该模型对50个化合物的IC50值和对应的pIC50值进行预测。同时，借助支持向量机(SVM)和Adaboost二分类模型，对化合物Caco-2、CYP3A4、hERG、HOB、MN的5种成分进行分别预测，并建立ADMET分类预测模型。最后，利用秃鹰搜索算法构建化合物筛选模型，使用黑鹰搜索算法融合前两个模型，解决各类复杂数值优化问题，以找到可行性药物操作变量范围。实验结果表明，所提出的预测模型具有很高的准确性，可应用于抗乳腺癌药物的研发。

关键词

乳腺癌，随机森林，ERa活性预测，ADMET分类预测，秃鹰搜索算法

Optimization Modeling of Anti-Breast Cancer Candidate Drugs Based on Bald Eagle Search

Kaichao Long, Xuefeng Yuan, Li Zhang*

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: Jun. 7th, 2023; accepted: Jul. 17th, 2023; published: Jul. 24th, 2023

Abstract

Breast cancer has replaced lung cancer as the most common cancer worldwide, and its mortality

*通讯作者。

文章引用: 龙楷潮, 袁学枫, 张利. 基于秃鹰搜索的抗乳腺癌候选药物优化建模[J]. 建模与仿真, 2023, 12(4): 3930-3942.
DOI: 10.12677/mos.2023.124359

rate remains high. Therefore, the selection of breast cancer drugs using techniques such as machine learning and intelligent optimization algorithms is of great significance to drive the development of breast cancer treatment drugs. In this paper, we propose a method based on the improved random forest algorithm to construct an ERa activity prediction model and select the top 20 most influential molecular descriptors for biological activity. Subsequently, using this model, we predict the IC50 values and corresponding pIC50 values of 50 compounds. Furthermore, with the aid of support vector machine (SVM) and Adaboost binary classification models, we predict the five components (Caco-2, CYP3A4, hERG, HOB, MN) of the compounds separately and establish an ADMET classification prediction model. Finally, we construct a compound screening model using the Bald Eagle search algorithm and integrate it with the previous two models using the Black Hawk search algorithm to address various complex numerical optimization problems and determine the feasible range of drug operating variables. Experimental results demonstrate that the proposed prediction model exhibits high accuracy and can be applied to the development of anti-breast cancer drugs.

Keywords

**Breast Cancer, Random Forest, ERa Activity Prediction,
ADMET Classification Prediction, Bald Eagle Search Algorithm**

1. 引言

乳腺癌目前发病率在 2~3/万之间，高居全球第一，而且还在呈上升趋势，年龄也越来越年轻化。在研究 ERα 基因缺失小鼠的实验结果中，发现 ERα 是治疗乳腺癌的重要靶标，因此能够拮抗 ERα 活性的化合物可能是治疗乳腺癌的候选药物[1]。但是想要成为候选药物，除了需要具备良好的生物活性外，还需要在人体内具备良好的药代动力学性质和安全性，合称为 ADMET (Absorption 吸收、Distribution 分布、Metabolism 代谢、Excretion 排泄、Toxicity 毒性)性质[2]。但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，因而还需要进行 ADMET 性质优化。

传统药物研发渠道的平均成本为 26 亿美元，大概耗时 12 年，因此如何在降低成本和时间的同时确保药物的有效性成为药物公司的重大难题，基于机器学习、深度学习辅助药物各个阶段的研发越来越成为各大公司的首选。基于图注意力网络，构造分子图作为分子结构特征的药物 ADMET 分类预测模型进行药物研发的虚拟筛选，据有良好的精准性[3]。采用 Chemoffice 2004 中的 MOPAC-PM3 算法筛选量化吡喃酮类化合物的量子化学结构，利用人工神经网络中的径向基网络建立分子结构描述符与生物活性间的相关模型，有效的提高了对吡喃酮类化合物结构的预测精度[4]。基于 RegNet-1d 模型和积分梯度法的 ERα 拮抗剂的生物活性预测方法，通过搭建 RegNet-1d 深度学习模型，并以积分梯度法为理论基础进行数据结构优化，变量对生物活性影响的相关性分布，以此筛选合适的分子描述符变量，时优化后的模型预测准确率略有下降但所需测量的数据量大大减少，节约了药物研发的时间和成本[5]。采用分子描述、支持向量机、遗传算法三种机器学习建立 ADMET 的 QSAR 预测模型，验证结果得出可推广应用至药物代谢、毒性评估等方面[6]。

本文采用“华为杯”第十八届中国研究生数学建模竞赛 D 题中的数据，包括 1974 个化合物样本，每个样本都有 729 个分子描述符变量，5 个 ADMET 性质数据。将从分子描述符出发构建预测模型，基于融合遗传算法的随机森林算法来预测化合物的 IC50 值和对应的 pIC50 值预测值。同时再借助 SVM 与 Adaboost 二分类模型，对化合物 Caco-2、CYP3A4、hERG、HOB、MN 的 5 中成分进行分别预测，建立 ADMET 分类预测模型，从而能找到既能满足较高的化合物活性，也能拥有较好的 ADMET 性质，助于抗乳腺癌药物的研发。

2. 构建 ERα 生物活性的定量预测模型

2.1. 特征选择

在不破坏原始数据可解释性的前提条件下，依据对分子描述符的显著性影响排名，进行特征选择。本文对各变量相关性进行分析，在最大程度上保留原始数据信息的同时，将分子变量描述符数据的维度从 729 维降至 20 维。在降维和筛选变量之前首先需要对全部 729 个变量数据进行整定。第一步：为减少计算量，删除冗余分子描述符，即需要过滤掉方差为 0 的特征[7]。第二步：对变量数据进行归一化处理使各个特征的尺度控制在相同的范围内，这样可以便于在计算分子描述符之间相关性。第三步：对以上预处理后的数据进行特征选择，将分子变量描述符数据的维度从 729 维降至 20 维。

第一步的整定算法流程，利用 Python 的 sklearn 包中 Variance Threshold 方法，他是一个简单的特征选择基准方法，该方法就是去除所有没有达到指定阈值的特征。默认是去除所有零方差的数据。第二步计算方法为：Min-Max 标准化(Min-Max Normalization) (线性函数归一化)也称为离差标准化，是对原始数据的线性变换，使得结果映射到 0~1 之间。利用第一步的剔除后变量数据导入第二步进行变量归一化处理，根据上述数据整定算法可以实现对原始数据效果更佳的筛选[8]，采用 Python 语言编程实现，得到数据整定结果，整定前后样本数皆为 1974，而操作变量数由 729 降为 504。第三步利用过滤式中的互信息法[9]以及 lasso 回归算法进行特征选择[10]，设计了综合筛选模型，并调用 Pandas 工具包对各变量之间的相关度分析，去除相关度低的部分变量，最终得到符合条件的主要影响变量。采用 Python 语言对上述两种特征选择方法实现，得出前 20 个对生物活性最具有显著影响的分子描述符(即变量)，maxHsOH'，'BCUTc-1h'，'minHBa'，'minwHBa'，'SaaCH'，'MLFER_A'，'maxHBa'，'MAXDN2'，'BCUTp-1h'，'gmin'，'maxHBd'，'maxHCsats'，'minssO'，'hmin'，'minHBint10'，'MDEO-12'，'minHBint6'，'ATSc4'，'MDEC-22'，'C2SP2'，构建这些化合物对 ERα 生物活性的定量预测模型。

2.2. 基于改进的随机森林算法对 ERα 的活性预测

随机森林算法属于 bagging 算法的一种，也属于 bagging 算法的一种加强算法，是将多棵 CART 回归树集成的一种有监督学习算法，其样本的数据集输入为式 x 为 SMILES，y 为输出的活性值，具体公式：

$$E=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\cdots,\left(x_m,y_m\right)\right\} \tag{2-1}$$

迭代次数为 t 次，即是对训练集进行 $t=1,2,\cdots,t$ 次分别采样，得到最终的集合 E，所得集合的算术平均值就是最后的模型输出。

在对模型进行训练时采取 7:3 的训练验证集比例，在训练过程中融入遗传算法加快模型收敛速度，寻找更优解，遗传算法是通过选择、交叉以及变异等机制，模拟出一个人工种群的进化过程，借鉴生物界自然选择和自然遗传机制的随机优化搜索算法，为避免决策树陷入过拟合，本文使用遗传算法来优化参数本文使用的决策树共有五个超参数：n_estimators (随机森林包含的弱分类器数量)、max_depth (树的最大深度)、min_samples_leaf (叶子节点包含的样本数)、min_samples_split (分枝所包含的最少样本个数)、

max_features (选的最大特征数)。将初始种群的是数量设置为 100，维度设为 5，随机初始化种群，适应度函数为MSE。经过 50 次迭代后 n_estimators = 15，max_depth = 15，min_samples_leaf = 2，min_samples_split = 3，max_features = 7 为最优解。当使用原始随机森林算法时得到的准确度(R^2)和均方误差(MSE)分别为 0.61，0.76，利用改进后的随机森林算法获得的结果为 0.73，0.52 其中准确度提高了 0.12，均方差减小了 0.24，由实验结果可得当使用改进后的随机森林算法拟合原数据能力更强。通过改进的随机森林算法预测模型对特征 1974 个化合物中 30%数据的 IC50 值和对应的 pIC50 值验证效果如下图 1 所示。

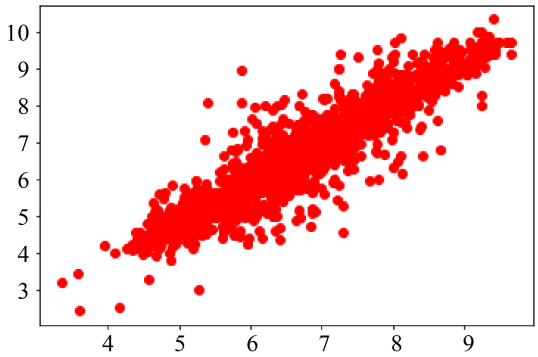

Figure 1. Validation of IC50 values and corresponding pIC50 values
图 1. IC50 值和对应的 pIC50 值验证效果

运用此模型，再对 50 个化合物的生物活性值进行 IC50 值和对应的 pIC50 值预测，具体预测结果如下表 1 所示。

Table 1. Prediction results of IC50 values and corresponding pIC50 values
表 1. IC50 值和对应的 pIC50 值预测结果

SMILES	IC50_nM	pIC50	SMILES	IC50_nM	pIC50
1	14.323	6.338	26	10.248	6.673
2	6.557	7.119	27	13.199	6.42
3	4.624	7.469	28	16.694	6.185
4	6.557	7.119	29	19.288	6.041
5	10.174	6.68	30	18.456	6.085
6	7.166	7.031	31	35.922	5.419
7	6.055	7.199	32	35.922	5.419
8	6.008	7.207	33	35.922	5.419
9	5.483	7.298	34	30.291	5.589
10	10.908	6.61	35	17.282	6.15
11	10.268	6.671	36	7.966	6.925
12	11.603	6.549	37	7.966	6.925
13	10.908	6.61	38	3.316	7.857
14	10.268	6.671	39	19.068	6.052

Continued

15	14.004	6.361	40	19.785	6.015
16	14.004	6.361	41	19.785	6.015
17	10.466	6.652	42	19.785	6.015
18	6.313	7.157	43	19.785	6.015
19	6.914	7.066	44	45.663	5.179
20	14.263	6.342	45	19.785	6.015
21	6.476	7.132	46	45.663	5.179
22	6.235	7.17	47	19.785	6.015
23	36.958	5.39	48	15.651	6.249
24	40.037	5.39	49	15.111	6.285
25	11.008	6.601	50	17.517	6.137

3. 构建 ADMET 性质预测模型

对于所提供的 1974 个化合物的 ADMET 是数据，分别构建化合物 Caco-2 (小肠上皮细胞渗透性)、CYP3A4 (能否够被 CYP3A4 代谢)、hERG (是否具有心脏毒性)、HOB (口服生物利用度)、MN(是否具有遗传毒性)的分类预测模型，然后使用所构建的 5 个分类预测模型，鉴于化合物标签均为 2 分类模型，因此我们使用广泛且经典 SVM 分类模型和 Adaboost 分类模型进行分析比较，并通过多个评价指标来选择该化合物相应的模型。

3.1. SVM 分类模型

由于数据样本是离散的，因此采用机器学习中的支持向量机(SVM)算法进行网络训练，空心圆点和黑心圆点代表两类不同类别的样本; H 为分类线，H1，H2 分别为平行于分类线的直线，它们经过离分类线最近的那些少量的样本点，两者间距离称为分类间隔。H 线将两个不同的类正确隔离，同时使分类间隔最大化。设样本集为$(x_i, y_i), i=1,2,\cdots,m,\ y_i \in \{-1,1\}$，并满足：$y_i[w\cdot x_i)-b]-1\geq 0$ 该分类的间隔等于$\frac{2}{\|w\|}$，其间隔最大等价于求$\|w\|^2$的值最小。满足上式且$\frac{1}{2}\|w\|^2$最小的分类平面叫做最优分类面，H1，H2 两条平行直线上的那些训练样本点称为支持向量，利用拉格朗日方法可以把求解最优分类面的原始问题转化为其对偶问题，即在条件$\sum_{i=1}^{n} a_i y_i = 0, a_i \geq 0,\ i=1,2,3,\cdots,n$下对 ai 求解以下最大的函数值如下：

$$Q(a)=\sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i, x_j) \tag{3-1}$$

ai 为原问题中与(1)式对应的拉格朗日乘子，该问题就是求解二次凸规划的优化问题，存在唯一最优解，可以证明，有些乘子 a_i 不为零，它也就是支持向量。求解该问题得到最优平面的 w^*和 b^*，此时最优分类函数

$$D(x)=\operatorname{sgn}\left((w^*\cdot x)-b^*\right)=\operatorname{sgn}\left(\sum_{i,j=1}^{n} a_i^* y_i (x_i\cdot x)-b^*\right) \tag{3-2}$$

求和只对支持向量进行；b^*是偏移量。根据泛函分析中的度量空间理论，倘若有一种核函数$k(x_i, x_j)$满足 Mercer 条件的核函数替换线性算法的内积就可以找到原输入空间中对应的非线性算法。如果用特征

空间的$\varphi(x)$代替x，则上式转化为下式

$$Q(a)=\frac{1}{2}\sum_{i,j=1}^{n}a_i a_j y_i y_j \varphi(x_i)\varphi(x_j) \tag{3-3}$$

而相应的分类函数变为

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}a_i^* y_i k(x_i\cdot x)-b^*\right) \tag{3-4}$$

SVM 分类算法中，如果定义不同的内积函数，就能实现多项式逼近、贝叶斯分类器、径向基函数(RBF)方法等选用不同的核函数就可以构造不同的 SVM [11]。在本模型中我们选择径向基函数(RBF)为模型的核核函数构造 SVM 分类器。

3.2. Adaboost 分类模型

在集成学习原理中，集成学习按照个体学习器之间是否存在依赖关系可以分为两类，第一个是个体学习器之间存在强依赖关系，另一类是个体学习器之间不存在强依赖关系。前者的代表算法就是是 boosting 系列算法。在 boosting 系列算法中，Adaboost 是最著名的算法之一。Adaboost 既可以用作分类，也可以用作回归，本次二分类任务中使用了其分类功能[12]。AdaBoost 二元分类问题算法流程如下。

输入为样本集$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，输出为{−1, +1}，弱分类器算法，弱分类器迭代次数K。输出为最终的强分类器$f(x)$。

1) 初始化样本集权重函数如下

$$D(1)=(w_{11},w_{12},\cdots,w_{1m});\ w_{1i}=\frac{1}{m};\ i=1,2,\cdots,m \tag{3-5}$$

2) 对于$k=1,2,\cdots,k$：

a) 使用具有权重D_k的样本集来训练数据，得到弱分类器 Gk(x)

b) 计算$G_k(x)$的分类误差率

$$e_k=P(G_k(x_i)\neq y_i)=\sum_{i=1}^{m}w_{ki}I(G_k(x_i)\neq y_i) \tag{3-6}$$

c) 计算弱分类器的系数

$$a_k=\frac{1}{2}\log\frac{1-e_k}{e_k} \tag{3-7}$$

d) 更新本集的权重分布

$$w_{k+1,i}=\frac{w_{ki}}{Z_K}\exp(-\alpha k y_i G_k(x_i)),\ i=1,2,\cdots,m \tag{3-8}$$

这里Z_k是规范化因子

$$Z_k=\sum_{i=1}^{m}w_{ki}\exp(-\alpha_k y_i G_k(x_i)) \tag{3-9}$$

3) 构建最终分类器为式：

$$f(x)=\operatorname{sign}\left(\sum_{k=1}^{K}\alpha_k G_k(x)\right) \tag{3-10}$$

对于 Adaboost 多元分类算法，其实原理和二元分类类似，最主要区别在弱分类器的系数上。比如 Adaboost SAMME 算法，它的弱分类器的系数如下所示

$$\begin{cases} \alpha_k = \frac{1}{2}\log\frac{1-e_k}{e_k} + \log(R-1) \end{cases} \tag{3-11}$$

3.3. 测试结果及分析

化合物 Caco-2、CYP3A4、hERG、HOB、MN 的分类预测模型通过建立的 SVM 模型和 Adaboost 模型，使用 7:3 的比列划分训练集和测试集，如图 2、图 3 所示。

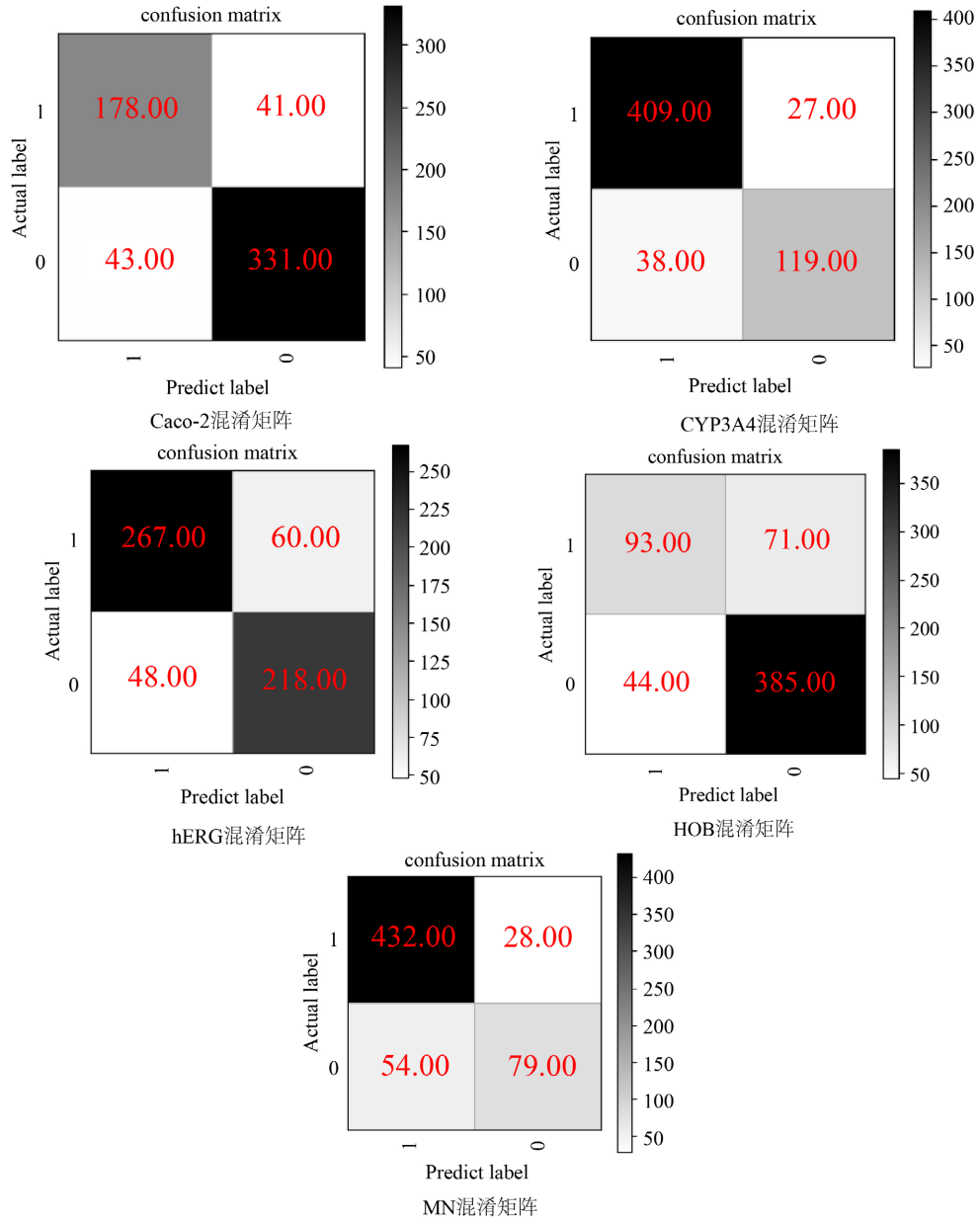

Figure 2. Confusion matrix of the Adaboost model on the test set
图 2. Adaboost 模型测试集混淆矩阵

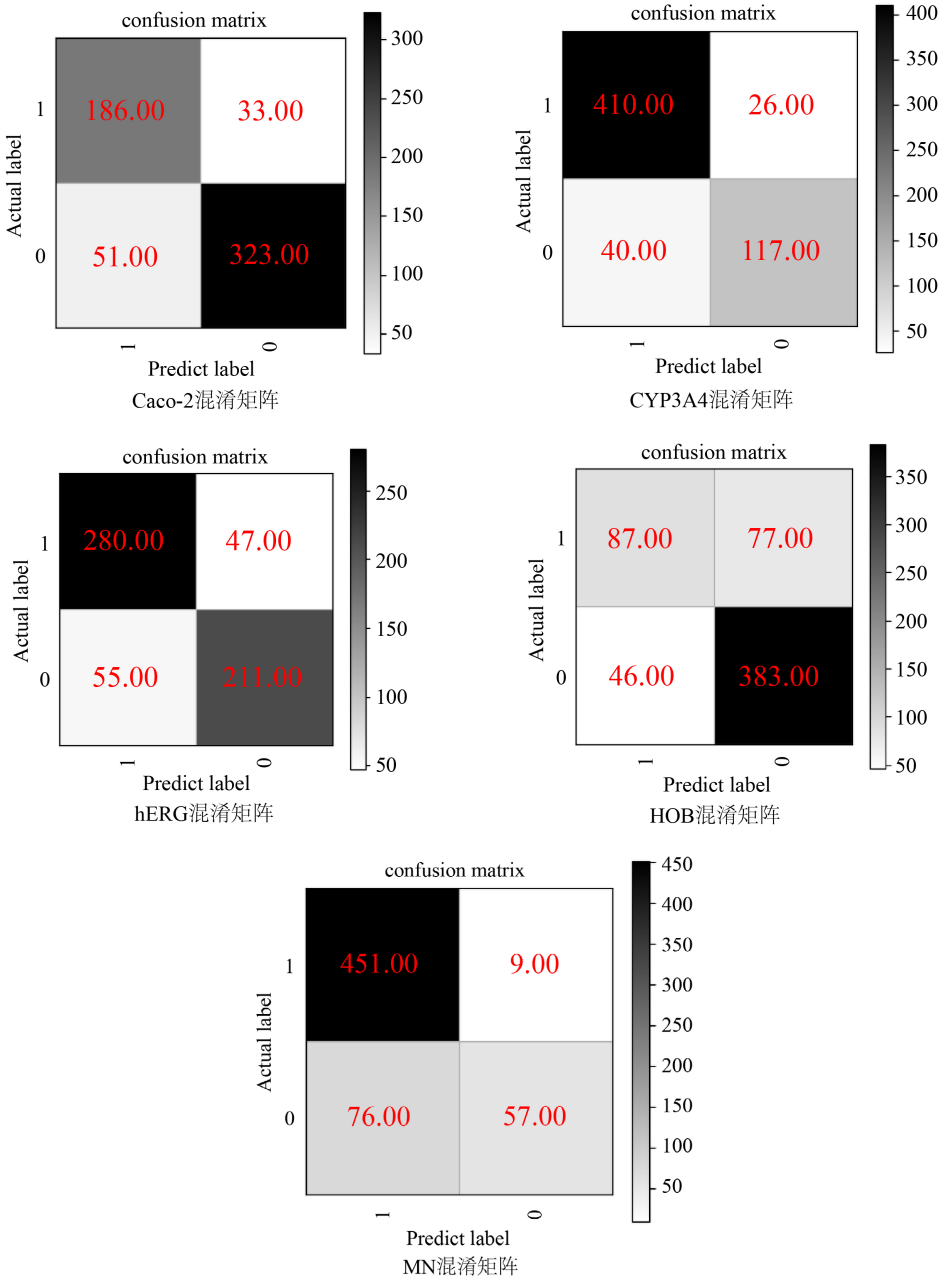

Figure 3. Confusion matrix of the SVM model on the test set
图 3. SVM 模型测试集混淆矩阵

如上所示图混淆矩阵主对角线为测试集中预测正确的样本，副对角线为测试集中预测错误的样本，通过实验对比各模型在化合物 Caco-2、CYP3A4、hERG、HOB、MN 的分类结果，以准确率，精度，召回率，F1 值 5 种评价指标已经药物的实际作用综合选择模型。实验结果如下表 2、表 3 所示。

所以在选择模型时不能只看某一方面数据，要做考虑到实际药物摄入与人体的友好性相关，所以在选择检测化合物 Caco-2 模型时，为缩小样本空间，确保药物的精度选择 Adaboost 模型，其精准率较 SVM 高 0.02%。在选择化合物 CYP3A4 时，考虑到该药物的代谢情况，优先选择精准率较高的 SVM 分类器。在选择化合物 hERG 时，考虑到该化合物对心脏有毒性，在精度相同时，选择 F1 值较大的，应为 F1 综

Table 2. Evaluation of the Adaboost model's prediction results
表 2. Adaboost 模型预测结果评估

	Adaboost				
	Caco-2	CYP3A4	hERG	HOB	MN
准确率/%	0.85	0.89	0.81	0.80	0.86
精度/%	0.80	0.91	0.84	0.67	0.88
召回率/%	0.81	0.93	0.81	0.56	0.94
F1 值/%	0.81	0.92	0.83	0.61	0.91

Table 3. Evaluation of the SVM model's prediction results
表 3. SVM 模型预测结果评估

	Adaboost				
	Caco-2	CYP3A4	hERG	HOB	MN
准确率/%	0.85	0.88	0.82	0.79	0.85
精度/%	0.78	0.93	0.84	0.65	0.85
召回率/%	0.84	0.94	0.85	0.53	0.98
F1 值/%	0.81	0.92	0.84	0.58	0.91

合考量了正确率与召回率。及在测试该化合物时选取 SVM 模型。在选择化合物化合物 HOB 时，考虑到该化合物口服生物利用度，选择精度较高的 Adaboost 模型。在选择化合物 MN 时，考虑到该化合物的遗传毒性，选择精度较高的 Adaboost 模型。

4. 秃鹰搜索(BES)与化合物筛选模型的建立

为了能从操作变量的取值区间内找到对抑制 ERα 具有更好的生物活性，同时具有更好的 ADMET 性质的化合物，将使用一种智能搜索算法，将模型 1 作为种群的适应度函数，通过迭代计算使种群向最佳适应度位置(药物活性最高时的 20 操作变量的值)移动，直至完成求解过程。其次，从智能搜索算法的历史最佳搜索位置中依次选取全局最优位置，将此位置输入模型 2 至模型 6 完成 ADMET 性质预测，直至筛选出 ADMET 性质满足要求的药物。

秃鹰搜索算法：本文模型一使用改进的随机森林建模，其内部包含多颗决策树作为弱分类器，由于决策树工作时呈现非线性，使得模型一总体上呈现较强的非线性[13]。在模型一中搜索最优解是一种复杂的非线性过程，而在这类问题上具有显著的竞争优势，因此选择这一算法来进行建模。

BES 寻优过程可分为三个阶段。在第一阶段(选择空间)，鹰选择猎物数量最多的空间，从中心点移动到选定的搜索区域，具有较强全局寻优能力；在第二阶段(空间搜索)，鹰在选定的空间内移动寻找猎物，具有较强全局寻优能力；在第三阶段(俯冲)，鹰从第二阶段确定的最佳位置摆动，并确定最佳狩猎，即全局最优位置。计算如下：

1) 选择阶段：

秃鹰在选择的搜索空间中识别并选择最佳的区域(根据食物的数量)，在那里它们可以捕食。公式(4-1)用数学方法表示了这种行为：

$$P_{i,\text{new}} = P_{\text{best}} + \alpha^{*} r\left(P_{\text{mean}} - P_i\right) \tag{4-1}$$

其中，这里是控制位置变化的参数，取值在[1.5, 2]之间。是[0, 1]之间的一个随机数。P_{best}是当前最优位置，P_{mean}表示利用前面所有点的信息。P_i为第只秃鹰的位置。

2) 搜索阶段：

秃鹰在选定的搜索空间内搜索猎物，并在螺旋形空间内以不同的方向移动，以加快搜索速度。俯冲的最佳位置用以下数学公式表示

$$P_{i,\text{new}} = P_i + y(i)*(P_i - P_{i+1}) + x(i)*(P_i - P_{\text{mean}}) \tag{4-2}$$

$$x(i) = \frac{xr(i)}{\max(|xr|)}, y(i) = \frac{yr(i)}{\max(|yr|)} \tag{4-3}$$

$$xr(i) = r(i)*\sin(\theta(i)),\ yr(i) = r(i)*\cos(\theta(i)) \tag{4-4}$$

$$\theta(i) = a^*\pi^*\text{rand}(c)\,r(i) = \theta(i) + R^*\text{rand} \tag{4-5}$$

其中：(i)与(i)分别为螺旋方程的极角与极径，a 与 R 是控制螺旋轨迹的参数，变化范围分别为(5, 10)、(0.5, 2)。rand 为(0, 1)内随机数，$x(i)$与 $y(i)$表示极坐标中秃鹰位置，取值均为(−1, 1)。秃鹰位置跟新如以下公式所示：

$$P_{i,\text{new}} = \text{rand}*P_{\text{best}} + x1(i)*(P_i - c1*P_{\text{mean}}) + y1(i)*(P_i - c2*P_{\text{best}}) \tag{4-6}$$

3) 俯冲阶段：

秃鹰从搜索空间的最佳位置摇摆到它们的目标猎物。所有的点也都朝着最好的点移动。下面三个公式从数学上说明了这种行为：式中:增加了秃鹰向最佳点和中心点的移动强度.

$$\theta(i) = a^*\pi^*\ \text{rand}\ , r(i) = \theta(i) \tag{4-7}$$

$$P_{i,\text{new}} = \text{rand}*P_{\text{best}} + x1(i)*(P_i - c1*P_{\text{mean}}) + y1(i)*(P_i - c2*P_{\text{best}}) \tag{4-8}$$

$$x1(i) = xr(i)/\max(|xr|),\ y1(i) = yr(i)/\max(|yr|) \tag{4-9}$$

俯冲过程，位置根据以下式格式跟新如下，其中 c_1 与 c_2 表示秃鹰向最佳与中心位置的运动强度，取值均为(1, 2)。

$$\begin{aligned} \delta_x &= x1(i)*(P_i - c_1*P_{\text{mean}}) \\ \delta_y &= y1(i)*(P_i - c_2*P_{\text{best}}) \end{aligned} \tag{4-10}$$

$$P_{i,new} = \text{rand}*P_{best} + \delta_x + \delta_y \tag{4-11}$$

药物筛选策略：药物筛选策略分为两部分，第一部分利用黑鹰搜索算法从模型一寻找全局最优位置，即最高 pIC50 值所对应的 20 个操作变量的值。第二部分用模型 2~6 对第一部分中得到的全局最优位解进行 ADMET 性质友好性筛选，挑选符合条件的解。

① 黑鹰搜索寻找模型一全局最优位置

设置初始变量的维度为 20，正好与选取的 20 个操作变量一一对应，使用模型 1 的预测结果作为种群适应度计算。种群大小设置为 100，迭代次数 50，算法中的超参数使用文献[14]提供的值。算法首先从 0 到 1 之间随机选取操作变量初始值，接着算法进入迭代求解，每次迭代都经历选择空间、空间搜索、俯冲三个阶段，直到达到最大迭代次数，或满足求解精度，黑鹰搜索将每次迭代的结果存放至历史最优解列表，它代表者黑鹰搜索寻优的轨迹，常用收敛曲线表示。随着迭代次数的增加，得到的最优解逐渐接近真实最优解，并且最终收敛在真实最优解附近。

② ADMET 性质友好性筛选

ADMET 性质友好性主要是指：满足化合物的小肠上皮细胞渗透性较好、能够被 CYP3A4 代谢、不具有心脏毒性、口服生物利用度较好、不具有遗传毒性，五者中的三个及以上指标[15]。也就是模型 2~6 预测所预测的结果中至少有满足以下条件中的至少三条：Caco-2 为‘1’、CYP3A4 为‘1’、hERG 为‘0’、HOB 为‘1’、MN 为‘0’。

从黑鹰搜索的历史最优解中从后往前依此取出一个最优解，并将最优解分别输入至模型 2~6，五个二分类模型根据输入值预测 Caco-2、CYP3A4、hERG、HOB、MN。

并且使用以下函数作为评分器，进行打分。其中 Modeli(x)表示模型 i 对化合物 x 的预测结果，target = {1, 1, 0, 1, 0}，评分器表示输入的化合物满足 ADMET 性质友好性中的数量。当评分器输出结果大于或等于 3 时，认为该化合物 DMET 性质友好，将其作为抗胰腺癌候选药物输出。如果该值低于 3，则放弃放弃该药物，并从黑鹰搜索的历史最优解中取出下一种药物(诺该化合物与前面值相同，则自动往后取出下一种化合物)如下式，判断器 ADMET 性质对人体友好性。

$$\text{score}(x)=\sum_{i=2}^{i=6}\text{score}_i(x) \tag{4-12}$$

5. 结果分析

运行上述程序次得到如下表 4 所示结果如下，其中 pIC50 为 9.113，活性较高。其 ADMET 性质中 Caco-2 为 1，CYP3A4 为 0，hERG 为 0，即该化合物对小肠上皮细胞渗透性较好、能够被 CYP3A4 代谢、不具有心脏毒性。虽然该化合物 MN 为 1、HOB 为 0，但总体上表现为 ADMET 友好性。

Table 4. Properties of candidate compounds
表 4. 候选化合物性质

Caco-2	CYP3A4	hERG	HOB	MN	pIC50
1	1	0	0	1	9.113

表 5 是黑鹰搜索的候选化合物原始坐标，由于原数据经过为归一化处理，因此需要将获得的候选物坐标还原成真实值。如表 6 所示，其值满足所选 20 个操作变量的取值范围，连续性等，故该化和物能够被选为真实候选药物。因此当选取的 20 个操作变量的值位于该化合物附近时，可认为其具有与该化合物相似的性质。

Table 5. Real coordinates of candidate compounds
表 5. 候选化合物真实坐标

maxHsOH	2.91E−01	maxHBd	7.76E−01
BCUTc-1h	4.80E−01	maxHCsats	1.28E+00
minHBa	1.19E+01	minssO	4.11E+00
minwHBa	1.63E−01	hmin	−5.10E−01
SaaCH	4.74E−01	minHBint10	8.94E+00
MLFER_A	3.31E−01	MDEO-12	2.26E+00
maxHBa	7.94E+00	minHBint6	3.24E−01

Continued

MAXDN2	3.57E+00	ATSc4	3.23E+00
BCUTp-1h	−2.73E+00	MDEC-22	3.52E−01
gmin	1.30E+01	C2SP2	2.36E−01

Table 6. Results of Black Hawk search
表 6. 黑鹰搜索结果

maxHsOH	0.341	maxHBd	0.91
BCUTc-1h	0.896	maxHCsats	0.998
minHBa	0.891	minssO	0.611
minwHBa	0.541	hmin	0.08
SaaCH	0.154	minHBint10	0.863
MLFER_A	0.008	MDEO-12	0.381
maxHBa	0.513	minHBint6	0.092
MAXDN2	0.474	ATSc4	0.135
BCUTp-1h	0.494	MDEC-22	0.459
gmin	0.567	C2SP2	0.007

根据小白鼠试验发现 ERα 活性的化合物是治疗乳腺癌的候选药物，但要想成为药物，就需要兼顾 ADMET 性质，因此本文首先采取基于改进的随机森林算法对 ERα 的活性预测建立模型一，再通过 SVM 分类模型和 Adaboost 分类模型构建化合物 Caco-2 (小肠上皮细胞渗透性)、CYP3A4 (能否够被 CYP3A4 代谢)、hERG (是否具有心脏毒性)、HOB (口服生物利用度)、MN (是否具有遗传毒性)的分类预测模型建立模型二，以上两种模型均使用 7:3 的比列划分训练集和测试集。最后通过秃鹰搜索算法将前两个模型结合建立最终的药物筛选模型，从而能得到最优解，实验结果表明利用此方法能对备筛化合物库进行预筛，去除 ADMET、分子稳定性、溶解性等性质较差的化合物，通过减少高通量筛选的压力，从而提高筛选效率；同时数据库质量的提升还可以降低药物开发后期的失败率，以避免化合物药代动力学引起的开发失败，降低开支。

参考文献

[1] Wang, B.J., Shen, Y.H., Liu, T.Y. and Li, T. (2021) ERα Promotes Transcription of Tumor Suppressor Gene ApoA-I by Establishing H3K27ac-Enriched Chromatin Microenvironment in Breast Cancer Cells. *Journal of Zhejiang University-Science B*, **22**, 1034-1044. https://doi.org/10.1631/jzus.B2100393

[2] Khamouli, S., *et al*. (2022) QSAR Modeling, Molecular Docking, ADMET Prediction and Molecular Dynamics Simulations of Some 6-Arylquinazolin-4-Amine Derivatives as DYRK1A Inhibitors. *Journal of Molecular Structure*, **1258**, Article ID: 132659. https://doi.org/10.1016/j.molstruc.2022.132659

[3] 顾耀文, 张博文, 郑思, 杨丰春, 李姣. 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.

[4] 俞青芬. 人工神经网络在吡喃酮类化合物生物活性预测中的应用[J]. 江汉大学学报(自然科学版), 2017, 45(5): 418-423. https://doi.org/10.16389/j.cnki.cn42-1737/n.2017.05.006

[5] 王玉成, 冯志宏, 赵娜娜, 汪鸣明, 叶晓东. 基于 RegNet-1d 模型和积分梯度法的 ERα 拮抗剂的生物活性预测方法[P]. 中国专利, CN114121177A. 2022-03-01.

[6] 沈杰. 药物 ADMET 理论预测方法开发和靶向雌激素受体的药物设计研究[D]: [博士学位论文]. 上海: 华东理工大学, 2011.

[7] 于娜. 常用的特征筛选方法研究[J]. 科技资讯, 2020, 18(36): 231-233.

[8] 李颜平, 吴刚. 基于典型数据集的数据预处理方法对比分析[J]. 沈阳工业大学学报, 2022, 44(2): 185-192.

[9] Hu, L., Gao, L.B., Li, Y.H., Zhang, P. and Gao, W.F. (2022) Feature-Specific Mutual Information Variation for Multi-Label Feature Selection. *Information Sciences*, **593**, 449-471. https://doi.org/10.1016/j.ins.2022.02.024

[10] 王璐, 孙聚波. Lasso 回归方法在特征变量选择中的应用[J]. 吉林工程技术师范学院学报, 2021, 37(12): 109-112.

[11] Naila, S., *et al*. (2020) A Rapid Recognition Method for Rice False Smut Based on HOG Features and SVM Classification. *Journal of Physics*: *Conference Series*, **1576**, Article ID: 012018. https://doi.org/10.1088/1742-6596/1576/1/012018

[12] Li, W. and Jiao, G. (2020) Prediction of Poor Students' Classification Based on Adaboost Algorithm Integrated Learning Model. *Journal of Physics Conference Series*, **1574**, Article ID: 012172. https://doi.org/10.1088/1742-6596/1574/1/012172

[13] 贾鹤鸣, 姜子超, 李瑶. 基于改进秃鹰搜索算法的同步优化特征选择[J/OL]. 控制与决策: 1-9. https://doi.org/10.13195/j.kzyjc.2020.1025, 2021-10-18.

[14] Alsattar, H.A., Zaidan, A.A. and Zaidan, B.B. (2020) Novel Meta-Heuristic Bald Eagle Search Optimisation Algorithm. *Artificial Intelligence Review*, **53**, 2237-2264. https://doi.org/10.1007/s10462-019-09732-5

[15] Shar, P.A., Tao, W.Y., Gao, S., *et al*. (2016) Pred-Binding: Large-Scale Protein-Ligand Binding Affinity Prediction. *Journal of Enzyme Inhibition and Medicinal Chemistry*, **31**, 1443-1450. https://doi.org/10.3109/14756366.2016.1144594