

1 **Integration of Machine Learning to Identify Diagnostic Genes in Leukocytes**

2 **for Acute Myocardial Infarction Patients**

3 Lin Zhang^{1*}, Yue Liu^{2*}, Kaiyue Wang¹, Xiangqin Ou³, Jiashun Zhou⁴, Houliang Zhang⁴,

4 Min Huang², Zhenfang Du^{2#}, Sheng Qiang^{2#}

5 ¹ State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional
6 Chinese Medicine, 10 Poyanghu Road, Jinghai, Tianjin 301617, P. R. China.

7 ² Department of Nephropathy, Zhangjiagang TCM Hospital Affiliated to Nanjing University of
8 Chinese Medicine, Zhangjiagang, Jiangsu 215600, P. R. China.

9 ³ The First Affiliated Hospital of Guizhou University of Traditional Chinese Medicine, Guiyang,
10 Guizhou 550025, P. R. China.

11 ⁴ Tianjin Jinghai District Hospital, 14 Shengli Road, Jinghai, Tianjin 301699, P. R. China.

12

13 * **Lin Zhang and Yue Liu contributed equally.**

14 # **Corresponding author**

15 **Sheng Qiang**, Department of Nephropathy, Zhangjiagang TCM Hospital Affiliated to Nanjing

16 University of Chinese Medicine, Zhangjiagang, Jiangsu 215600, P. R. China.

17 Email address: qiangsheng660@163.com.

18 **Zhenfang Du**, Department of Nephropathy, Zhangjiagang TCM Hospital Affiliated to Nanjing

19 University of Chinese Medicine, Zhangjiagang, Jiangsu 215600, P. R. China.

20 Email address: zyydzf@163.com.

21

22 **Abstract:**

23 **Background:** Acute myocardial infarction (AMI) has two clinical characteristics: high missed
24 diagnosis and dysfunction of leukocytes. Transcriptional RNA on leukocytes is closely related to
25 the course evolution of AMI patients. We hypothesized that transcriptional RNA in leukocytes
26 might provide potential diagnostic value for AMI. Integration machine learning (IML) was first
27 used to explore AMI discrimination genes. The following clinical study was performed to validate
28 the results.

29 **Methods:** A total of four AMI microarrays (derived from the Gene Expression Omnibus) were
30 included in this study (220 sample size), and the controls were identified as patients with stable
31 coronary artery disease (SCAD). At a ratio of 5:2, GSE59867 was included in the training set,
32 while GSE60993, GSE62646, and GSE48060 were included in the testing set. IML was explicitly
33 proposed in this research, which is composed of six machine learning algorithms, including
34 support vector machine (SVM), neural network (NN), random forest (RF), gradient boosting
35 machine (GBM), decision trees (DT), and least absolute shrinkage and selection operator
36 (LASSO). IML had two functions in this research: filtered optimized variables and predicted the
37 categorized value. Furthermore, 40 individuals were recruited, and the results were verified.

38 **Results:** Thirty-nine differentially expressed genes (DEGs) were identified between controls and
39 AMI individuals from the training sets. Among the thirty-nine DEGs, IML was used to process
40 the predicted classification model and identify potential candidate genes with overall normalized
41 weights >1 . Finally, Two genes (AQP9 and SOCS3) show their diagnosis value with the area
42 under the curve (AUC) > 0.9 in both the training and testing sets. The clinical study verified the

43 significance of AQP9 and SOCS3. Notably, more stenotic coronary arteries or severe Killip
44 classification indicated higher levels of these two genes, especially SOCS3. These two genes
45 correlated with two immune cell types, monocytes and neutrophils.

46 **Conclusion:** AQP9 and SOCS3 in leukocytes may be conducive to identifying AMI patients with
47 SCAD patients. AQP9 and SOCS3 are closely associated with monocytes and neutrophils, which
48 might contribute to advancing AMI diagnosis and shed light on novel genetic markers. Multiple
49 clinical characteristics, multicenter, and large-sample relevant trials are still needed to confirm its
50 clinical value.

51 **Keywords:** Acute Myocardial Infarction, Diagnostic Gene Identification, Machine Learning,
52 AQP9, SOCS3, Immune Cell Correlation

53
54
55
56
57
58
59
60
61
62
63

64 **1 Introduction**

65 Acute myocardial infarction (AMI), the most severe form of cardiovascular disease, is associated
66 with [1, 2] millions of deaths annually around the world [3, 4]. Generally, the diagnosis of AMI
67 includes clinical syndrome, electrocardiogram, and serum changes in enzyme levels [5]. However,
68 AMI is easily misdiagnosed because of the following three aspects: nonclassic clinical symptoms
69 [6, 7], atypical underappreciation [8], and an untimely serum peak. Because of the above three
70 problems, a previous study [9] reported that the missed diagnosis rate of AMI is higher than 0.9%.
71 The diagnosis and treatment of AMI must be prompt; otherwise, it may trigger irreversible results.
72 Therefore, exploring new markers of AMI to decrease missed diagnoses is essential and urgent.

73

74 Leukocytes play an important and varied role in the entire evolution of AMI. During the acute
75 injury phase of AMI, leukocytes promote a severe inflammatory cascade response through the
76 polarization of M1 macrophages [10]. During the repair phase of AMI, M2 macrophages in
77 leukocytes suppress inflammation and mediate the repair of injured myocardium [11].
78 Furthermore, leukocyte alteration positively correlates with AMI severity and, inversely, with
79 patient survival [12, 13].

80

81 RNAs are involved in the evolution of AMI. For example, miR-155 correlated positively with the
82 concentration of inflammatory cytokines - IL-6 and TNF- α [14] in AMI. Neutrophil-derived
83 S100A8/A9 amplify granulopoiesis and cardiac injury in AMI mice [15]. Conversely, M2
84 macrophage-derived exosomes carry miR-1271-5p [16] to alleviate AMI-related cardiac injury. In

85 conclusion, RNA on leukocytes plays a different role in the evolution of AMI, possibly related to
86 different leukocyte subtypes. However, numerous studies have focused on integrating target
87 interventions [12, 17] and leukocyte complications [17, 18]. Few studies have focused on the
88 diagnostic value of leukocytes' RNA. Because the leukocytes' RNA is involved in the evolution of
89 AMI, these RNA might have diagnosing value for AMI patients. The diagnosis value might be
90 related to various leukocyte subtypes.

91

92 Machine learning (ML) helps humans learn patterns from complex data to predict future
93 behavioural outcomes and trends. ML was widely utilized in variable filtering. A previous study
94 used a single ML algorithm or two integrated ML algorithms (e.g., support vector machine [18] or
95 least absolute shrinkage and selection operator [19]) to optimize variables. Still, these approaches
96 may have missed potential genes [20]. Compared with a single ML algorithm, the integrated ML
97 (IML) approach [21-23] we developed is more advantageous in variable screening and model
98 building. IML helps identify potential genes mistakenly deleted by a single ML and find more
99 meaningful variables [21]. IML integrates the advantages of a single ML, and its predictive
100 classification value is better [23]. Based on a favourable filtration value in transcriptomics of IML,
101 IML might be used to comprehensively explore the diagnostic value in AMI patients.

102

103 In summary, we aim to explore the potential diagnostic value of transcriptome within leukocytes
104 for identifying AMI patients. Because of IML's good variable screening and excellent predictive
105 value, IML was first used to mine diagnostic genes in AMI leukocytes with multiple microarrays.

106 Single microarray data might have inherent biases in capturing the entire transcriptomic landscape,
107 so multiple microarrays are integrated after resolving batch effects to reduce bias and validate
108 each other. And clinical validation was added to confirm the result. The relationship between
109 transcriptome and leukocyte subtypes was unclear, so the correlation between immune cells and
110 target transcriptome was subsequently accomplished. We expect to explore the functional roles of
111 the identified genes in AMI pathophysiology, investigating their potential as therapeutic targets.

112

113 **2 Methods**

114 **2.1 Data acquisition**

115 The raw data were obtained from the Gene Expression Omnibus (GEO, March 27, 2022). AMI
116 patients have similar symptoms to SCAD patients, which were set as the controls. An increasing
117 leukocyte may influence the result of other cardiovascular diseases (*e.g.*, stroke [24, 25] and
118 heart failure [26]), to be excluded. AMI is easily misdiagnosed as SCAD. Leukocytes are also
119 altered in other cardiovascular diseases. Based on the above, the following inclusion and
120 exclusion criteria were set: I) inclusion criteria—(i) diagnosed as AMI patients on admission; (ii)
121 transcriptome was obtained from leukocytes in blood; (iii) initial data were free and accessible;
122 and (iv) the control individuals were diagnosed with stable coronary artery disease (SCAD); and
123 II) exclusion criteria—(i) other cardiovascular diseases suspected and (ii) blood were taken more
124 than one day after hospitalization.

125

126 **2.2 Data processing**

127 To ensure the reliability of the data, the R package *sva* (version 3.46.0) was applied to data
128 integration to minimize the batch effects with the *ComBat* function and parametric adjustments.
129 Regarding the distribution ratio of previous literature (1.64:1 [27] to 5:1 [28]) and to minimize the
130 branching effect, this research was distributed in the training or testing sets at a ratio of 5:2.
131 GSE59867 was included in the training set. In contrast, GSE60993, GSE62646, and GSE48060
132 were included in the testing set. In brief, the training set was applied to explore candidate
133 diagnostic genes, and the testing set was used for validation. Based on the differential DEGs,
134 three functional enrichment analyses were developed via the Kyoto Encyclopedia of Genes and
135 Genomes Gene Set Enrichment Analysis (KEGG-GSEA), Gene Ontology (GO), and Disease
136 Ontology (DO). In addition, the GO terms included three branches: molecular function (MF),
137 biological process (BP), and cellular components (CC). Notably, the novel IML served two
138 functions: developing classification ML and exploring the candidate variable. Finally, the above
139 candidate genes were verified in the testing group and clinical study, and an immune analysis
140 among the candidate genes was performed. CIBERSORT was processed for immune correlation
141 analysis in the *corrplot* R package (version 0.92). And the primary code was link with
142 <https://github.com/Linzhang-BiuBiuBiu/ML-for-diagnosis-genes..git>.

143

144 **2.3 Searching for DEGs**

145 Because the same gene may have multiple sequences, the transcriptome will appear to have
146 several expression data for the same genes. For the same genes, *limma* (version 3.54.0) was
147 employed to identify the DEGs with the average gene expression. According to the Benjamini and

148 Hochberg method, two thresholds were established: an absolute value of fold change
149 ($|\log FC| > 0.7$ (previous studies were 0.5 [29]-1 [23]) and a false discovery rate [30] < 0.05 .

150

151 **2.4 IML of six ML algorithms**

152 Classification models of IML, composed of six ML algorithms, were processed, covering support
153 vector machine (SVM), neural network (NN), random forest (RF), gradient boosting machine
154 (GBM), decision trees (DT), and least absolute shrinkage and selection operator (LASSO). In
155 brief, IML was used to identify candidate genes with the overall normalized weights. The six ML
156 algorithms were developed to optimize parameter settings, model development in the training sets,
157 and validation in the testing sets. For stability, all ML algorithms were tenfold cross-validated.
158 Notably, an accuracy value was applied to evaluate the predictive classification value, and a
159 higher accuracy value showed a better classification value of the six ML algorithms.

160

161 LASSO [31] minimizes the sum of squares of the residuals when the sum of the absolute values
162 of the regression coefficients is less than a constant, producing specific regression coefficients
163 equal to 0 and filtering variables. LASSO was processed with the *glmnet* (version 4.1-6) R
164 package. *cv.glmnet* was utilized to majorize lambda. For the parameters, the scale of "lambda"
165 was set between 0 and 100 with "binomial" and "class". Based on the minimum lambda, *glmnet*
166 was processed to the LASSO with alpha and a "binomial" method in training sets.

167

168 SVM aim to find the separating hyperplane [32] that divides the dataset correctly with the largest

169 geometric interval. SVM was developed with the *e1071* R package (version 1.7–12). *tune.svm*
170 was adopted to optimize the settings parameter with the kernel of "linear", and the cost between 1
171 and 20.

172

173 DT [33] is based on a tree structure that judges (one or more) sample attributes sequentially, from
174 top to bottom, up to the leaf nodes of the decision tree and derives the final result. DT was
175 processed with *rpart* (version 4.1.19) and *rpart.plot* (version 3.1.1). Based on the "class" method
176 and a cp value of 0.001, the *rpart* function was adopted for the DT model.

177

178 RF use a " bagging " technique [34] to construct complete decision trees in parallel by randomly
179 self-sampling dataset samples and features. RF was completed with the R package *randomForest*
180 (version 4.7-1.1). First, the *tuneRF* function was adopted to optimize 0-700 trees with one step
181 size. RF was developed based on the minimum error rate to optimize the number of trees.

182

183 NN outputs model [35] by inputting multiple nonlinear models and weighted interconnections
184 between various models. NN was processed with *neuralnet* (version 1.44.2) with *neuralnet*
185 function, five layers (an input, an output, and three hidden layers), err.fct of "sse", and the linear.

186

187 GBM serially generates a series of weak learners [36], which are directly used to form the final
188 model by combining them. Compared with the other 5 ML algorithms, GBM processed more
189 steps and was prone to making mistakes. The GMB was developed with *h2o* (version 3.38.0.1).

190 First, the Java operating environment was installed, which is the virtual environment of GBM.
191 Essential for running the memory setting in *h2o.init*, the model memory of GBM was adjusted to
192 8G. The h2o data type in GBM was inevitable, and the *as.h2o* function was utilized to transform
193 the data format. Next, *h2o.gbm* tuned the parameters and developed the model with the
194 "Bernoulli" distribution, 200 trees, a learning rate of 0.001, and a sample rate of 90%.

195

196 Furthermore, with the weights of the above six ML algorithms in DEGs, the normalized sum
197 weight of IML was calculated as follows: overall weights = $\text{abs(RF)}/\text{abs(RFmax)} +$
198 $\text{abs(SVM)}/\text{abs(SVMmax)} + \text{abs(LASSO)}/\text{abs(LASSOmax)} + \text{abs(NN)}/\text{abs(NNmax)} +$
199 $\text{abs(GBM)}/\text{abs(GBMmax)} + \text{abs(DT)}/\text{abs(DTmax)}$. For instance, if the weight of interleukin-6 in
200 six ML algorithms was 30, -22, 20, -2, 320, and -8, the maximum absolute value weights in the
201 six ML algorithms were 60, 88, 80, 8, 640, and 16. Therefore, the overall weight of interleukin-6
202 was $|30/60| + |-22/88| + |20/80| + |-2/8| + |320/640| + |-8/16| = 2.25$. With normalized overall weights >1,
203 the candidate genes were estimated by the area under the curve (AUC).

204

205 **2.5 Clinical validation**

206

207 The clinical trial was performed according to the Declaration of Helsinki guidelines. All AMI and
208 SCAD patients provided individual written informed consent from October 10 2022, to December
209 31 2022, and the Ethics Review Committee of Jinghai District Hospital approved the study. There
210 was no increase in the cost of treatment for the patients, no addition of other intervention in the

211 treatment, and the blood samples used were taken from the discarded blood of the patients after
212 their routine blood tests on the same day. If the patient did not have a routine blood test on that
213 day, then the patient was excluded. All AMI patients underwent coronary angiography, and blood
214 samples were collected in anticoagulant tubes on admission. Density gradient centrifugation [37,
215 38] was performed for leukocyte isolation. In brief, 8 mL of Ficoll solution was added to 8 mL of
216 anticoagulated blood, and the upper plasma layer was discarded after centrifugation. The white
217 cell layer at the isolate's junction was aspirated, added to 10mL of saline, and centrifuged; the
218 bottom layer was the leukocytes. RNA, isolated from leukocytes, was synthesized with reverse
219 transcription kits (Takara, Shiga, Japan). Quantitative reverse transcription PCR was executed on
220 an ABI7900HI (Thermo Fisher Scientific). According to previous literature, the relative content of
221 the candidate genes was scaled to the reference gene (GAPDH [39]), and **Table 1** lists the primer
222 sequences.

223 **Table 1** List of primers for real-time PCR analysis in GAPDH, AQP9, and SOCS3.

224 **3 Results**

225 **3.1 Included datasets**

226 A total of 4 datasets (**Table 2**) (220 sample sizes), namely, GSE59867, GSE60993, GSE62646,
227 and GSE48060, were integrated for this study. The training set was obtained from GSE59867 (46
228 controls and 111 AMI patients) based on a raw ratio of 5:2. Furthermore, the testing set was
229 integrated with the other three datasets (28 controls and 35 AMI patients), namely, GSE60993,
230 GSE62646, and GSE48060. The following analysis is presented in **Fig 1**.

231

232 **Table 2** Fundamental information in the four datasets.

233

234 **Fig. 1** The workflow of this study contains four parts: GEO datasets for training and testing sets, machine
235 learnings for classification and variable filtration, diagnosis value verification, and immune correlation.

236 **3.2 DEG identification**

237 Thirty-nine DEGs were identified (**Table S1**) in a training set from 17,049 RNAs. Compared to
238 the control group (SCAD), 28 genes were upregulated (SOCS3, HP, ECRP, AQP9, FAM20A,
239 CES1, STAB1, NRG1.1, NRG1, DYSF, RNASE1, RNASE2, ASGR2, CYP1B1, MERTK,
240 FCGR1A.2, MIR21, FCGR1A.1, TCN2, VSIG4, PPARG, FCGR1A, SLED1, S100A9, FMN1.1,
241 CD163, TMEM176A, and SERPINB2) and 11 genes were downregulated (KLRC3, KLRD1,
242 KLRA1P, DTHD1, KLRC4, MYBL1, CLC, KLRC2, KLRC4-KLRK1, SNORD20, and
243 SNORD45B) in AMI individuals (**Fig. 2**).

244

245 **Fig. 2** Heatmap and volcano plot of 39 DEGs in the AMI and control groups. **A** Red in the heat map indicates high
246 expression, and a blue indicates low expression. **B** Green in the volcano map suggests lower expression, and red
247 indicates high expression.

248 **3.3 Functional analysis**

249 Based on the above DEGs, 45 GSEA terms (**Table S2**) were identified, and the top 5 are shown in
250 **Fig. 3A-B**; 160 GO terms (**Table S3**) were identified, and the top 5 are shown in **Fig. 3C**; and the
251 top 10 of 57 DO terms (**Table S4**) are shown in **Fig. 3D**. In GSEA-KEGG of AMI, the top 3 were
252 Fc gamma R-mediated phagocytosis, Huntington disease, and Leishmania infection. In GO, the

253 top 3 in BP were the stimulatory C-type lectin receptor signalling pathway, response to lectin, and
254 cellular response to lectin. In DO terms, the top 3 were atherosclerosis, arteriosclerotic
255 cardiovascular disease, and arteriosclerosis.

256

257

258 **Fig. 3** Functional analysis of GSEA, GO, and DO terms. **A** The top 5 GSEA-KEGG pathways in controls. **B** The
259 top 5 GSEA-KEGG pathways in AMI patients. **C** The top 5 GO terms in BP, CC, and MF. **D** The top 15 DO terms.

260

261 **3.4 IML of six ML algorithms**

262 Six ML algorithms (**Fig. 4**) and their accuracies (**Table 3**) were assessed. Eight genes were
263 identified in LASSO (**Fig. 4A**), and the training and testing sets' accuracy value was 70.70%
264 (**Table 3**). In SVM, 13 genes were filtered (**Fig. 4B**), and the accuracies were 88.46% and 91.84%,
265 respectively. The error rate of RF (**Fig. 4C**) decreased with an increasing number of trees. Until
266 161 trees, the error rate of RF was minimized, and the accuracy of the two sets was 98.09% and
267 100%. In DT (**Fig. 4D**), the gene expression of 9.8 in AQP9 could discriminate the control and
268 AMI groups, while the accuracies were unstable, 94.27%, and 75.52%. In GBM (**Fig. 4E**), 6-fold
269 methods were established to optimize the diagnosis genes, but unstable accuracies, such as the
270 above ML algorithms, were 93.30% and 85.71%. In the NN (**Fig. 4F**), although sufficient for
271 discriminating the controls and AMI patients with three hidden layers, the accuracy was either
272 83.74% or 71.43%. Among the above ML algorithms, the raw weights of 39 DEGs were
273 identified (**Table S5**). Interestingly, RF had the highest and most stable accuracy value among all

274 ML algorithms. The normalized overall weights (**Table 4**) were calculated to filter the candidate
275 variables. Twenty-six genes (ASGR2, SOCS3, AQP9, PPARG, RNASE1, DYSE, S100A9,
276 FCGR1A, VSIG4, STAB1, MYBL1, KLRD1, ECRP, TCN2, FAM20A, MERTK, HP, RNASE2,
277 DTHD1, CLC, SNORD20, CD163, NRG1, SNORD45B, CYP1B1, and KLRC2) were identified
278 because of overall weights >1 (**Table 4**).

279

280 **Table 3** Accuracy of six MLs based on 39 DEGs in the training and test sets.

281

282 **Fig.4** Six ML algorithms for classification with 39 DEGs. A LASSO for eight candidate genes and the error bars
283 mean the fluctuation range of Binomial Deviance; B SVM for 13 candidate genes. C RF discriminated between
284 the control and AMI groups. And The red, black, and green lines represent the Con, out-of-bag (OOB), and AMI
285 groups respectively. D DT discriminated between the control and AMI groups. E A 6-fold GBM submodel was
286 constructed. The heat map illustrates the importance of genes in each respective submodel. The intensity of the
287 color corresponds to the significance of the gene in the particular submodel. F NN discriminated between the
288 control and AMI groups. All 39 DEGs were involved in modeling in NN, and there are ten because of space
289 limitations. If an edge is colored red, it indicates a positive correlation, meaning that the current feature positively
290 affects the classification result. Conversely, if the edge is gray, it implies a negative correlation. Furthermore, the
291 thickness of the edge signifies the weight's magnitude.

292

293 **Table 4** Overall weights of six classification models were constructed to optimize the candidate diagnostic genes.

294

295 With the basis of overall normalized weights >1 , 26 candidate genes were filtered for subsequent
296 diagnosis in AMI and control groups in the training and testing sets. Among the 26 genes, 10 were
297 excluded because of no differentiation in the testing set. Sixteen genes were significant in the two
298 sets (**Fig. 5**).

299

300 **Fig. 5** The 16 DEGs also differed in the testing set.

301

302 **3.5 Diagnosis value of candidate genes**

303 Sixteen candidate genes were included in the following ROC analysis. The AUC values of
304 SOCS3, AQP9, and ASGR2 were greater than 0.85 in both the training and testing sets. In
305 particular, 2 genes, SOCS3 and AQP9, were greater than 0.9 (**Fig. 6**). The AUC value of the two
306 genes indicated a potential diagnostic value in AMI.

307

308 **Fig. 6** ROC curves for AQP9, SOCS3, and ASGR2 in the training and testing sets.

309

310 **3.6 Correlation analysis**

311 Immune correlation was performed with the 220 samples (**Fig. 7**). The infiltration landscape (**Fig.**
312 **7A**) showed 22 immune distributions in the control and AMI groups. Nine types of immune cells
313 (CD8 T cells, naive CD4 T cells, regulatory T cells (Tregs), resting NK cells, monocytes, M0
314 macrophages, M2 macrophages, eosinophils, and neutrophils) infiltrated significantly between the
315 control and AMI groups (**Fig. S1**). Moreover, the correlations between 22 immunized cells and

316 the two diagnostic genes, AQP9 and SOCS3, based on *Spearman* analysis (**Fig. 7B-C**) showed
317 significant correlations with 9 immune cells (monocytes, neutrophils, CD8 T cells, resting NK
318 cells, naive CD4 T cells, eosinophils, M2 macrophages, activated dendritic cells, and memory B
319 cells). More importantly, two immune cell types (monocytes and neutrophils) possessed a higher
320 correlation coefficient (**Fig. 7B-C**) than the other 7 immune cell types (**Fig. S2-S3**). In particular,
321 the correlation coefficients of monocytes (**Fig. 7B-C**) were highest for the two genes (0.56 for
322 SOCS3 and 0.76 for AQP9).

323

324 **Fig. 7** Immune correlation analysis of AQP9 and SOCS3 between the control and AMI groups. **A** The stacked
325 column graph between the control and AMI groups. **B** The violin plot showed 7 immune cell types infiltrated
326 differently between the control and AMI groups. **C** The lollipop map of the different immune cell types in AQP9
327 and SOCS3. * mean <0.05, ** mean <0.01, ***mean<0.001.

328

329 **3.7 Clinical validation**

330 Finally, 40 individuals (20 SCAD and 20 AMI patients) were recruited. The general information of these
331 individuals was shown in **Table 5**. Among 39 clinical characteristics were summarized, and 13 had significance
332 between the SCAD and AMI patients, including WBC, NeP, MonP, Lym, GAT, D-dimer, CRP, SOCS3, AQP9,
333 LDH, cTnT, CK-MB, and Albumin.

334

Table 5 The general characteristics of the 40 patients.

335 The relative RNA levels (**Fig. 8A**) of AQP9 and SOCS3 were both significant. The SOCS3 content of coronary
336 arteries differed by the number of lesions (**Fig. 8B**): three lesions showed significantly higher SOCS3 than two

337 and one (Fig. 8B). Patients with III-IV Killip classification had higher SOCS3 compared to those with I-II (Fig.
338 8C). Although more stenotic coronary arteries were associated with higher levels of AQP9, the difference was
339 less significant than for SOCS3 (Fig. 8B). In addition, different Killip classifications associated with AQP9
340 possessed no significant differences (Fig. 8C). Furthermore, the 9 significant clinical features were analysed
341 with *Pearson* correlation test (Fig. S4). And SOCS3 had a positive correlation with AQP9. Both genes had a
342 negative correlation with Albumin.

343

344 **Fig. 8** Relative RNA levels of AQP9 and SOCS3 in AMI patients and SCAD controls. **A** The relative content of
345 SOCS3 and AQP9 in AMI patients and SCAD controls. **B** The comparison of AQP9 and SOCS3 in the number of
346 coronary arteries with different stenoses in AMI. **C** The comparison of AQP9 and SOCS3 in various Killip
347 classifications in AMI. * mean <0.05, ** mean <0.01, *** mean<0.001, ns mean no significance.

348

349 **4 Discussion**

350

351 To our knowledge, our work is the first to filter AMI diagnosis genes based on the overall
352 normalized weights of IML. Four microarrays with 220 samples were adopted for data analysis,
353 and further clinical studies were performed to validate the results. Two genes, AQP9 and SOCS3,
354 showed an AUC >0.9 in both the training set and testing set (Fig. 6). Both genes showed a typical
355 and highest correlation coefficient (Fig. S2-3) in monocytes. The clinical study verified the
356 significance between AMI patients and healthy controls, indicating a potential diagnostic value of
357 AQP9 and SOCS3. Compared with previous studies, we reached similar conclusions that AQP9

358 presented diagnostic value for AMI [40, 41], and we further explored the immune correlation of
359 AQP9. Additionally, Prof. Zhu [42] identified SOCS3 as an immune-related gene in AMI, and we
360 expanded it to have diagnostic value. More importantly, this study is the first to reveal the RNA
361 correlation of AQP9 and SOCS3, especially SOCS3, between the number of stenotic coronary
362 arteries and the Killip classification.

363

364 AQP9, a cell membrane protein, transports water down the concentration gradient. ERK1/2 can
365 be reversed in AMI rats by silencing AQP9, attenuating cardiomyocytes' inflammatory response
366 and apoptosis and upregulating cardiac function [43]. The above research indicated the crucial
367 role of AQP9 in the pathogenesis of AMI. In human polymorphonuclear leukocytes, AQP9-
368 related inflammation may result from the NK- κ B [44] and F-actin polymerization [45]. In our
369 work, the ROC curve of AQP9 was > 0.9 . Therefore, AQP9 might be a potential genetic marker
370 for diagnosing AMI with SCAD.

371

372 SOCS3 is increased in AMI mice [29] and regulates the T-cell repertoire with STAT3/SOCS3
373 signalling [46]. More importantly, cardiac-specific silencing of SOCS3 triggers sustained STAT3
374 and decreases myocardial apoptosis [47]. Therefore, SOCS3 is the dominant negative modulator
375 [48] of Th17 via STAT3 [49]. Apoptosis regulates the pathophysiological evaluation of AMI [50].
376 In vitro, SOCS3 can trigger the apoptosis of mammary cells [51], and knocking out SOCS3
377 regulates the expression of apoptosis in 3T3-L1 preadipocytes [52]. The above research
378 emphasized the immune regulation of SOCS3 and the regulation of apoptosis with STAT3. In our

379 work, the ROC curve of SOCS3 was > 0.9 . Therefore, SOCS3 might be an effective genetic
380 marker for diagnosing AMI.

381

382 Additionally, the CIBERSORT algorithm showed that the proportion of neutrophils and
383 monocytes in the AMI group was higher than in the control group. The progression of AMI is
384 correlated with immune disorder. For example, the white blood cell count correlates highly with
385 in-hospital mortality after AMI [53]. Neutrophils are increased in peripheral blood, and
386 researchers have emphasized that neutrophils-lymphocytes [54, 55] and monocytes/macrophages
387 [56] can be easily acquired factors for the prognosis of AMI. Macrophages were dominant in
388 infarcted myocardium, especially over the first week of AMI [57]. However, NK cells have
389 diminished cytotoxic function [58], and the targeted regulation of NK cells may indicate a
390 dominant role in the cure of AMI. At the beginning of AMI, inflammation deteriorates with
391 increased neutrophils and monocytes [59], and inflammation decreases over time with the reduced
392 function of NK cells. Innate immunity is a vital regulatory factor in the inflammatory,
393 proliferative, and maturation phases [3, 60, 61]. AMI leads to a deteriorated inflammatory process.
394 Currently, novel therapeutic interventions targeting the immune system may regulate slant
395 inflammation, which is conducive to resolving pathological conditions. In a previous clinical trial
396 of 182 NSTEMI patients (a subtype of AMI), the patient's intake of IL-1 blockers decreased acute
397 inflammation [62]. Another immune study showed that short-term blockade of S100A9
398 downregulates inflammation [63] in permanent coronary ischaemia mice. However, the above
399 immune interventions are still experimental and not in the clinic. In summary, regulating immune

400 cells along with the progression of AMI and immune intervention in AMI might be a potential
401 target.

402 AQP9 expression was highest in human polymorphonuclear leukocytes [45] compared with the
403 spleen and liver, suggesting a possible correlation between AQP9 and immunity or inflammation.

404 AQP9 regulates water flow on leukocytes [64], which regulates cellular morphology and motility,
405 a change that facilitates the migration of leukocytes to inflammatory sites. Similar to our result,

406 Hawang [65] indicated the correlation between AQP9 and neutrophil granulocytes. Research [29,
407 60, 61] emphasizes the correlation between SOCS3 and neutrophils in inflammation. In our

408 research, both genes had a higher correlation with two immune cells, neutrophils and monocytes.

409 The immune cell correlation indicated that the targeted gene therapy of immune cells may benefit

410 the course of AMI—potential feasibility of using AQP9 and SOCS3 as therapeutic targets or
411 predictors of treatment response.

412 ML algorithms are widely performed for various cardiovascular diseases, such as optimizing
413 variables, classification, and regression. For variable filtration, numerous studies take only

414 single or double ML algorithms (e.g., weighted gene coexpression network analysis [66], LASSO,
415 and SVM). However, only the single or double ML algorithms might unconsciously delete the

416 potential genes. For example, AQP9 will be ignored if we only take DT because the weights of
417 AQP9 were zero in DT (**Table 4**). Taking only a single ML might miss some potential genes. For

418 example, although LASSO can detect candidate genes with big data when highly correlated
419 features exist, the LASSO regression method tends to select one of them and ignore all the other

420 features, leading to the instability of the results [67]. In pigmented skin lesions [68], SVM and

421 NN displayed their talent classification value. In preoperative postsurgical mortality [69], GBM is
422 optimized rather than DT, RF, and SVM. Various ML algorithms may show different weights even
423 in the same variable (**Table 4**). Necessarily, the overall normalized weights of IML were taken to
424 filter genes. Surprisingly, IML explores two potential, unreported diagnostic genes in AMI. In our
425 study, IML has good value in both variable screening and model prediction.

426

427 Inevitably, three limitations exist in this work, although the best efforts were taken to eliminate
428 them. Primarily, small sample size verification might possess some bias. So, multicentre
429 collaborations or leveraging larger external datasets is crucial for further verification. Although
430 testing sets and clinical validation were developed to assess the stability of the diagnostic value,
431 the bias of single-centre validation might exist. More confirmation, clinical trials and animal
432 experiments are indispensable for solid verification. Next, the ML algorithms contained
433 limitations (e.g., the black box phenomenon [70]), especially NN, which has numerous layers [71].
434 The set of operations an ML performs in making a prediction is unknown, even if a human knows
435 precisely what the model is doing at each step of the decision-making process. The operations
436 performed cannot be described in terms of human-understandable semantics. And the
437 Interpretability techniques for ML models always catch the eye of developers, which enhances the
438 transparency and reliability of the ML. Finally, limited clinical features were obtained (e.g., age
439 [72], ethnicity, and race [73]). Clinical features could potentially enhance the predictive accuracy
440 of the diagnostic model and provide a more comprehensive understanding of AMI. For example,
441 various combinations (*e.g.*, sex, smoking or not, and laboratory indicators) of clinical variables

442 [74] are calibrated to analyze the relationship between the target variable and the outcome.

443

444 **5 Conclusion**

445

446 Based on the overall normalized weights of IML, the research successfully merges four

447 microarrays and uncovers hidden diagnostic genes AQP9 and SOCS3 for leukocytes of AMI

448 patients. AQP9 and SOCS3 are closely associated with monocytes and neutrophils, which might

449 contribute to advancing AMI diagnosis and shedding light on novel genetic markers, including

450 AMI pathogenesis, targeted therapies, and potential precision medicine. Although clinical

451 validation copies the result again. Multiple clinical characteristics, multicenter, and large-sample

452 relevant trials are still needed to confirm its clinical value.

453

454 **Supplementary Materials:**

455 Table S1. The 39 DEGs in healthy controls and AMI patients.

456 Table S2. GSEA enrichment of 45 terms.

457 Table S3. GO enrichment of 160 terms.

458 Table S4. DO enrichment of 41 terms.

459 Table S5. Primary weight of DEGs in the six classification ML algorithms.

460 Fig. S1 Difference between the 22 immune cells.

461 Fig. S2. Correlation analysis of SOCS3 in 7 immune cell types.

462 Fig. S3. Correlation analysis of AQP9 in 7 immune cell types.

463 Fig S4. The Correlation analysis of 9 clinical variables.

464

465 **Authors' contributions:** LZ and YL wrote the original draft. LZ, KYW, YL, JSZ, and HLZ
466 performed the research. LZ, YL, XZ, and KYW analyzed the data. SQ, MH, JSZ, and HLZ
467 designed the experiment and revised the manuscript.

468

469 **Funding:** The research was funded by Suzhou Science & Technology Development Plan
470 (SYSD2019222). Zhangjiagang science and technology plan project (ZKS2135), Youth science and
471 technology project of Zhangjiagang Municipal Health Commission (ZJGQNKJ202211).

472

473 **Consent for publication:** This study has not been published before, and this publication has been
474 approved by all authors.

475

476 **Ethics approval and consent to participate**

477 The clinical trial part was approved by the Ethics Review Committee of Jinghai District Hospital.

478

479 **Availability of data and material**

480 The datasets presented in this study can be found online. The names of the repositories and GEO
481 numbers can be found below: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59867>;
482 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60993>;<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62646>;<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4806>

484 [Q.](#)

485

486 **Competing of interests:**

487 The authors declare that they have no conflicts of interest.

488

489 **Abbreviation:**

490 AUC: Area under the Curve; AMI: Acute Myocardial Infarction; IML: Integration Machine

491 Learning; DEGs: Differently Expressed Genes; KEGG-GSEA: Kyoto Encyclopedia of Genes and

492 Genomes-Gene Set Enrichment Analysis; GO: Gene Ontology; DO: Disease Ontology; MF:

493 Molecular Function; BP: Biological Process; CC: Cellular Components; SVM: Support Vector

494 Machine; ML: Machine Learning; LASSO: Least Absolute Shrinkage and Selection Operator; RF:

495 Random Forest; GBM: Gradient Boosting Machine; DT: Decision Trees; NN: Neural Network.

496

497 **Acknowledgments**

498 We thank Suzhou Science & Technology Development Plan.

499

500 **References**

- 501 1. **Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases**
502 **and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-**
503 **2017: a systematic analysis for the Global Burden of Disease Study 2017.** *Lancet* 2018,
504 **392**:1859-1922.
- 505 2. Murray CJ, Barber RM, Foreman KJ, Abbasoglu Ozgoren A, Abd-Allah F, Abera SF,
506 Aboyans V, Abraham JP, Abubakar I, Abu-Raddad LJ, et al: **Global, regional, and national**
507 **disability-adjusted life years (DALYs) for 306 diseases and injuries and healthy life**
508 **expectancy (HALE) for 188 countries, 1990-2013: quantifying the epidemiological**
509 **transition.** *Lancet* 2015, **386**:2145-2191.

- 510 3. Reed GW, Rossi JE, Cannon CP: **Acute myocardial infarction.** *Lancet* 2017, **389**:197-210.
- 511 4. Anderson JL, Morrow DA: **Acute Myocardial Infarction.** *N Engl J Med* 2017, **376**:2053-
- 512 2064.
- 513 5. Levine GN, Bates ER, Bittl JA, Brindis RG, Fihn SD, Fleisher LA, Granger CB, Lange RA,
- 514 Mack MJ, Mauri L, et al: **2016 ACC/AHA Guideline Focused Update on Duration of Dual**
- 515 **Antiplatelet Therapy in Patients With Coronary Artery Disease: A Report of the**
- 516 **American College of Cardiology/American Heart Association Task Force on Clinical**
- 517 **Practice Guidelines: An Update of the 2011 ACCF/AHA/SCAI Guideline for**
- 518 **Percutaneous Coronary Intervention, 2011 ACCF/AHA Guideline for Coronary Artery**
- 519 **Bypass Graft Surgery, 2012 ACC/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the**
- 520 **Diagnosis and Management of Patients With Stable Ischemic Heart Disease, 2013**
- 521 **ACCF/AHA Guideline for the Management of ST-Elevation Myocardial Infarction, 2014**
- 522 **AHA/ACC Guideline for the Management of Patients With Non-ST-Elevation Acute**
- 523 **Coronary Syndromes, and 2014 ACC/AHA Guideline on Perioperative Cardiovascular**
- 524 **Evaluation and Management of Patients Undergoing Noncardiac Surgery.** *Circulation*
- 525 2016, **134**:e123-155.
- 526 6. Huang CC, Liao PC: **Heart Attack Causes Head-Ache - Cardiac Cephalalgia.** *Acta*
- 527 *Cardiol Sin* 2016, **32**:239-242.
- 528 7. Di Stefano R, Di Bello V, Barsotti MC, Grigoratos C, Armani C, Dell'Omodarme M, Carpi A,
- 529 Balbarini A: **Inflammatory markers and cardiac function in acute coronary syndrome:**
- 530 **difference in ST-segment elevation myocardial infarction (STEMI) and in non-STEMI**
- 531 **models.** *Biomed Pharmacother* 2009, **63**:773-780.
- 532 8. Wei EY, Hira RS, Huang HD, Wilson JM, Elayda MA, Sherron SR, Birnbaum Y: **Pitfalls in**
- 533 **diagnosing ST elevation among patients with acute myocardial infarction.** *J*
- 534 *Electrocardiol* 2013, **46**:653-659.
- 535 9. Moy E, Barrett M, Coffey R, Hines AL, Newman-Toker DE: **Missed diagnoses of acute**
- 536 **myocardial infarction in the emergency department: variation by patient and facility**
- 537 **characteristics.** *Diagnosis (Berl)* 2015, **2**:29-40.
- 538 10. Peet C, Ivetic A, Bromage DI, Shah AM: **Cardiac monocytes and macrophages after**
- 539 **myocardial infarction.** *Cardiovasc Res* 2020, **116**:1101-1112.
- 540 11. Heidt T, Courties G, Dutta P, Sager HB, Sebas M, Iwamoto Y, Sun Y, Da Silva N, Panizzi P,
- 541 van der Laan AM, et al: **Differential contribution of monocytes to heart macrophages in**
- 542 **steady-state and after myocardial infarction.** *Circ Res* 2014, **115**:284-295.
- 543 12. Jiang K, Tu Z, Chen K, Xu Y, Chen F, Xu S, Shi T, Qian J, Shen L, Hwa J, et al: **Gasdermin**
- 544 **D inhibition confers antineutrophil-mediated cardioprotection in acute myocardial**
- 545 **infarction.** *J Clin Invest* 2022, **132**.
- 546 13. Liang Y, Chen H, Wang P: **Correlation of Leukocyte and Coronary Lesion Severity of**
- 547 **Acute Myocardial Infarction.** *Angiology* 2018, **69**:591-599.
- 548 14. Kazimierczyk E, Eljaszewicz A, Zembko P, Tarasiuk E, Rusak M, Kulczynska-Przybik A,
- 549 Lukaszewicz-Zajac M, Kaminski K, Mroczko B, Szmitkowski M, et al: **The relationships**
- 550 **among monocyte subsets, miRNAs and inflammatory cytokines in patients with acute**
- 551 **myocardial infarction.** *Pharmacol Rep* 2019, **71**:73-81.
- 552 15. Sreejit G, Abdel-Latif A, Athmanathan B, Annabathula R, Dhyani A, Noothi SK, Quaiife-Ryan

- 553 GA, Al-Sharea A, Pernes G, Dragoljevic D, et al: **Neutrophil-Derived S100A8/A9 Amplify**
554 **Granulopoiesis After Myocardial Infarction.** *Circulation* 2020, **141**:1080-1094.
- 555 16. Long R, Gao L, Li Y, Li G, Qin P, Wei Z, Li D, Qian C, Li J, Yang G: **M2 macrophage-**
556 **derived exosomes carry miR-1271-5p to alleviate cardiac injury in acute myocardial**
557 **infarction through down-regulating SOX6.** *Mol Immunol* 2021, **136**:26-35.
- 558 17. Li R, Jin J, Liu E, Zhang J: **A novel circulating biomarker lnc-MALAT1 for acute**
559 **myocardial infarction: Its relationship with disease risk, features, cytokines, and major**
560 **adverse cardiovascular events.** *J Clin Lab Anal* 2022, **36**:e24771.
- 561 18. **Global, regional, and national incidence, prevalence, and years lived with disability for**
562 **354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic**
563 **analysis for the Global Burden of Disease Study 2017.** *Lancet* 2018, **392**:1789-1858.
- 564 19. Lu Y, Wu Q, Liao J, Zhang S, Lu K, Yang S, Wu Y, Dong Q, Yuan J, Zhao N, Du Y:
565 **Identification of the distinctive role of DPT in dilated cardiomyopathy: a study based on**
566 **bulk and single-cell transcriptomic analysis.** *Ann Transl Med* 2021, **9**:1401.
- 567 20. Di Z, Di M, Fu W, Tang Q, Liu Y, Lei P, Gu X, Liu T, Sun M: **Integrated Analysis Identifies**
568 **a Nine-microRNA Signature Biomarker for Diagnosis and Prognosis in Colorectal**
569 **Cancer.** *Front Genet* 2020, **11**:192.
- 570 21. Wang K, Zhang L, Li L, Wang Y, Zhong X, Hou C, Zhang Y, Sun C, Zhou Q, Wang X:
571 **Identification of Drug-Induced Liver Injury Biomarkers from Multiple Microarrays**
572 **Based on Machine Learning and Bioinformatics Analysis.** *Int J Mol Sci* 2022, **23**.
- 573 22. Zhang L, Lin Y, Wang K, Han L, Zhang X, Gao X, Li Z, Zhang H, Zhou J, Yu H, Fu X:
574 **Multiple-model machine learning identifies potential functional genes in dilated**
575 **cardiomyopathy.** *Front Cardiovasc Med* 2022, **9**:1044443.
- 576 23. Zhang L, Mao R, Lau CT, Chung WC, Chan JCP, Liang F, Zhao C, Zhang X, Bian Z:
577 **Identification of useful genes from multiple microarrays for ulcerative colitis diagnosis**
578 **based on machine learning methods.** *Sci Rep* 2022, **12**:9962.
- 579 24. DeLong JH, Ohashi SN, O'Connor KC, Sansing LH: **Inflammatory Responses After**
580 **Ischemic Stroke.** *Semin Immunopathol* 2022, **44**:625-648.
- 581 25. Denorme F, Portier I, Rustad JL, Cody MJ, de Araujo CV, Hoki C, Alexander MD, Grandhi R,
582 Dyer MR, Neal MD, et al: **Neutrophil extracellular traps regulate ischemic stroke brain**
583 **injury.** *J Clin Invest* 2022, **132**.
- 584 26. Swirski FK, Nahrendorf M: **Leukocyte behavior in atherosclerosis, myocardial infarction,**
585 **and heart failure.** *Science* 2013, **339**:161-166.
- 586 27. Hiremath A, Shiradkar R, Fu P, Mahran A, Rastinehad AR, Tewari A, Tirumani SH, Purysko
587 A, Ponsky L, Madabhushi A: **An integrated nomogram combining deep learning, Prostate**
588 **Imaging-Reporting and Data System (PI-RADS) scoring, and clinical variables for**
589 **identification of clinically significant prostate cancer on biparametric MRI: a**
590 **retrospective multicentre study.** *Lancet Digit Health* 2021, **3**:e445-e454.
- 591 28. Wang Y, Guan Q, Lao I, Wang L, Wu Y, Li D, Ji Q, Wang Y, Zhu Y, Lu H, Xiang J: **Using**
592 **deep convolutional neural networks for multi-classification of thyroid tumor by**
593 **histopathology: a large-scale pilot study.** *Ann Transl Med* 2019, **7**:468.
- 594 29. Zhu X, Yin T, Zhang T, Zhu Q, Lu X, Wang L, Liao S, Yao W, Zhou Y, Zhang H, Li X:
595 **Identification of Immune-Related Genes in Patients with Acute Myocardial Infarction**

- 596 **Using Machine Learning Methods.** *J Inflamm Res* 2022, **15**:3305-3321.
- 597 30. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer:**
598 **is there a unique set?** *Bioinformatics* 2005, **21**:171-178.
- 599 31. Tang G, Qi L, Sun Z, Liu J, Lv Z, Chen L, Huang B, Zhu S, Liu Y, Li Y: **Evaluation and**
600 **analysis of incidence and risk factors of lower extremity venous thrombosis after**
601 **urologic surgeries: A prospective two-center cohort study using LASSO-logistic**
602 **regression.** *Int J Surg* 2021, **89**:105948.
- 603 32. Zhou S: **Sparse SVM for Sufficient Data Reduction.** *IEEE Trans Pattern Anal Mach Intell*
604 2022, **44**:5560-5571.
- 605 33. Rosenblatt WH, Yanez ND: **A Decision Tree Approach to Airway Management Pathways**
606 **in the 2022 Difficult Airway Algorithm of the American Society of Anesthesiologists.**
607 *Anesth Analg* 2022, **134**:910-915.
- 608 34. Utkin LV, Konstantinov AV: **Attention-based random forest and contamination model.**
609 *Neural Netw* 2022, **154**:346-359.
- 610 35. Kriegeskorte N, Golan T: **Neural network models and deep learning.** *Curr Biol* 2019,
611 **29**:R231-r236.
- 612 36. Dash TK, Chakraborty C, Mahapatra S, Panda G: **Gradient Boosting Machine and Efficient**
613 **Combination of Features for Speech-Based Detection of COVID-19.** *IEEE J Biomed*
614 *Health Inform* 2022, **26**:5364-5371.
- 615 37. Jaatinen T, Laine J: **Isolation of mononuclear cells from human cord blood by Ficoll-**
616 **Paque density gradient.** *Curr Protoc Stem Cell Biol* 2007, **Chapter 2**:Unit 2A.1.
- 617 38. Tan YS, Lei YL: **Isolation of Tumor-Infiltrating Lymphocytes by Ficoll-Paque Density**
618 **Gradient Centrifugation.** *Methods Mol Biol* 2019, **1960**:93-99.
- 619 39. Sugiyama Y, Yamazaki K, Kusaka-Kikushima A, Nakahigashi K, Hagiwara H, Miyachi Y:
620 **Analysis of aquaporin 9 expression in human epidermis and cultured keratinocytes.**
621 *FEBS Open Bio* 2014, **4**:611-616.
- 622 40. Chen J, Yu L, Zhang S, Chen X: **Network Analysis-Based Approach for Exploring the**
623 **Potential Diagnostic Biomarkers of Acute Myocardial Infarction.** *Front Physiol* 2016,
624 **7**:615.
- 625 41. Shao G: **Integrated RNA gene expression analysis identified potential immune-related**
626 **biomarkers and RNA regulatory pathways of acute myocardial infarction.** *PLoS One*
627 2022, **17**:e0264362.
- 628 42. Yang Y, Liu P, Teng R, Liu F, Zhang C, Lu X, Ding Y: **Integrative bioinformatics analysis**
629 **of potential therapeutic targets and immune infiltration characteristics in dilated**
630 **cardiomyopathy.** *Ann Transl Med* 2022, **10**:348.
- 631 43. Huang X, Yu X, Li H, Han L, Yang X: **Regulation mechanism of aquaporin 9 gene on**
632 **inflammatory response and cardiac function in rats with myocardial infarction through**
633 **extracellular signal-regulated kinase1/2 pathway.** *Heart Vessels* 2019, **34**:2041-2051.
- 634 44. Takeuchi K, Hayashi S, Matumoto T, Hashimoto S, Takayama K, Chinzei N, Kihara S,
635 Haneda M, Kirizuki S, Kuroda Y, et al: **Downregulation of aquaporin 9 decreases catabolic**
636 **factor expression through nuclear factor- κ B signaling in chondrocytes.** *Int J Mol Med*
637 2018, **42**:1548-1558.
- 638 45. Matsushima A, Ogura H, Koh T, Shimazu T, Sugimoto H: **Enhanced expression of**

- 639 **aquaporin 9 in activated polymorphonuclear leukocytes in patients with systemic**
640 **inflammatory response syndrome.** *Shock* 2014, **42**:322-326.
- 641 46. Baker BJ, Akhtar LN, Benveniste EN: **SOCS1 and SOCS3 in the control of CNS immunity.**
642 *Trends Immunol* 2009, **30**:392-400.
- 643 47. Negoro S, Kunisada K, Fujio Y, Funamoto M, Darville MI, Eizirik DL, Osugi T, Izumi M,
644 Oshima Y, Nakaoka Y, et al: **Activation of signal transducer and activator of transcription**
645 **3 protects cardiomyocytes from hypoxia/reoxygenation-induced oxidative stress through**
646 **the upregulation of manganese superoxide dismutase.** *Circulation* 2001, **104**:979-981.
- 647 48. Yoshimura A, Naka T, Kubo M: **SOCS proteins, cytokine signalling and immune**
648 **regulation.** *Nat Rev Immunol* 2007, **7**:454-465.
- 649 49. Chen Z, Laurence A, Kanno Y, Pacher-Zavisin M, Zhu BM, Tato C, Yoshimura A,
650 Hennighausen L, O'Shea JJ: **Selective regulatory function of Socs3 in the formation of IL-**
651 **17-secreting T cells.** *Proc Natl Acad Sci U S A* 2006, **103**:8137-8142.
- 652 50. Scarabelli TM, Stephanou A, Pasini E, Comini L, Raddino R, Knight RA, Latchman DS:
653 **Different signaling pathways induce apoptosis in endothelial cells and cardiac myocytes**
654 **during ischemia/reperfusion injury.** *Circ Res* 2002, **90**:745-748.
- 655 51. Le Provost F, Miyoshi K, Vilotte JL, Bierie B, Robinson GW, Hennighausen L: **SOCS3**
656 **promotes apoptosis of mammary differentiated cells.** *Biochem Biophys Res Commun* 2005,
657 **338**:1696-1701.
- 658 52. Chhabra JK, Chattopadhyay B, Paul BN: **SOCS3 dictates the transition of divergent time-**
659 **phased events in granulocyte TNF- α signaling.** *Cell Mol Immunol* 2014, **11**:105-106.
- 660 53. Dutta P, Nahrendorf M: **Monocytes in myocardial infarction.** *Arterioscler Thromb Vasc Biol*
661 2015, **35**:1066-1070.
- 662 54. Lin G, Dai C, Xu K, Wu M: **Predictive value of neutrophil to lymphocyte ratio and red**
663 **cell distribution width on death for ST segment elevation myocardial infarction.** *Sci Rep*
664 2021, **11**:11506.
- 665 55. Sasmita BR, Zhu Y, Gan H, Hu X, Xue Y, Xiang Z, Huang B, Luo S: **Prognostic value of**
666 **neutrophil-lymphocyte ratio in cardiogenic shock complicating acute myocardial**
667 **infarction: A cohort study.** *Int J Clin Pract* 2021, **75**:e14655.
- 668 56. Kervinen H, Mänttari M, Kaartinen M, Mäkynen H, Palosuo T, Pulkki K, Kovanen PT:
669 **Prognostic usefulness of plasma monocyte/macrophage and T-lymphocyte activation**
670 **markers in patients with acute coronary syndromes.** *Am J Cardiol* 2004, **94**:993-996.
- 671 57. Yan X, Anzai A, Katsumata Y, Matsubashi T, Ito K, Endo J, Yamamoto T, Takeshima A,
672 Shinmura K, Shen W, et al: **Temporal dynamics of cardiac immune cell accumulation**
673 **following acute myocardial infarction.** *J Mol Cell Cardiol* 2013, **62**:24-35.
- 674 58. Ortega-Rodríguez AC, Marín-Jáuregui LS, Martínez-Shio E, Hernández Castro B, González-
675 Amaro R, Escobedo-Urbe CD, Monsiváis-Urenda AE: **Altered NK cell receptor repertoire**
676 **and function of natural killer cells in patients with acute myocardial infarction: A three-**
677 **month follow-up study.** *Immunobiology* 2020, **225**:151909.
- 678 59. Leuschner F, Rauch PJ, Ueno T, Gorbатов R, Marinelli B, Lee WW, Dutta P, Wei Y, Robbins
679 C, Iwamoto Y, et al: **Rapid monocyte kinetics in acute myocardial infarction are**
680 **sustained by extramedullary monocytopenesis.** *J Exp Med* 2012, **209**:123-137.
- 681 60. Nahrendorf M: **Myeloid cell contributions to cardiovascular health and disease.** *Nat Med*

- 682 2018, **24**:711-720.
- 683 61. Swirski FK, Nahrendorf M: **Cardioimmunology: the immune system in cardiac**
684 **homeostasis and disease.** *Nat Rev Immunol* 2018, **18**:733-744.
- 685 62. Yellon DM, Hausenloy DJ: **Myocardial reperfusion injury.** *N Engl J Med* 2007, **357**:1121-
686 1135.
- 687 63. Ridker PM, Everett BM, Thuren T, MacFadyen JG, Chang WH, Ballantyne C, Fonseca F,
688 Nicolau J, Koenig W, Anker SD, et al: **Antiinflammatory Therapy with Canakinumab for**
689 **Atherosclerotic Disease.** *N Engl J Med* 2017, **377**:1119-1131.
- 690 64. Moniaga CS, Watanabe S, Honda T, Nielsen S, Hara-Chikuma M: **Aquaporin-9-expressing**
691 **neutrophils are required for the establishment of contact hypersensitivity.** *Sci Rep* 2015,
692 **5**:15319.
- 693 65. Wang H, Dou S, Wang C, Gao W, Cheng B, Yan F: **Identification and Experimental**
694 **Validation of Parkinson's Disease with Major Depressive Disorder Common Genes.** *Mol*
695 *Neurobiol* 2023, **60**:6092-6108.
- 696 66. Radulescu E, Jaffe AE, Straub RE, Chen Q, Shin JH, Hyde TM, Kleinman JE, Weinberger
697 DR: **Identification and prioritization of gene sets associated with schizophrenia risk by**
698 **co-expression network analysis in human brain.** *Mol Psychiatry* 2020, **25**:791-804.
- 699 67. Choi BY, Bair E, Lee JW: **Nearest shrunken centroids via alternative genewise shrinkages.**
700 *PLoS One* 2017, **12**:e0171068.
- 701 68. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M: **A comparison of**
702 **machine learning methods for the diagnosis of pigmented skin lesions.** *J Biomed Inform*
703 2001, **34**:28-36.
- 704 69. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR: **Utilizing Machine Learning Methods**
705 **for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit**
706 **Admission.** *Ann Surg* 2020, **272**:1133-1139.
- 707 70. Regazzoni F, Chapelle D, Moireau P: **Combining data assimilation and machine learning**
708 **to build data-driven models for unknown long time dynamics-Applications in**
709 **cardiovascular modeling.** *Int J Numer Method Biomed Eng* 2021, **37**:e3471.
- 710 71. Peng JC, Ran ZH, Shen J: **Seasonal variation in onset and relapse of IBD and a model to**
711 **predict the frequency of onset, relapse, and severity of IBD based on artificial neural**
712 **network.** *Int J Colorectal Dis* 2015, **30**:1267-1273.
- 713 72. Kalkan IH, Dağlı U, Oztaş E, Tunç B, Ulker A: **Comparison of demographic and clinical**
714 **characteristics of patients with early vs. adult vs. late onset ulcerative colitis.** *Eur J Intern*
715 *Med* 2013, **24**:273-277.
- 716 73. Jiang L, Xia B, Li J, Ye M, Deng C, Ding Y, Luo H, Ren H, Hou X, Liu H, et al: **Risk factors**
717 **for ulcerative colitis in a Chinese population: an age-matched and sex-matched case-**
718 **control study.** *J Clin Gastroenterol* 2007, **41**:280-284.
- 719 74. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, Zhu W, Sama I, Tadel M,
720 Campagnari C, et al: **Improving risk prediction in heart failure using machine learning.**
721 *Eur J Heart Fail* 2020, **22**:139-147.

722

723

724

Table 1 List of primers for real-time PCR analysis in GAPDH, AQP9, and SOCS3.

Gene	Primer sequences	
GAPDH	F: TGTGGGCATCAATGGATTTGG	R: ACACCATGTATTCCGGGTCAAT
AQP9	F: GCCATCGGCCTCCTGATTAT	R: GCCCACTACAGGAATCCACC
SOCS3	F: TCCAAACAGGGGACACTTCG	R: GGGGGTGTGACCATTTCCTT

725

Table 2 Fundamental information in the 4 datasets.

ID	Public time	Institution	Platform	Country	C	A	Microarray/RNA-seq method
					o	M	
					n	I	
GSE5 9867	21- May- 15	Institute of Biochemistry and Biophysics	GPL 6244	Poland	46	11 1	Affymetrix GCS 3000 GeneArray Scanner
GSE6 0993	23- May- 15	Ajou University of Korea	GPL 6884	South Korea	7	7	HumanHT-12 v3 Expression BeadChip
GSE6 2646	23- Oct-14	Institute of Biochemistry and Biophysics	GPL 6244	Poland	0	28	Affymetrix GCS 3000 GeneArray Scanner
GSE4 8060	28- Feb- 14	Mayo Clinic	GPL 570	USA	21	0	GeneChip Scanner 3000 7G

726

727

Table 3 Accuracy of six MLs based on 39 DEGs in the training and test sets.

MLs	Training sets (%)	Testing sets (%)
LASSO	70.7	70.7
SVM	88.46	91.84
RF	98.09	100
DT	94.27	75.52
GBM	93.3	85.71
NN	83.74	71.43

728

Table 4 Overall weights of six classification models were constructed to optimize the candidate diagnostic genes.

ID	SVM	RF	NN	GBM	DT	LASSO	Overall weights
ASGR2	1	1	0.97	0.21	1	0.61	4.79
SOCS3	0.98	0.34	0.52	0.24	0.59	0.61	3.28
AQP9	0.61	0.1	0.68	1	0	0.52	2.91
PPARG	0.76	0.15	1	0.25	0	0.5	2.66
RNASE1	0.74	0.22	0.24	0.1	0.41	0.74	2.45
DYSF	0.2	0.68	0.01	0.57	0	0.72	2.18
S100A9	0.53	0.09	0.01	0.74	0	0.63	2
FCGR1A	0.17	0.51	0	0.57	0	0.68	1.93
VSIG4	0.44	0.3	0.1	0.19	0.01	0.86	1.9
STAB1	0.47	0.58	0.05	0.14	0	0.61	1.85
MYBL1	0.68	0.15	0.2	0	0.26	0.52	1.81
KLRD1	0.26	0.65	0.01	0.05	0	0.73	1.7
ECRP	0.44	0.24	0.11	0	0.34	0.54	1.67
TCN2	0.46	0.27	0.07	0	0	0.78	1.58
FAM20A	0.31	0.08	0.15	0	0	1	1.54
MERTK	0.19	0.21	0.01	0.1	0.14	0.71	1.36
HP	0.09	0.78	0	0	0	0.45	1.32
RNASE2	0.16	0.42	0.01	0	0	0.7	1.29
DTHD1	0.13	0.45	0.05	0	0	0.66	1.29
CLC	0.11	0.72	0.02	0	0	0.36	1.21
SNORD20	0.14	0.24	0.01	0.13	0.1	0.5	1.12
CD163	0.15	0.29	0	0.11	0	0.57	1.12
NRG1	0.2	0.25	0.02	0	0	0.63	1.1
SNORD45B	0.12	0.64	0.01	0	0	0.33	1.1
CYP1B1	0.14	0.25	0	0	0	0.66	1.05
KLRC2	0.07	0.51	0	0	0	0.46	1.04

TMEM176A	0.08	0.67	0	0	0	0.24	0.99
SLED1	0.09	0.24	0.02	0.05	0	0.49	0.89
FCGR1A.2	0.23	0	0	0.62	0	0	0.85
SERPINB2	0.08	0.21	0	0	0	0.54	0.83
FCGR1A.1	0.18	0	0	0.62	0	0	0.8
KLRC4	0.13	0.21	0	0	0	0.43	0.77
KLRA1P	0.1	0.07	0	0.08	0	0.51	0.76
MIR21	0.08	0.09	0.01	0	0	0.5	0.68
CES1	0.12	0.05	0.03	0	0	0.47	0.67
KLRC4-KLRK1	0.07	0	0	0.08	0	0.43	0.58
KLRC3	0.07	0.1	0	0	0	0.39	0.56
NRG1.1	0.13	0	0	0	0	0	0.13
FMN1.1	0.07	0	0.01	0	0	0	0.08

729

730

731

732

733

734

735

736

737

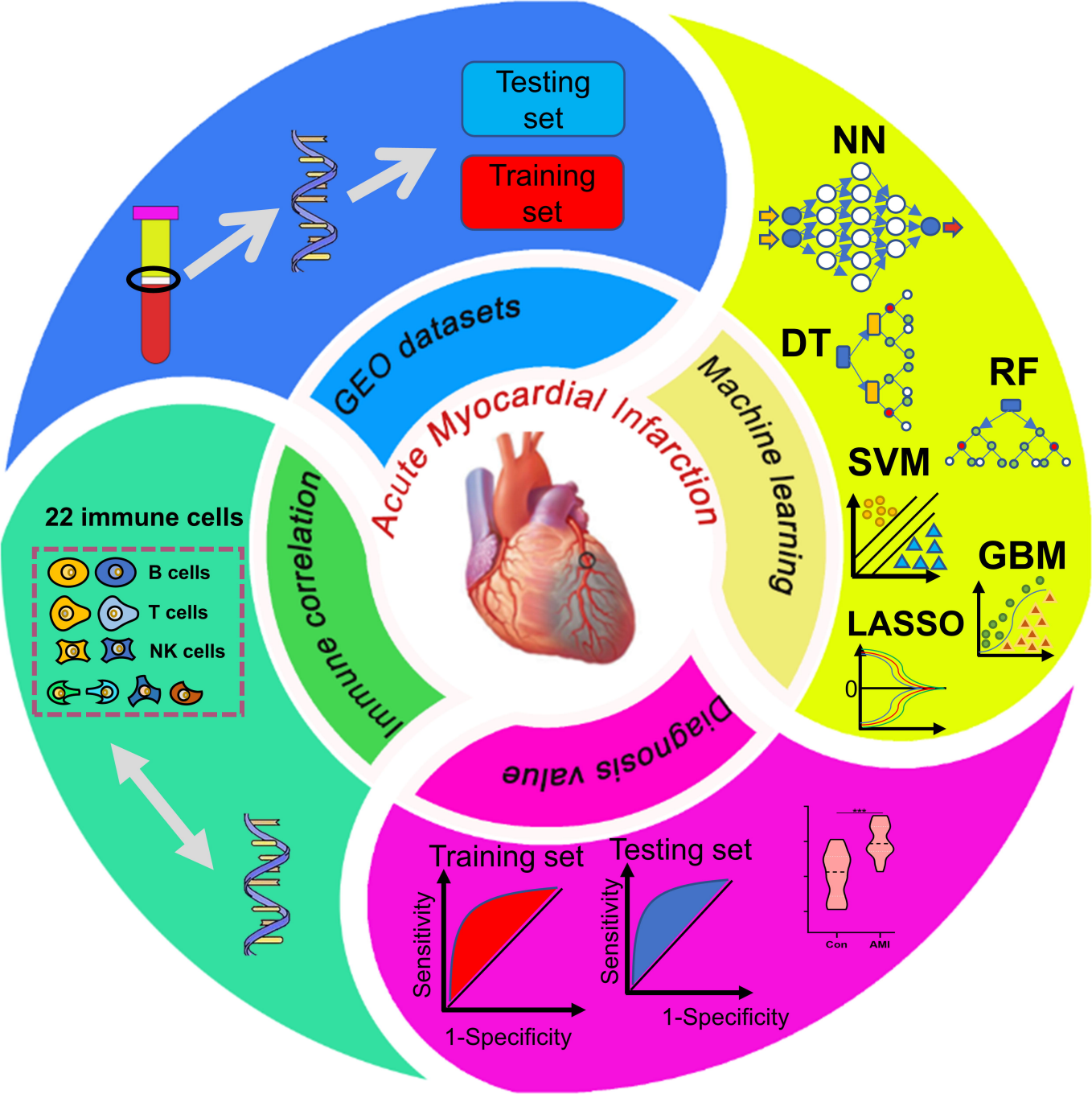
738

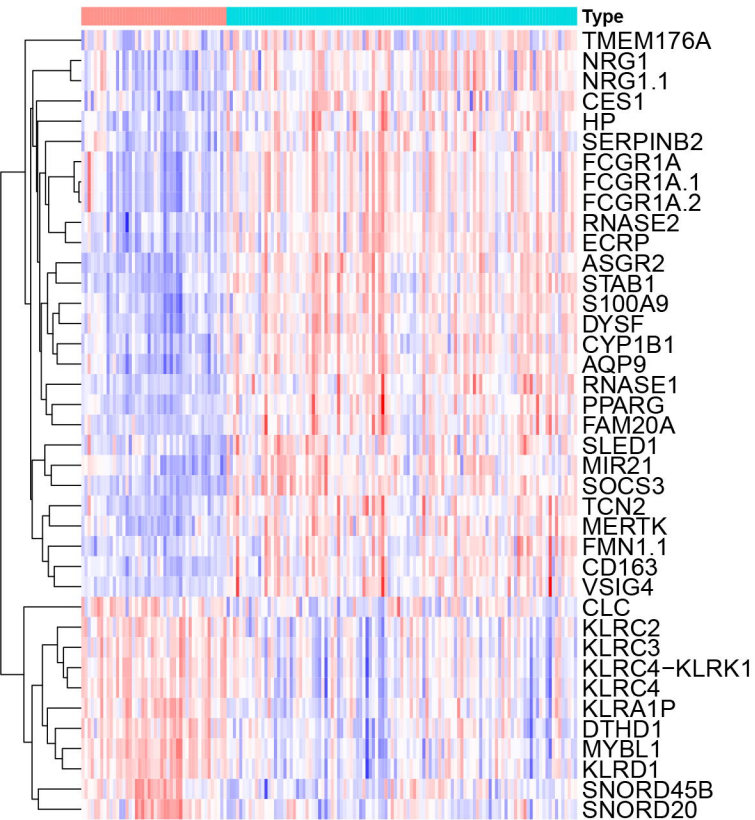
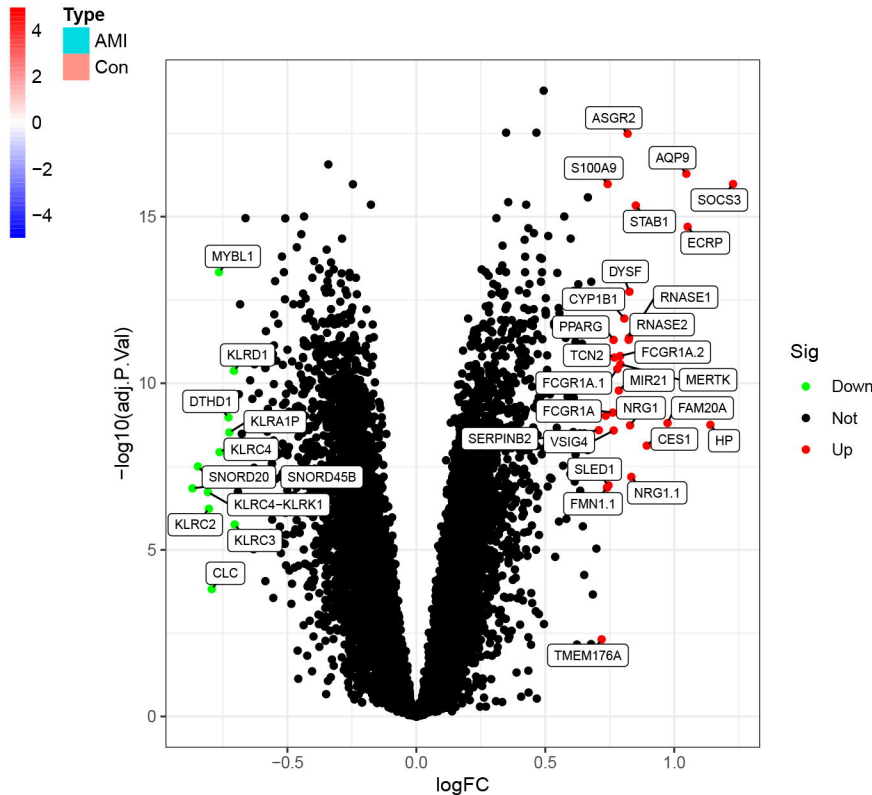
Table 5 The general characteristics of the 40 patients.

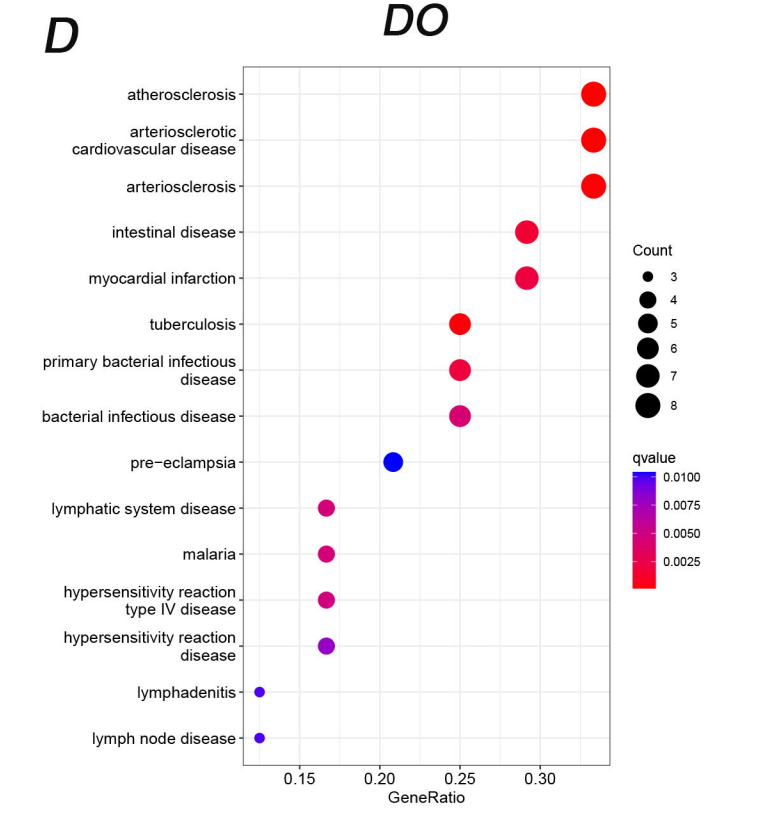
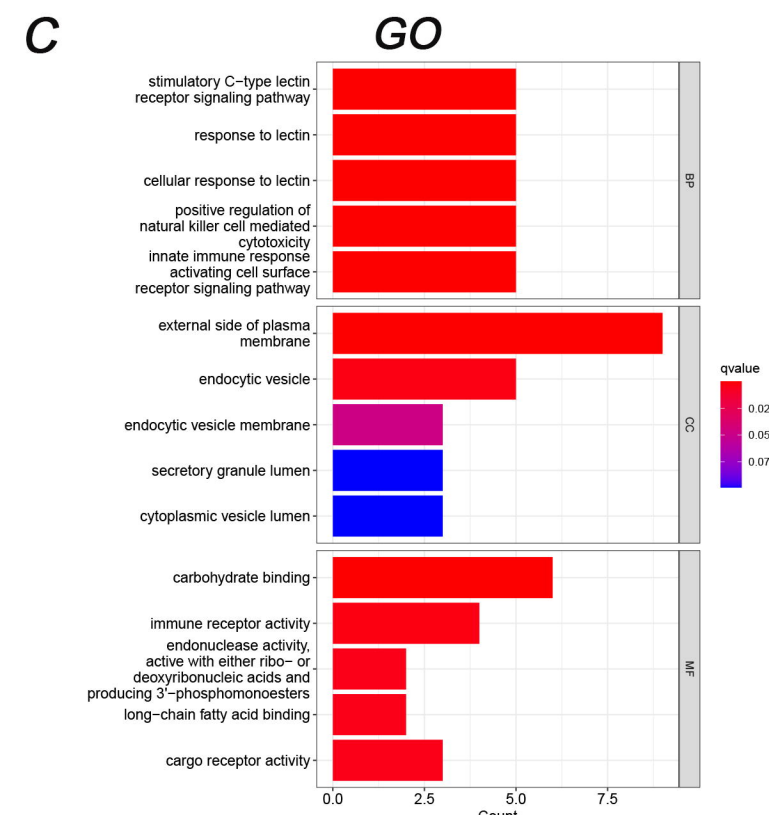
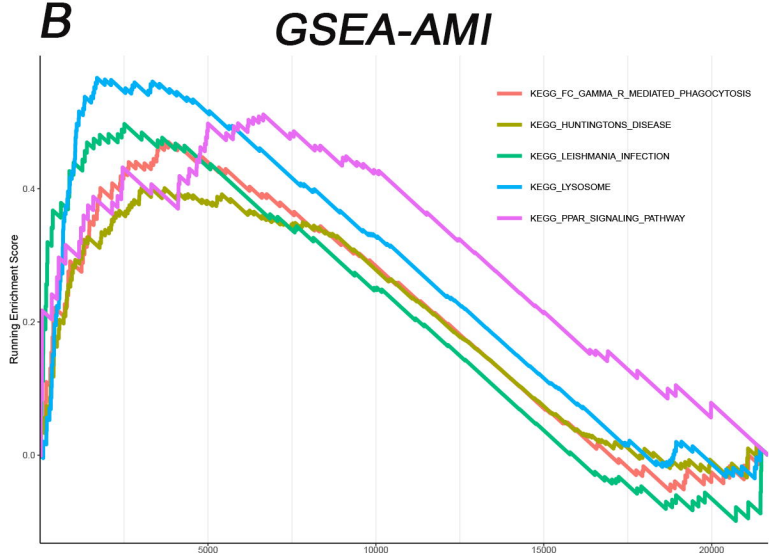
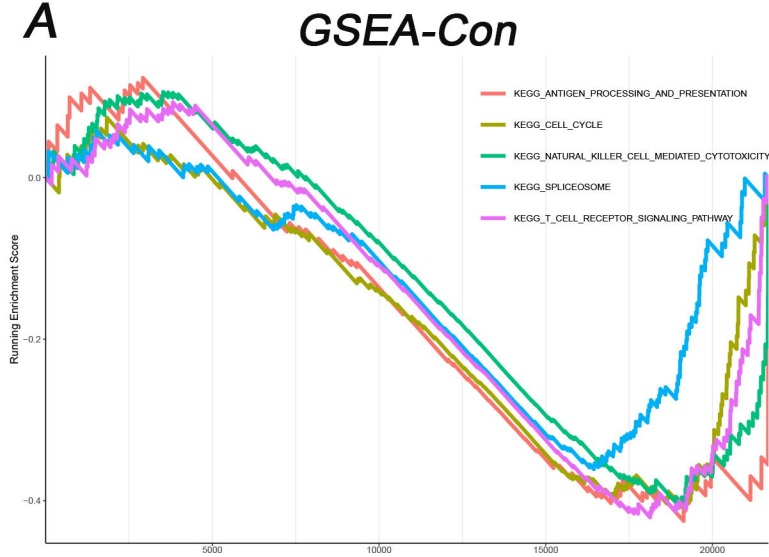
Characteristic	SCAD (<i>n</i> = 20)	AMI (<i>n</i> = 20)	<i>P</i> -value
Hypertension, %	16.00(80)	16.00(80)	>0.05

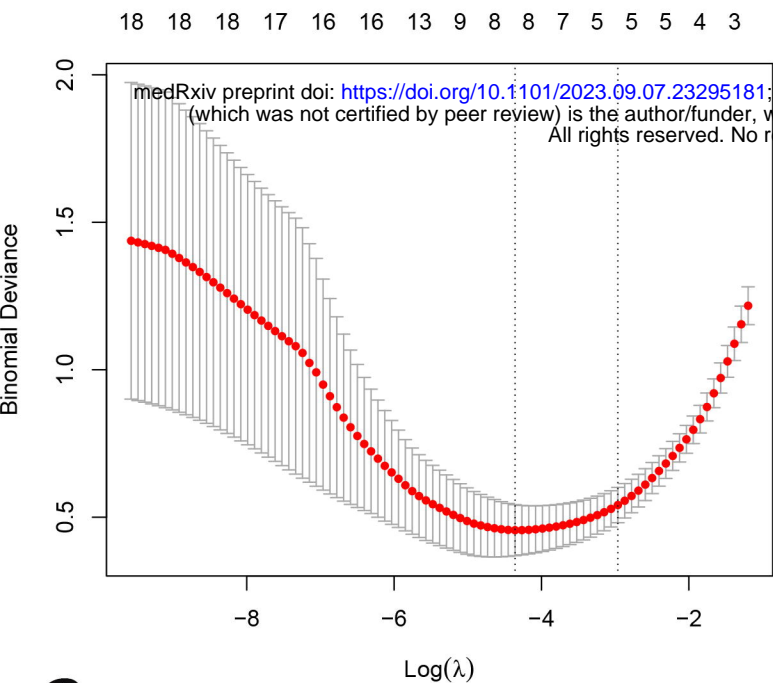
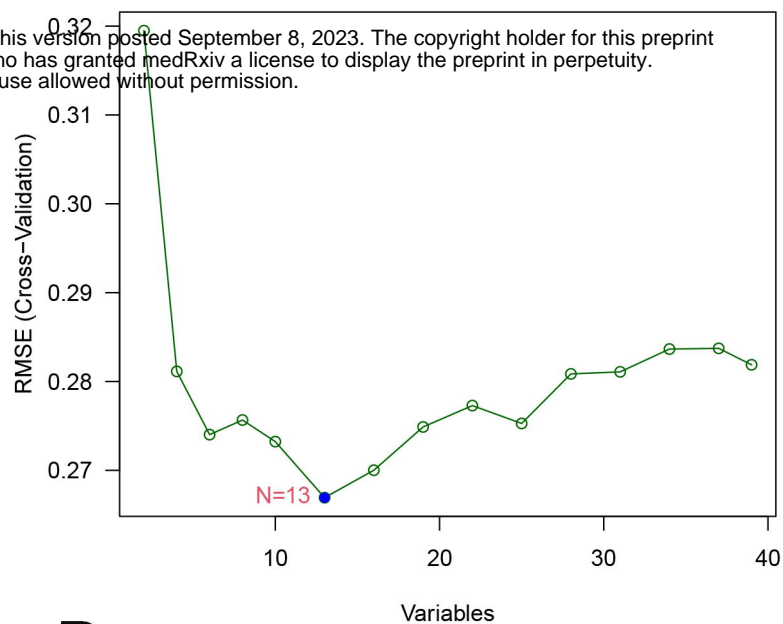
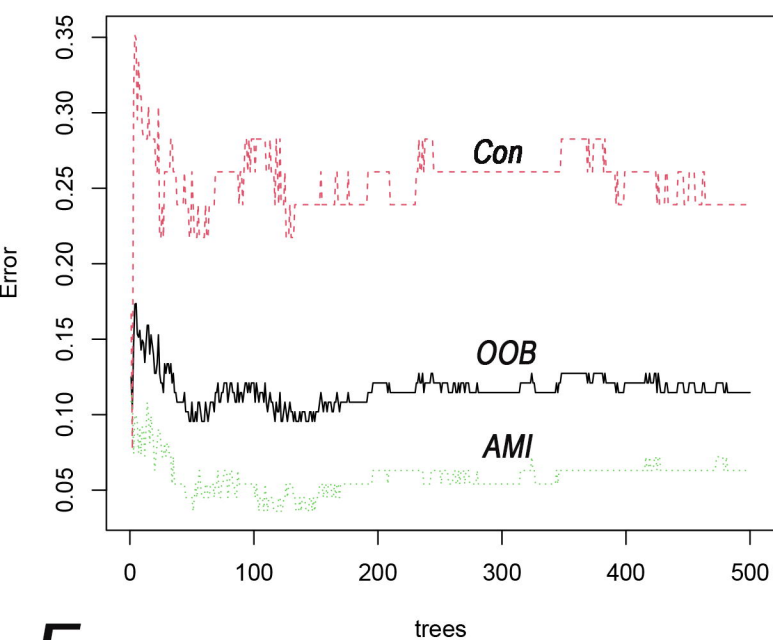
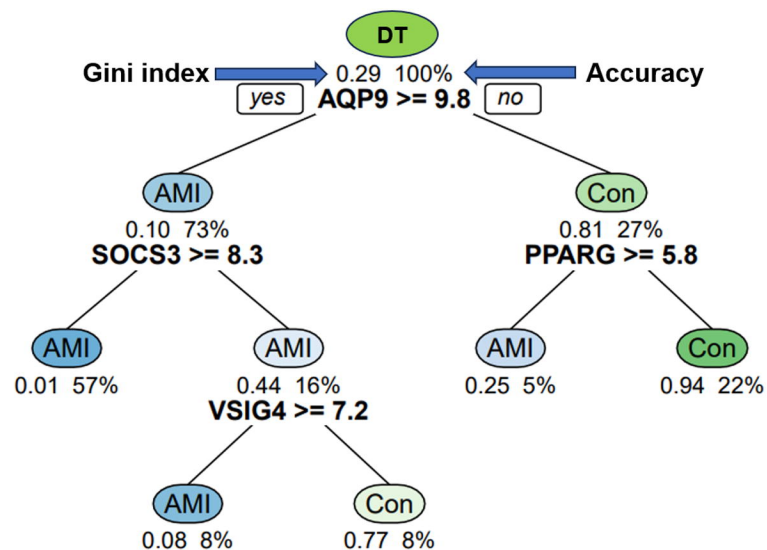
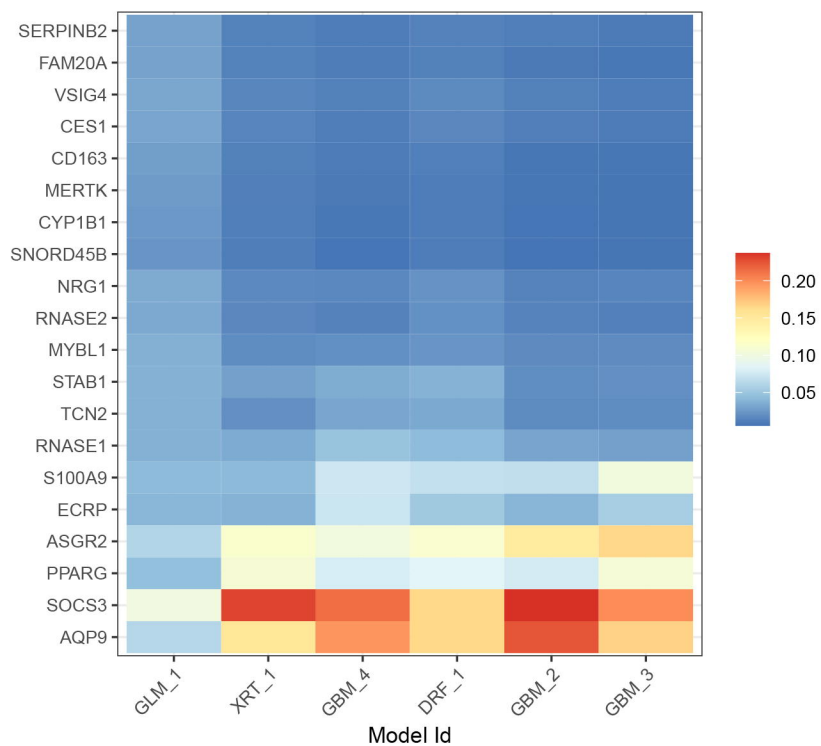
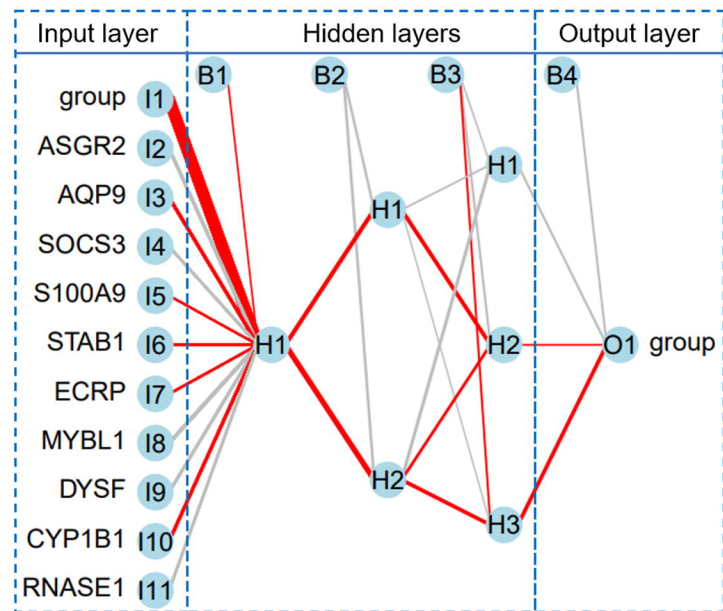
Diabetes mellitus, %	6.00(30)	7.00(35)	>0.05
Stroke, %	4.00(20)	4.00(20)	>0.05
Hyperlipemia, %	4.00(20)	6.00(30)	>0.05
Age, year	66 (63, 72)	70 (58, 77)	>0.05
Sex (Male), %	10.00(50.00)	12.00(60.00)	>0.05
RBC, million cells/ μ L	4.37 (3.80, 4.49)	4.03 (3.55, 4.35)	>0.05
WBC, 1000 cells/ μ L	6.02 (4.85, 6.84)	8.33 (6.84, 11.24)	<0.001
NeP, %	69 (58, 74)	80 (75, 86)	<0.001
MonP, %	8.00 (6.15, 9.23)	5.55 (3.98, 7.68)	<0.05
Mon, 1000 cells/ μ L	0.44 (0.35, 0.48)	0.42 (0.19, 0.73)	>0.05
Lym, 1000 cells/ μ L	1.37 (1.06, 1.77)	0.96 (0.62, 1.42)	<0.05
RDW, %	13.00 (12.55, 13.30)	13.55 (12.88, 14.98)	>0.05
PDW, %	11.85 (10.48, 13.90)	13.25 (11.68, 16.15)	>0.05
Pla, 1000 cells/ μ L	214 (163, 245)	219 (173, 244)	>0.05
MCHC, g/L	334 (329, 342)	329 (319, 338)	>0.05
Hg, g/L	131 (113, 140)	119 (109, 134)	>0.05
GAT, U/L	16 (14, 21)	28 (17, 51)	<0.05
D-dimer, mg/L	0.46 (0.27, 0.69)	1.01 (0.62, 2.70)	<0.01
CRP, mg/L	1 (1, 2)	12 (7, 26)	<0.001
SOCS3	1.57 (1.22, 1.76)	1.97 (1.86, 2.20)	<0.001
AQP9	0.90 (0.85, 1.03)	1.44 (1.16, 1.66)	<0.001
LDH, U/L	152 (141, 194)	260 (228, 303)	<0.001
cTnT, μ g/mL	12 (9, 18)	140 (92, 264)	<0.001
CK-MB, U/L	2 (1, 4)	19 (9, 33)	<0.001
LDL, mmol/L	1.83 (1.57, 2.68)	2.38 (1.74, 3.62)	>0.05
HDL, mmol/L	1.04 (0.96, 1.16)	1.11 (0.96, 1.34)	>0.05
TC, mmol/L	3.59 (2.87, 4.52)	4.15 (3.27, 5.80)	>0.05
TG, mmol/L	1.13 (0.71, 1.58)	0.98 (0.89, 1.22)	>0.05
Glucose, mg/L	5.36 (4.73, 5.81)	6.05 (5.12, 9.15)	>0.05
Cys, μ mol/L	11.7 (10.1, 16.4)	14.2 (8.7, 22.1)	>0.05
Albumin, g/L	41.8 (38.9, 43.3)	38.6 (34.9, 40.3)	<0.01
Total protein, g/L	65 (63, 67)	65 (60, 67)	>0.05
GGT, U/L	13 (11, 22)	17 (12, 33)	>0.05
IBIL, μ mol/L	3.65 (2.50, 6.03)	4.95 (3.50, 7.30)	>0.05
DBIL, μ mol/L	4.00 (2.58, 4.78)	4.30 (2.85, 7.75)	>0.05
IBIL, μ mol/L	7.8 (5.2, 10.7)	10.0 (6.3, 17.4)	>0.05
Globulin, g/L	24.7 (21.2, 25.8)	25.6 (23.5, 26.9)	>0.05
ALP, U/L	72 (61, 87)	87 (72, 108)	>0.05

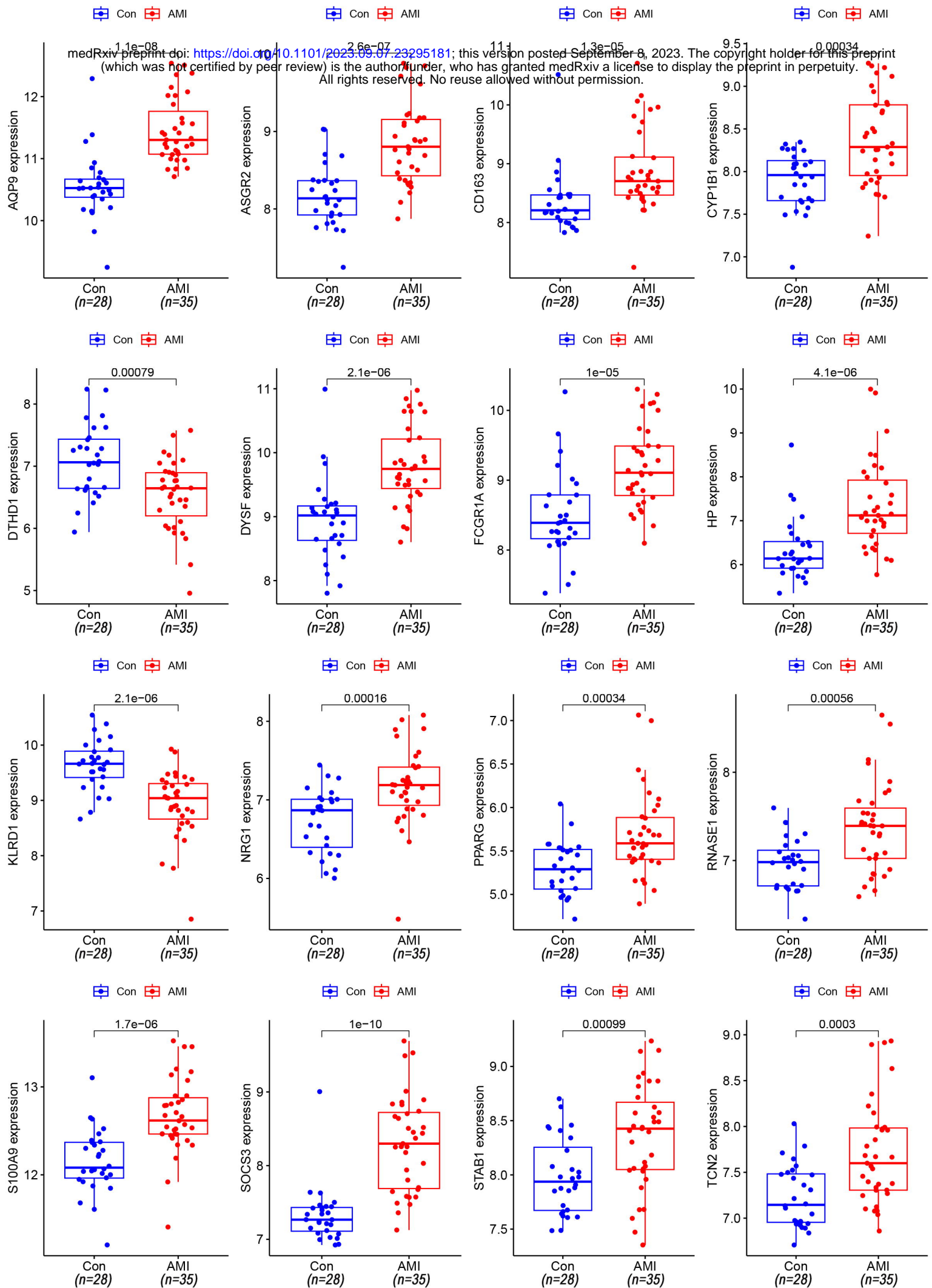
RBC; red blood cell count, WBC; white blood cell count, NeP; neutrophils percentage, MonP; monocyte percentage, Mon; monocyte count, Lym; lymphocyte count, RDW; red blood cell distribution width, PDW; platelet distribution width, Pla; platelet count, MCHC; mean corpuscular haemoglobin concentration, Hg; haemoglobin, GAT; glutamic transaminase, CRP; c-reactive protein, LDH; lactate dehydrogenase, cTnT; cardiac troponin t, CK-MB; creatine kinase isoenzymes, LDL; low-density lipoprotein, HDL; high-density lipoprotein, TC; total cholesterol, TG; total triglycerides, Cys; homocysteine, GGT; gammaglutaminase, IBIL; indirect bilirubin, DBIL; direct bilirubin, IBIL; total bile acid, ALP; alkaline phosphatase.



A**B**

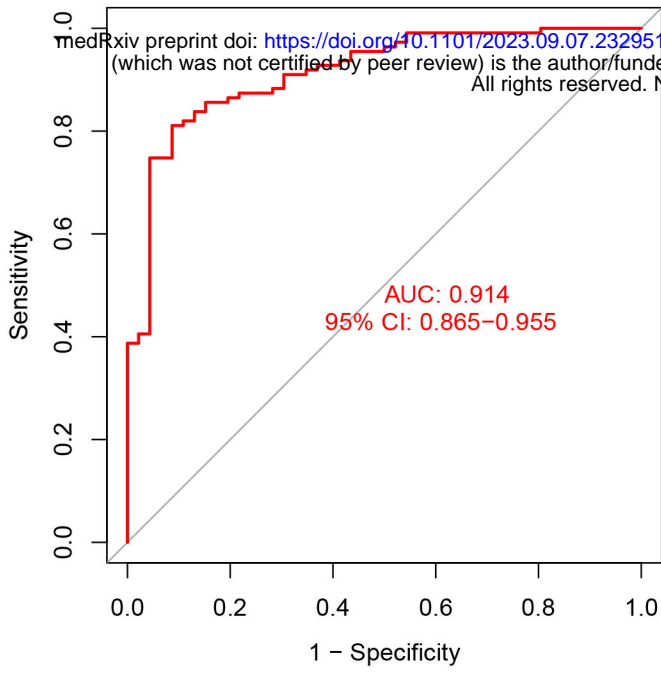


A**B****C****D****E****F**

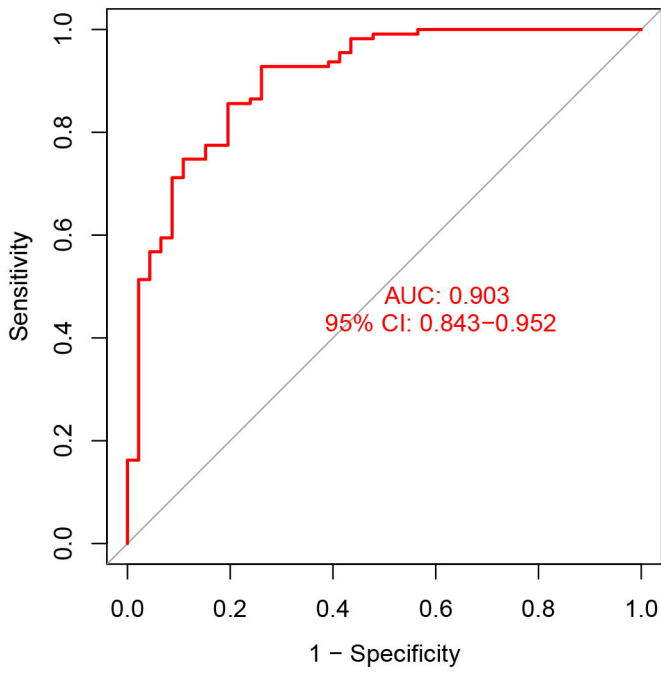


Training set (n=157)

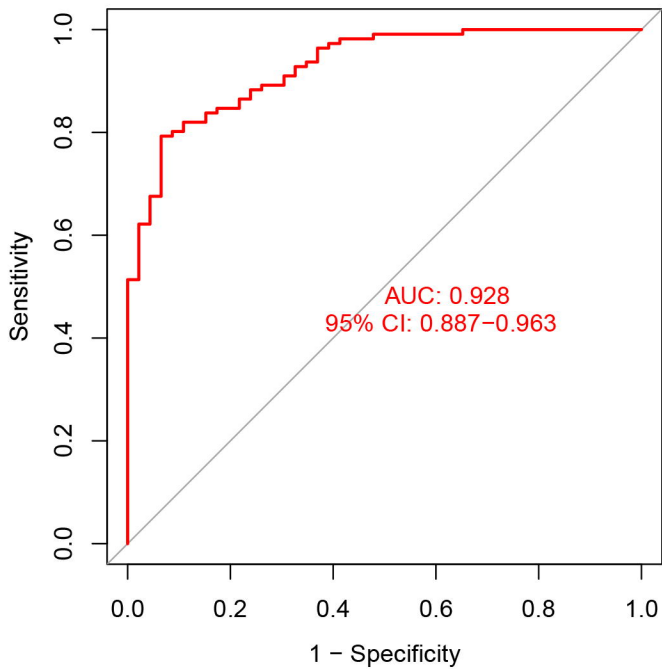
SOCS3



AQP9

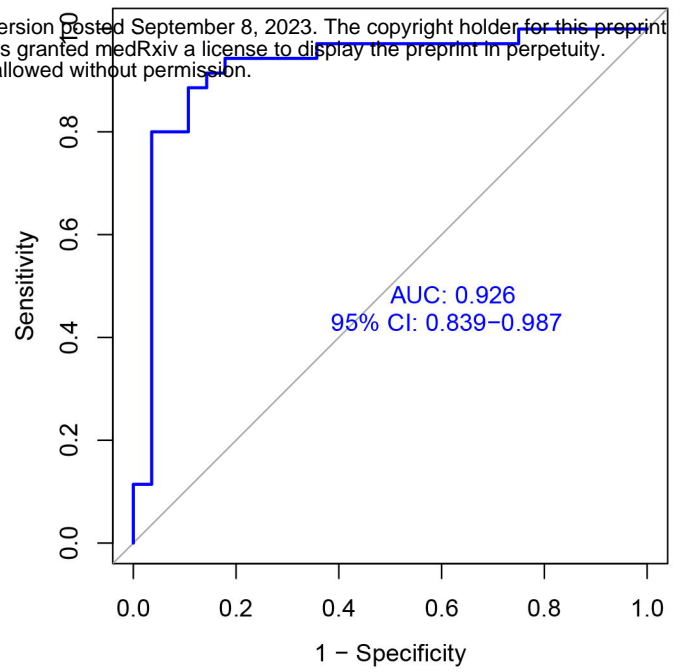


ASGR2

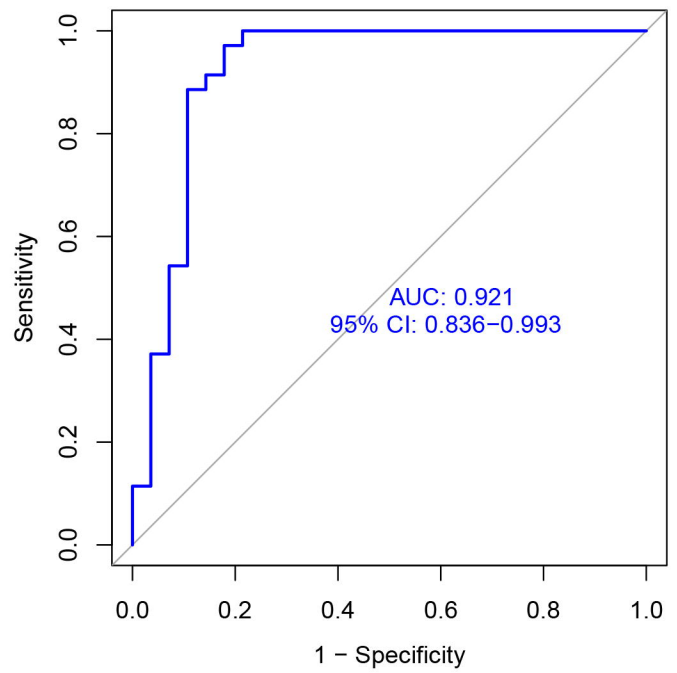


Testing set (n=63)

SOCS3



AQP9



ASGR2

