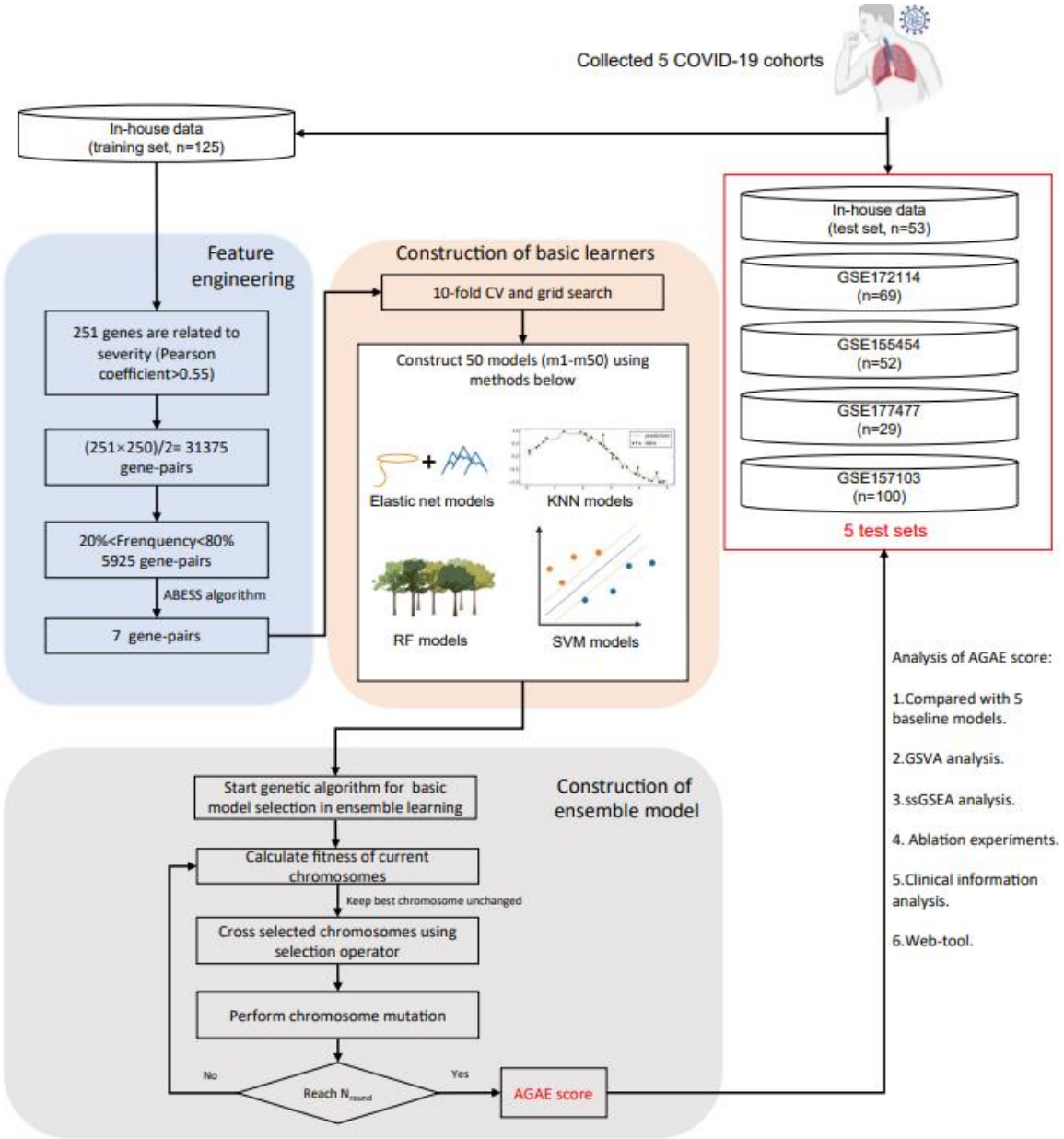
AGAE score：通过最优子集筛选算法和遗传算法辅助的集成学习方法建立了一种稳定的新冠患者严重程度打分

孔维恺忻1，朱杰1，毕素珍2，黄丽婷2，吴鹏3,4，朱素杰2

芬兰赫尔辛基大学转化医学研究所1
青岛大学转化医学研究院2
华中科技大学同济医学院3
华中科技大学同济医学院附属同济医院4

Kong, Weikaixin , Jie Zhu , Suzhen Bi , Liting Huang , Peng Wu , and Su-Jie Zhu . 2023. “Adaptive Best Subset Selection Algorithm and Genetic Algorithm Aided Ensemble Learning Method Identified a Robust Severity Score of COVID-19 Patients.” *iMeta* e126. https://doi.org/10.1002/imt2.126

正文



补充图1. AGAE评分的工作流程

(A)

Algorithm 1: The algorithm of AGAE

Input: Training set, $\mathcal{T}_1$.

Output: The trained AGAE score model $\mathcal{M}_{trained}$.

The feature engineering. Select the signaficant genes whose absolute Pearson coefficient with severity level is greater than 0.55 in in $\mathcal{T}_1$. These genes are labelled as $\mathcal{G}_1$;

Pair the genes of $\mathcal{G}_1$ (**Equation 1**) in $\mathcal{T}_1$ set to avoid batch effect and then select the gene-pairs whose frenquency of "1" label is greater than 0.2 and less than 0.8 in the training set and get the gene-pair set $\mathcal{G}_2$;

Perform ABESS algorithm (**Equation 4-11**) on $\mathcal{G}_2$ in $\mathcal{T}_1$ and then get the gene-pair set $\mathcal{G}_3$;

The construction of basic learners. Construct 4 kinds of severity prediction models. Firstly, perform 10-fold CV and grid search in the $\mathcal{T}_1$ set and then built $\mathcal{N}_{gene}$ models using the whole $\mathcal{T}_1$ set and these models are sorted by MSE value in CV from low to high $(m_1 ... m_N)$;

The construction of AGAE score model using genetic algorithm. Set up an ensemble learning model using the m_1 model parameters. Perform the genetic algorithm in the $\mathcal{T}_1$ set. And set the parameters: $\mathcal{N}_{mg}$, $\mathcal{N}_{round}$, $\mathcal{N}_{chromosome}$;

for *training iterations* $i = 1, ..., N_{round}$ **do**

  Compute current fitness value $f(x)$ by **Equation 13** and keep the chromosome with highest fitness unchanged ;

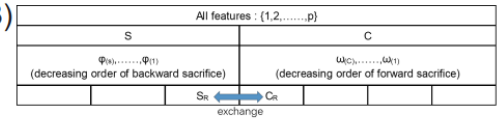
  Selecte chromosomes according to Selection operator (**Equation 14**);

  Cross selected chromosomes ;

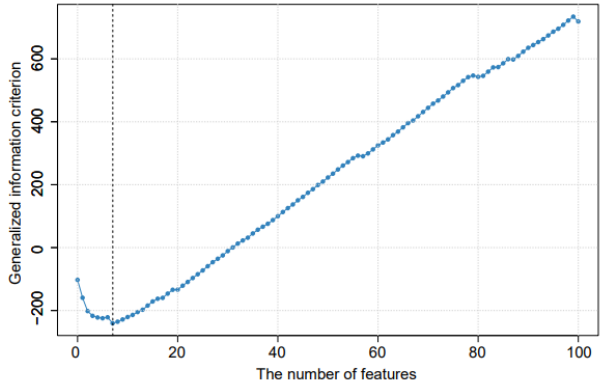
  Generate mutated chromosomes according to $\mathcal{N}_{mg}$ (**Equation 15**);

endfor

(B)



(C)



(D)

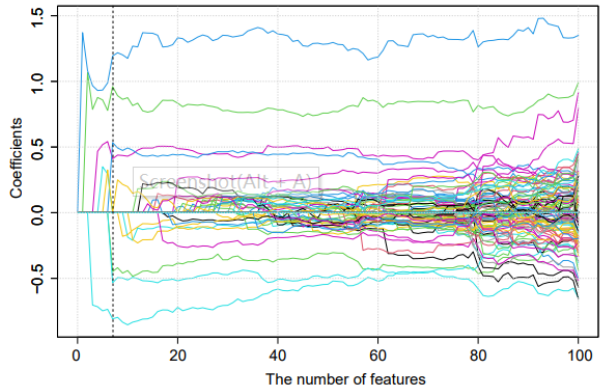
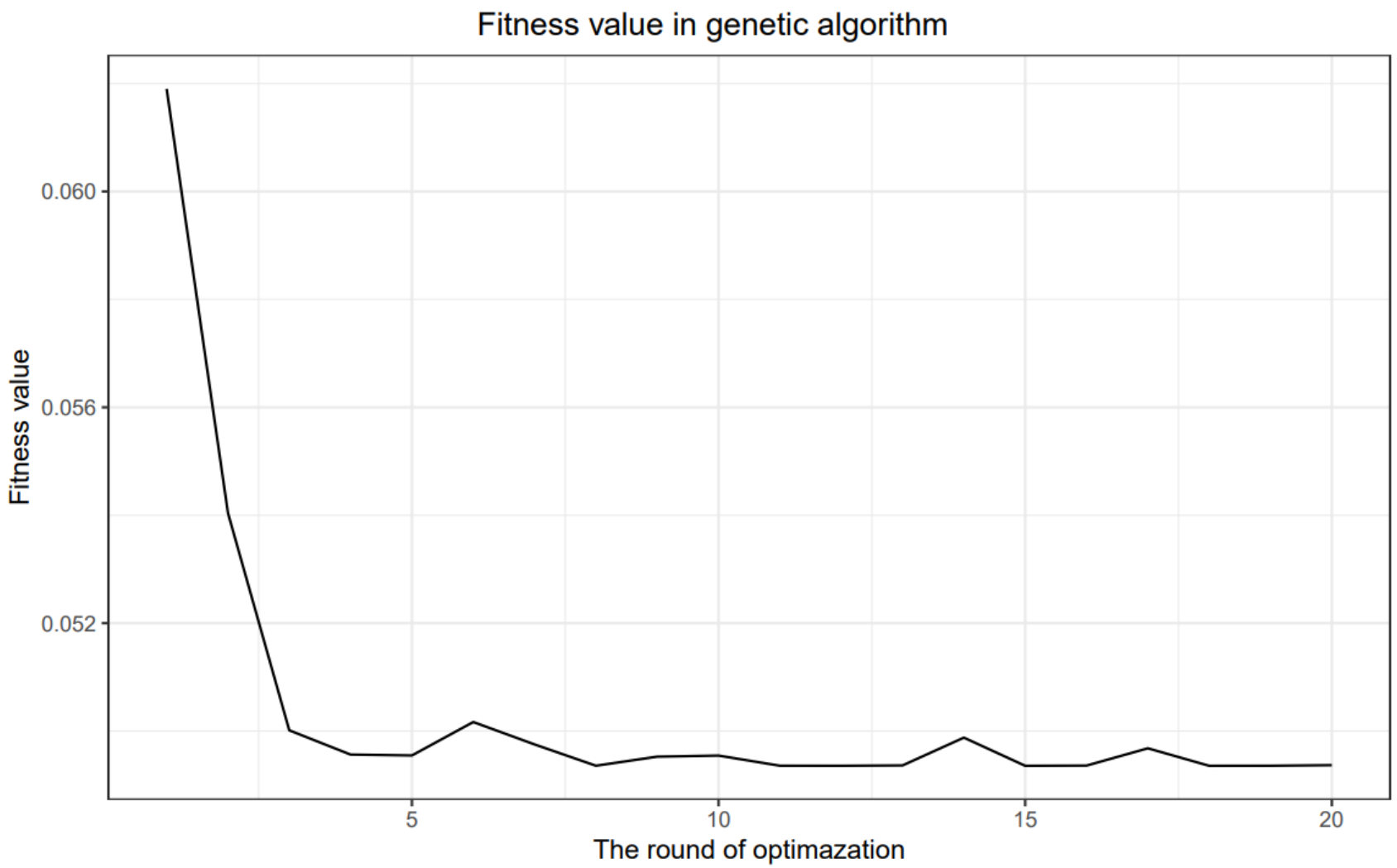


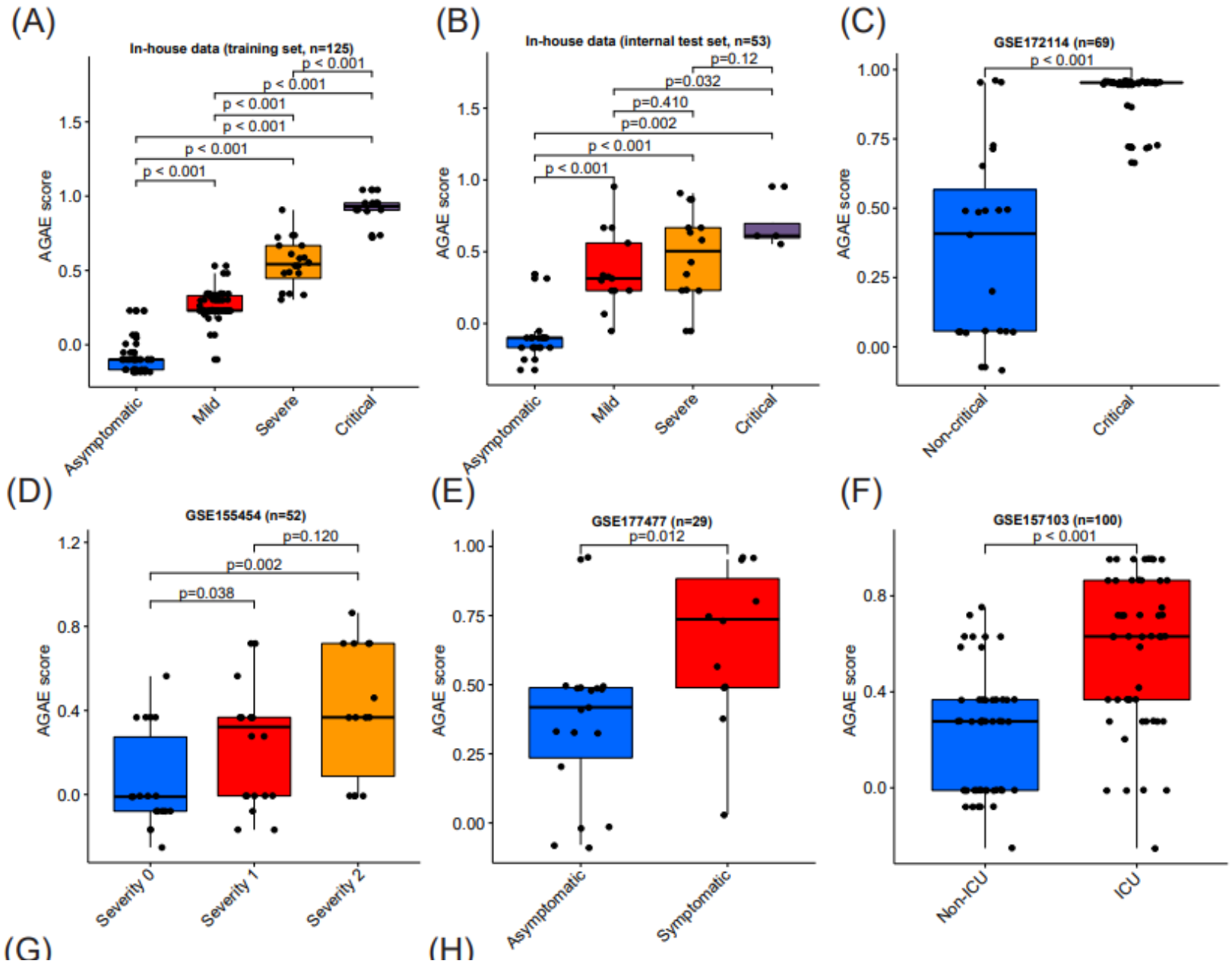
图 1. (A) AGAE 评分算法。 (B) ABESS 算法的图示。 (C) 使用 ABESS 算法选择特征时广义信息准则 (GIC) 的变化。 (D) 选择特征时的特征系数。

正文



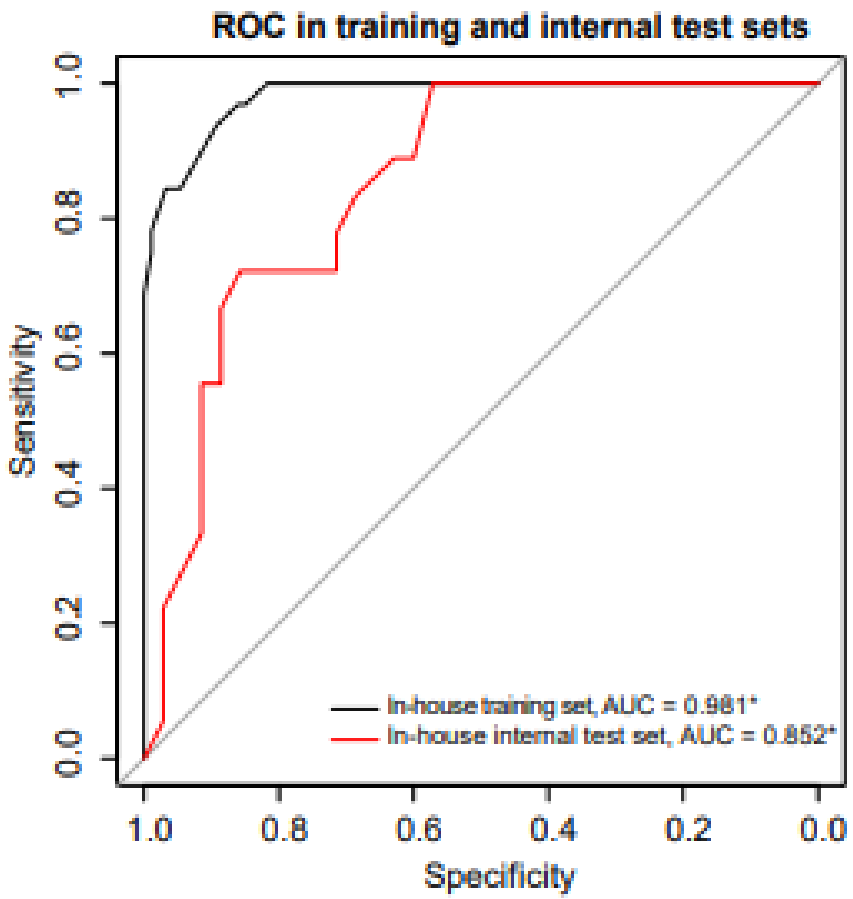
补充图 2. 遗传算法中适应度值的变化。

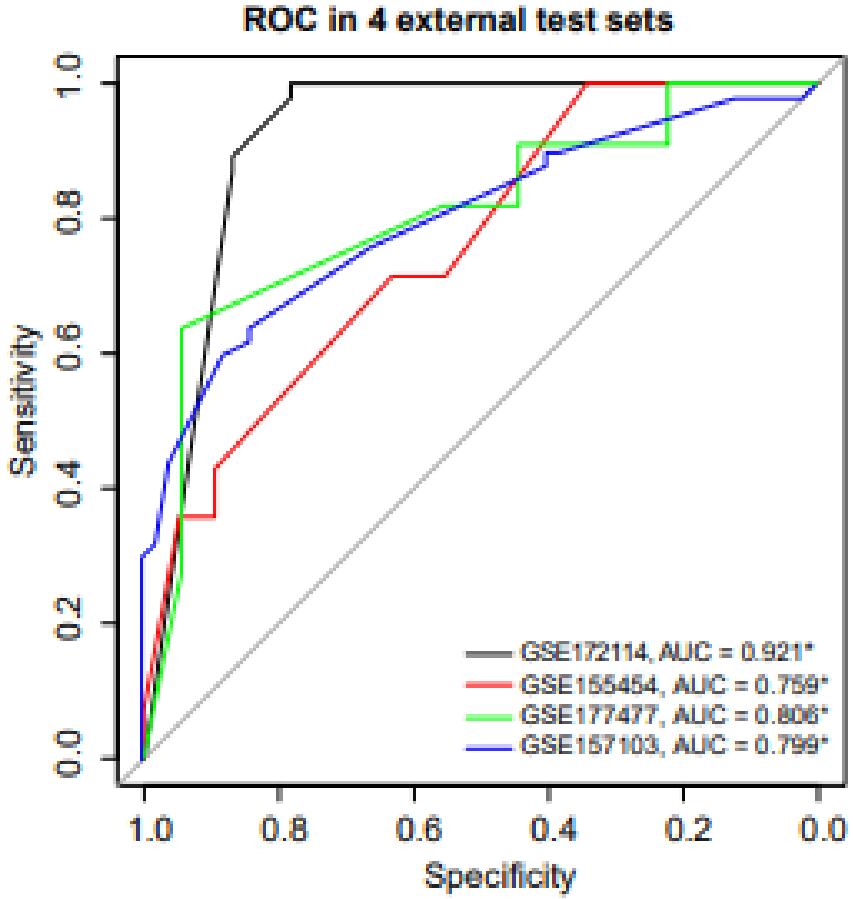
正文



AGAE 得分在训练集和 5 个测试集上的表现。

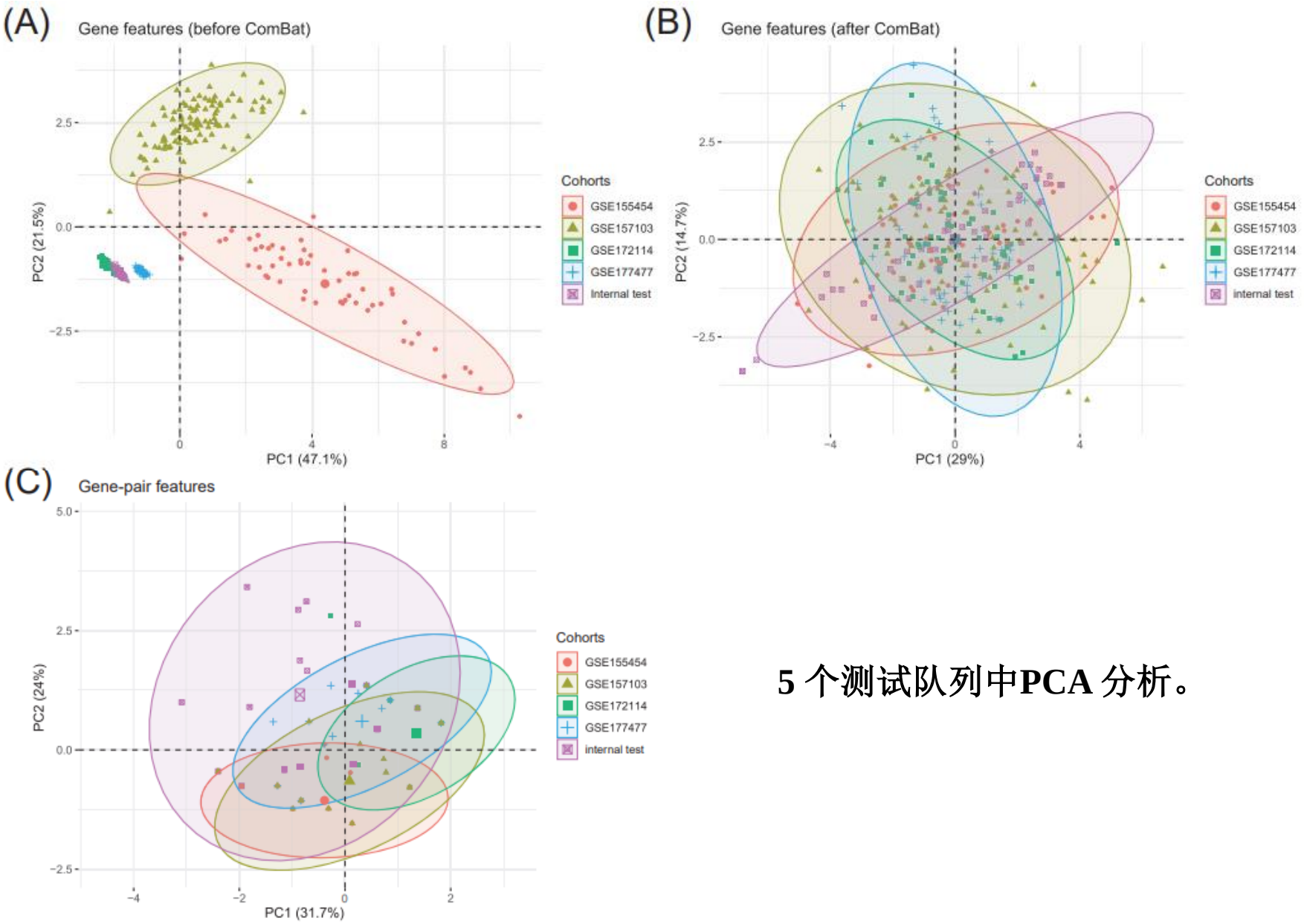
正文





AGAE 得分在训练集和 5 个测试集上的表现。

正文



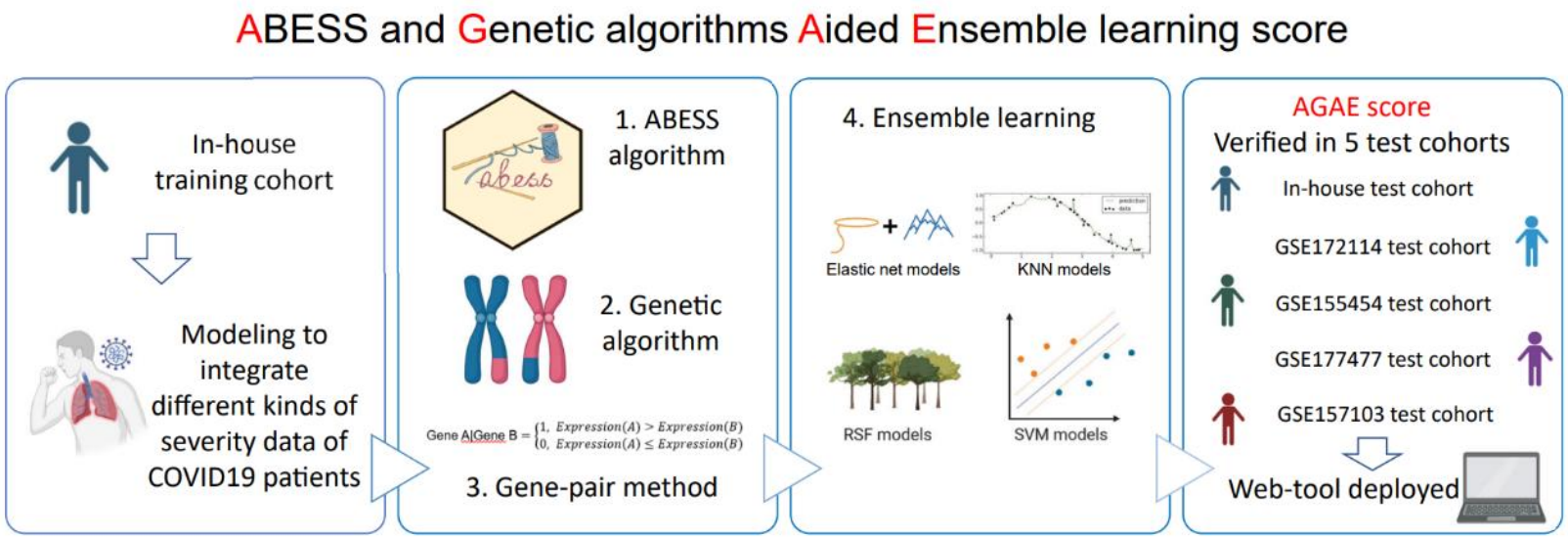
5 个测试队列中PCA 分析。

正文

补充表 4. AGAE 评分的 ROC-AUC 和 5 个测试集中的 5 个基线模型。

Models	AGAE score	m3	m21	m30	m32	m6A score
Methods	Ensemble learning	Elastic net	KNN	RF	KNN	LASSO
Internal test set	0.852	0.852	0.814	0.844	0.815	0.692
GSE172114	0.921	0.893	0.921	0.895	0.889	0.508
GSE155454	0.759	0.731	0.688	0.724	0.703	0.590
GSE177477	0.806	0.699	0.745	0.780	0.747	0.727
GSE157103	0.799	0.793	0.791	0.789	0.779	0.743
Average ROC-AUC	0.827	0.794	0.792	0.806	0.787	0.652
p	-	0.076	0.032	0.008	0.002	0.025

总结



•AGAE score是一种基于集成学习的新冠病人疾病严重程度打分，它在5个测试集中的平均ROC-AUC数值达到0.827

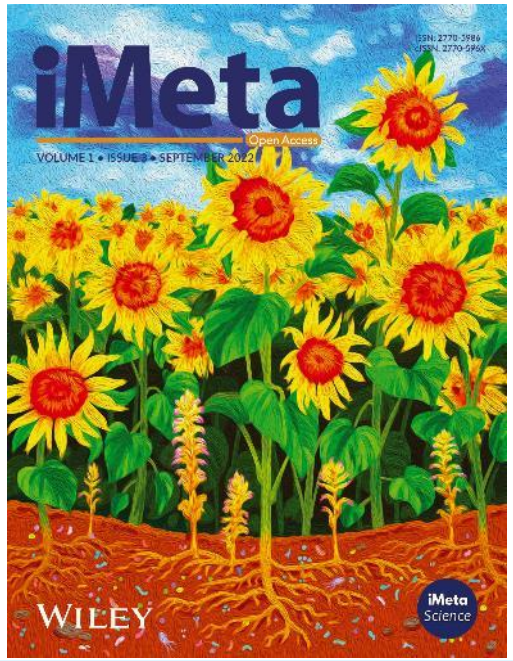
•我们提供了我们的内部数据集，这一数据集被划分为训练集和一个内部测试集。
•AGAE score被部署为网站，方便临床工作者使用“https://kwkxbioinfor.shinyapps.io/COVID19/”

•通过消融实验和PCA，我们发现基因配对方法能够有效消除批次效应，最优子集筛选算法和遗传算法能够有效提升模型的准确性和鲁棒性

Kong, Weikaixin , Jie Zhu , Suzhen Bi , Liting Huang , Peng Wu , and Su-Jie Zhu . 2023. “Adaptive Best Subset Selection Algorithm and Genetic Algorithm Aided Ensemble Learning Method Identified a Robust Severity Score of COVID-19 Patients.” *iMeta* e126. https://doi.org/10.1002/imt2.126

iMeta: 整合宏组学重新认识生命和环境科学

iMeta WILEY



"*iMeta*" 是由威立、肠菌分会和本领域数百位华人科学家合作出版的开放获取期刊，主编由中科院微生物所刘双江研究员和荷兰格罗宁根大学傅静远教授共同担任。目的是发表原创研究、方法和综述以促进宏基因组学、微生物组和生物信息学发展。目标是发表前10%(IF > 15)的高影响力论文。期刊特色包括视频投稿、可重复分析、图片打磨、青年编委、前3年免出版费、50万用户的社交媒体宣传等。2022年的三月、六月和九月期已正式在线出版发行!

主页: http://www.imeta.science
出版社: https://wileyonlinelibrary.com/journal/imeta

投稿: https://mc.manuscriptcentral.com/imeta

office@imeta.science

iMeta 宣传片