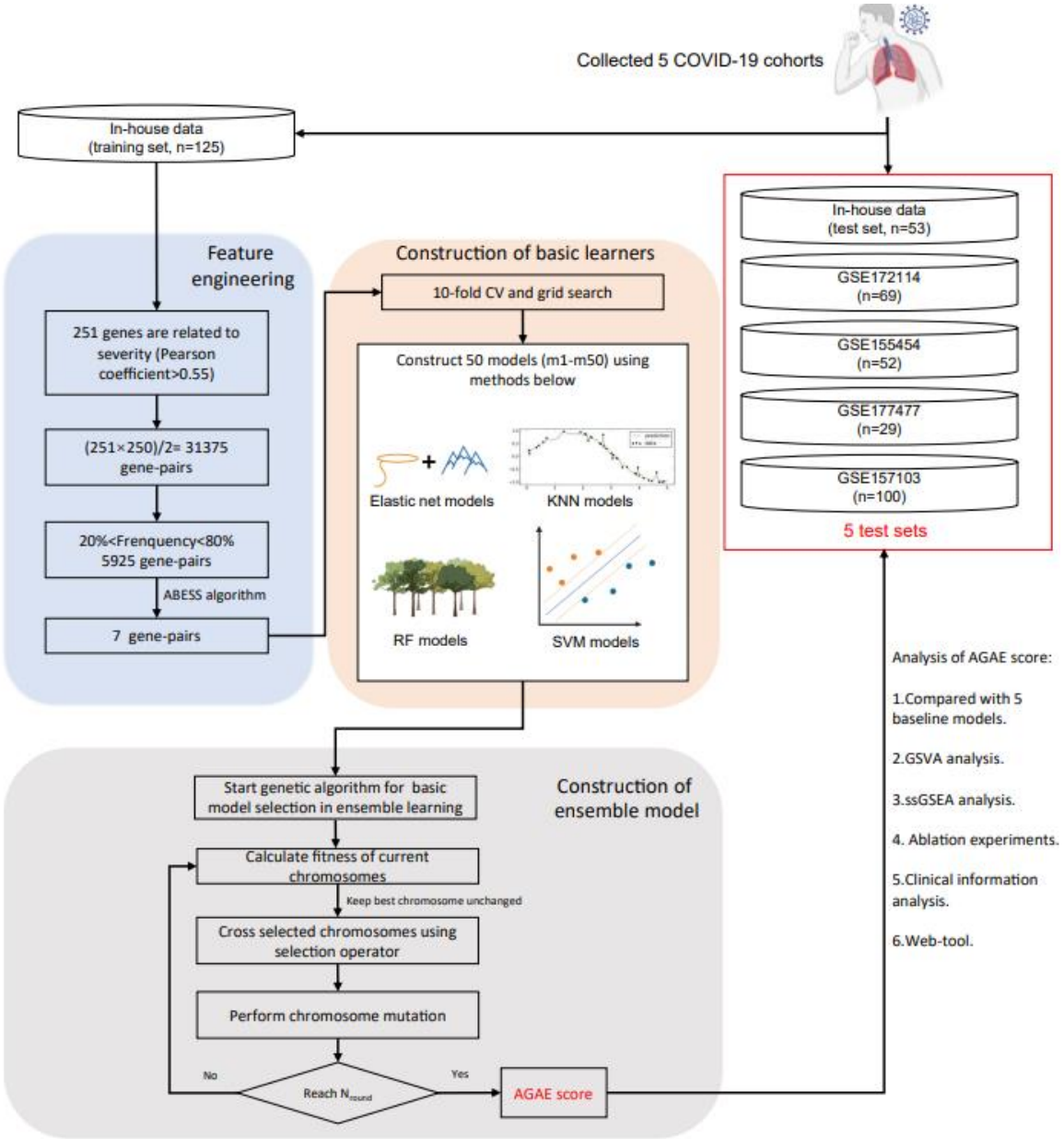
Adaptive Best Subset Selection Algorithm and Genetic Algorithm aided Ensemble Learning Method Identified a Robust Severity Score of COVID-19 Patients

Weikaixin Kong 1, Jie Zhu 1, Suzhen Bi 2, Liting Huang 2,
Peng Wu 3, Sujie Zhu 2

Institute for Molecular Medicine Finland (FIMM) 1
Institute of Translational Medicine, Qingdao University 2
Tongji Hospital, Huazhong University of Science and Technology3

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

Results



SF1. Workflow of AGAE score

(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 ;

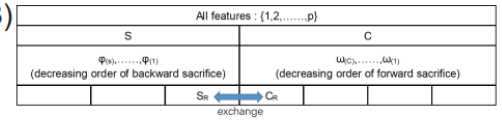
  Selecte chromosomes according to Selection operator (**Equation 14**);

  Cross selected chromosomes ;

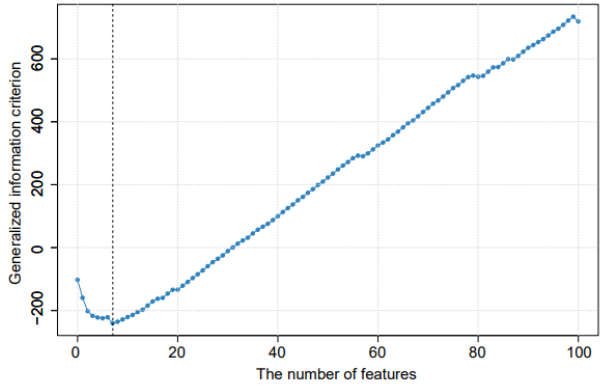
  Generate mutated chromosomes according to $\mathcal{N}_{mg}$ (**Equation 15**);

endfor

(B)



(C)



(D)

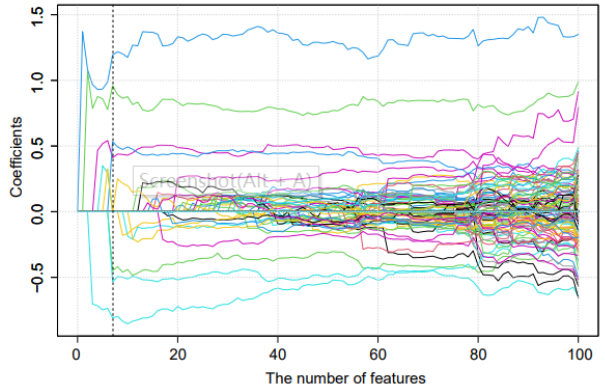
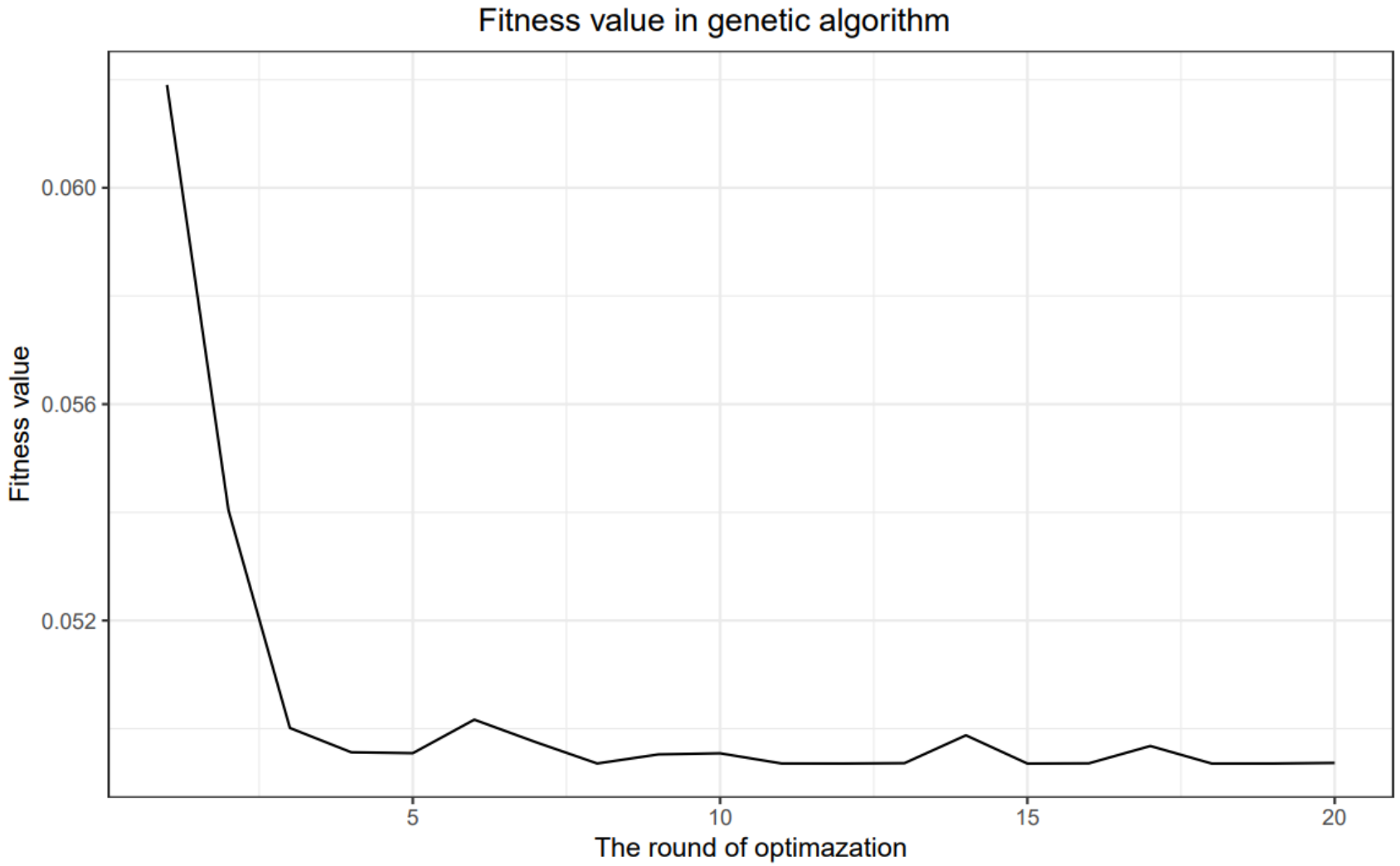


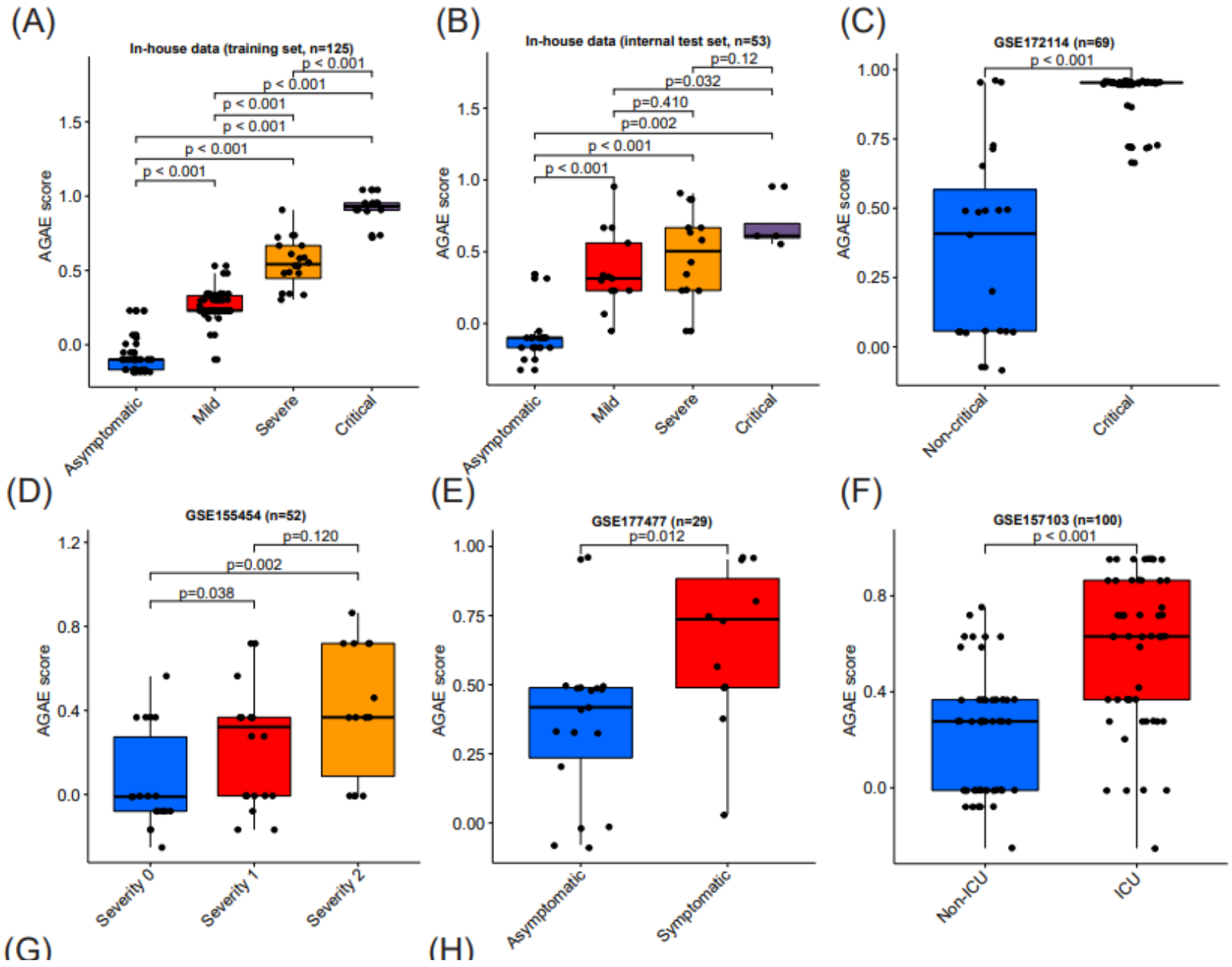
Figure 1. (A) The algorithm of AGAE score. (B) The illustration of the ABESS algorithm. (C) The change in generalized information criterion (GIC) when selecting features using ABESS algorithm. (D) The coefficients of features when selecting features

Results



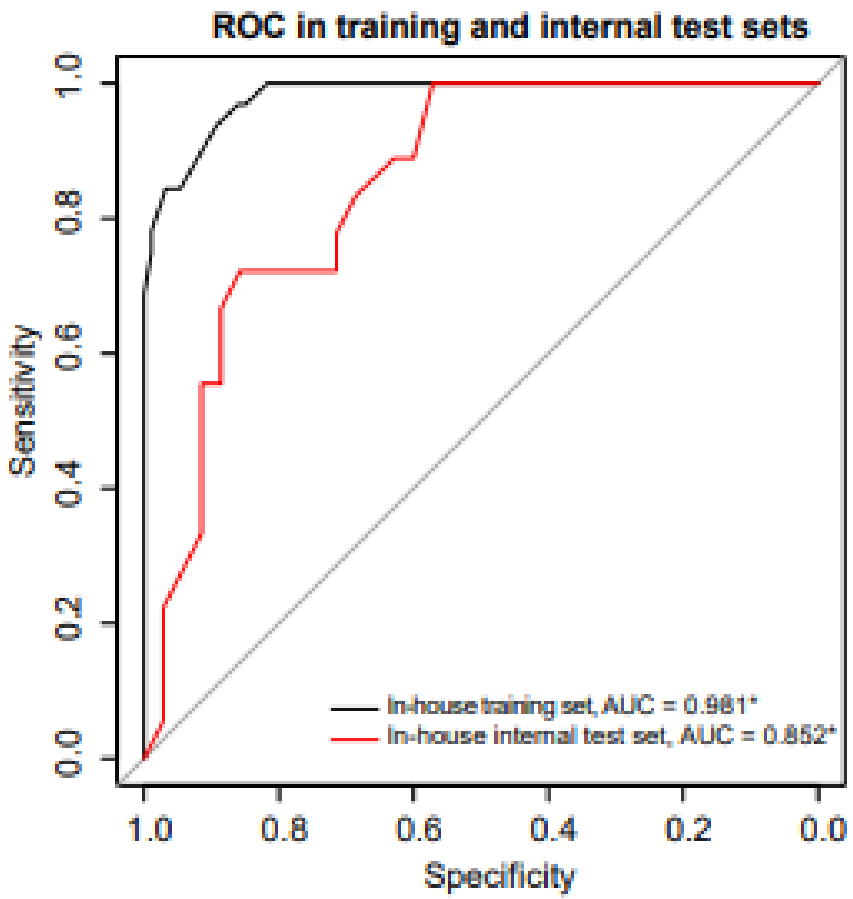
SF2. The change of fitness value in the genetic algorithm.

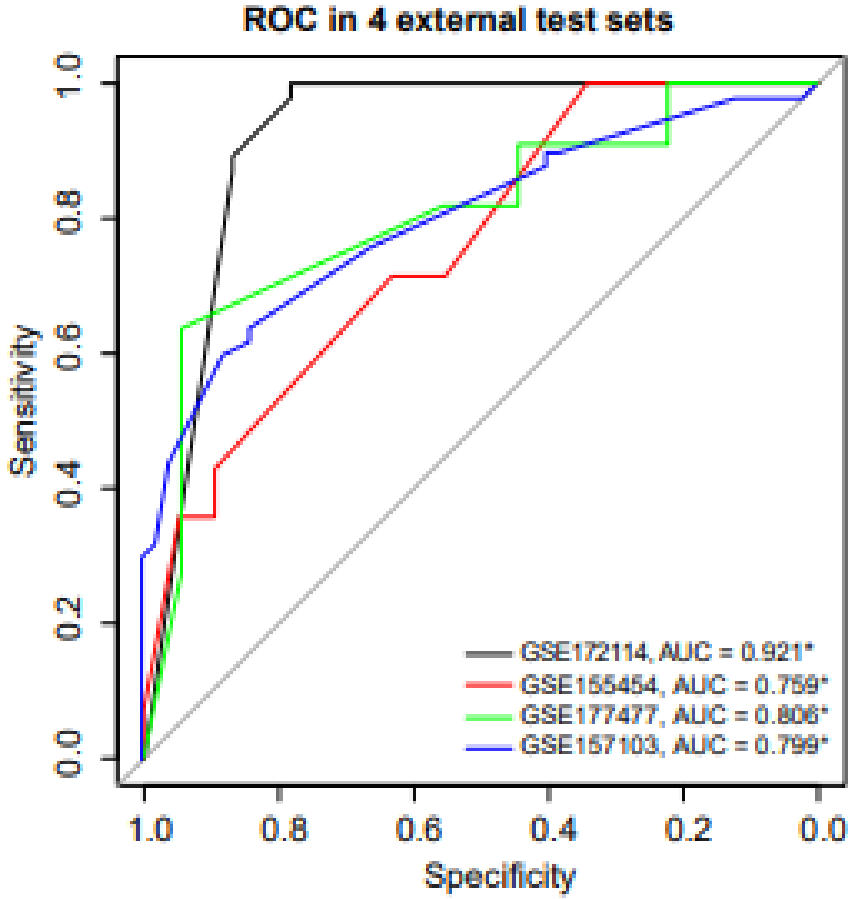
Results



The performance of AGAE score in the training set and 5 test sets.

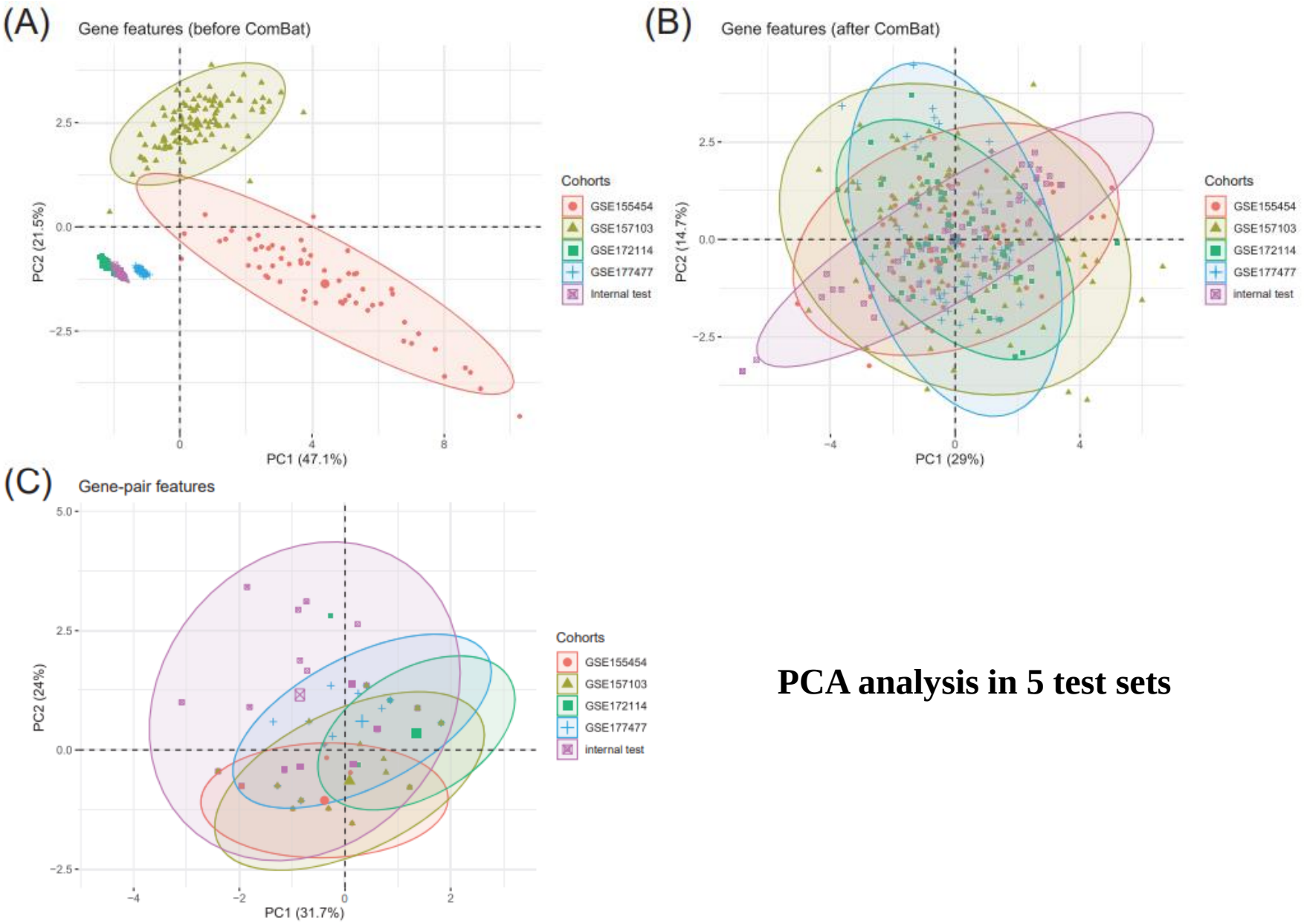
Results





ROCs of AGAE score in the training set and 5 test sets.

Results



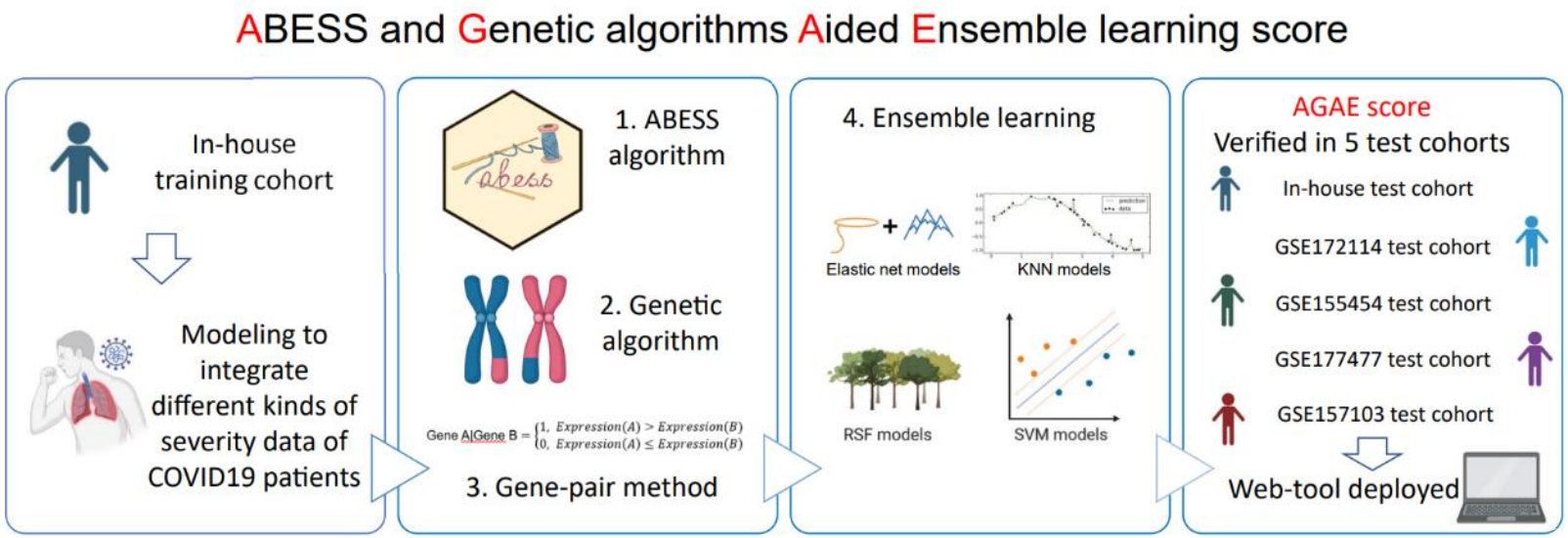
PCA analysis in 5 test sets

Results

ST4. The ROC-AUCs of AGAE score and 5 baseline models in the 5 test sets.

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

Summary

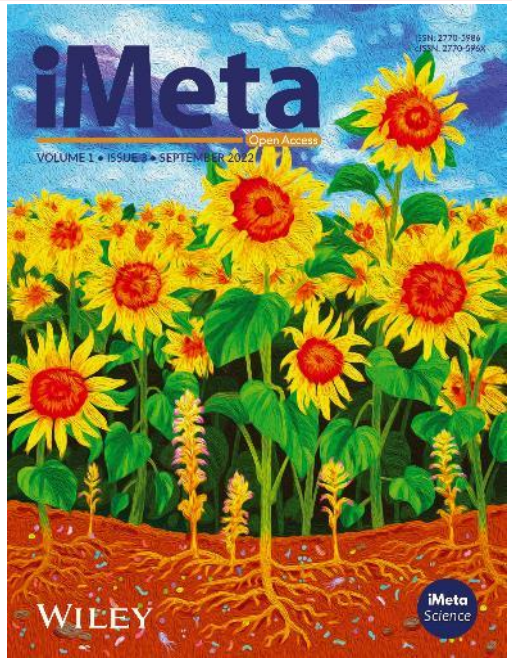


- We have developed a new machine learning ensemble model to predict the severity of COVID-19 cases and achieved an average ROC-AUC of 0.827 in five test sets.
- We provided our in-house COVID-19 cohort (n=178), which was used as a training set (n=125) and an internal test set (n=53).
- The gene-pairing feature extraction method in AGAE score can reduce the degree of batch effect. The adaptive best subset selection algorithm and genetic algorithm can improve the accuracy and effectiveness of the AGAE score.
- An easy-to-use web-tool based on AGAE score was established (https://kwkxbioinfor.shinyapps.io/COVID19/).

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: Integrated meta-omics to change the understanding of the biology and environment

WILEY



"***iMeta***" is an open-access Wiley partner journal launched by scientists of the Chinese Academy of Sciences. iMeta aims to promote metagenomics, microbiome, and bioinformatics research by publishing original research, methods, or protocols, and reviews. The goal is to publish high-quality papers (Top 10%, IF > 15) targeting a broad audience. Unique features include video submission, reproducible analysis, figure polishing, APC waiver, and promotion by social media with 500,000 followers. Three issues were released in March, June , and September 2022.

Society: http://www.imeta.science
Publisher: https://wileyonlinelibrary.com/journal/imeta
Submission: https://mc.manuscriptcentral.com/imeta

office@imeta.science
iMeta

iMetaScience
iMetaScience