Molecular Therapy
Nucleic Acids
Original Article

A polygenic stacking classifier revealed the complicated platelet transcriptomic landscape of adult immune thrombocytopenia

Chengfeng Xu,[1,4] Ruochi Zhang,[3,4] Meiyu Duan,[3] Yongming Zhou,[1] Jizhang Bao,[1] Hao Lu,[1] Jie Wang,[1] Minghui Hu,[1] Zhaoyang Hu,[2] Fengfeng Zhou,[3] and Wenwei Zhu[1]

[1]Department of Hematology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, 110 Ganhe Road, Hongkou District, Shanghai 200437, China; [2]Fun-Med Pharmaceutical Technology (Shanghai) Co., Ltd., RM. A310, 115 Xinjunhuan Road, Minhang District, Shanghai 201100, China; [3]College of Computer Science and Technology, Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China

Immune thrombocytopenia (ITP) is an autoimmune disease with the typical symptom of a low platelet count in blood. ITP demonstrated age and sex biases in both occurrences and prognosis, and adult ITP was mainly induced by the living environments. The current diagnosis guideline lacks the integration of molecular heterogenicity. This study recruited the largest cohort of platelet transcriptome samples. A comprehensive procedure of feature selection, feature engineering, and stacking classification was carried out to detect the ITP biomarkers using RNA sequencing (RNA-seq) transcriptomes. The 40 detected biomarkers were loaded to train the final ITP detection model, with an overall accuracy 0.974. The biomarkers suggested that ITP onset may be associated with various transcribed components, including protein-coding genes, long intergenic noncoding RNA (lincRNA) genes, and pseudogenes with apparent transcriptions. The delivered ITP detection model may also be utilized as a complementary ITP diagnosis tool. The code and the example dataset is freely available on http://www.healthinformaticslab.org/supp/resources.php

INTRODUCTION

Immune thrombocytopenia (ITP), previously known as immune thrombocytopenic purpura, is an acquired immune-mediated disease characterized by a blood platelet count less than 100×10^9 per liter.[1] ITP may develop in both children and adults, and female young adults are more prevalent among ITP patients.[2,3] Pediatric ITP may be fundamentally different from adult ITP since the rate of chronic ITP in adults is much higher than that in children.[4] Symptoms like platelet aggregations in ITP patients may be partly treated by anti-platelet glycoprotein VI phage antibodies,[5,6] while the phages originated as bacterial virulence factors.[7]

ITP has three clinical phases.[1] The first 3 months after the diagnosis is the newly diagnosed phase. The second phase refers to persistent ITP lasting between 3 and 12 months after diagnosis. The last phase is the chronic ITP phase, in which the patient carries the symptoms beyond 12 months. The first phase is sometimes called acute ITP, and patients may develop severe bleeding symptoms that require immediate interventions.[8] Most adult ITP patients will progress into the chronic phase.[9] The heterogeneous causes of thrombocytopenia make ITP diagnosis a major challenge in haematology.[10]

Different molecular biomarkers are observed to be associated with ITP diagnosis and prognosis. Most of the transcriptomic biomarkers are investigated in T cells. The interleukin (IL) genes IL-10 and IL-17 were differentially expressed in CD4+ T cells in the corticosteroid refractory ITP.[11] The C-X-C motif chemokine ligand 13 (CXCL13) has elevated expression levels in the plasma of ITP children, and its transcription regulation may be repressed by the CD4+ T cells with miR-125-5p inhibitors.[12] Serum proteins may also serve as good indicators of ITP treatment responses. The protein level of haptoglobin (Hp) in serums is measured by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometer (MS) technology and is observed to be positively correlated with the platelet count after the invasive treatment splenectomy.[13]

This study presents the largest cohort of platelet ITP samples and generates RNA sequencing (RNA-seq) transcriptomes for the

Received 23 December 2021; accepted 1 April 2022;
https://doi.org/10.1016/j.omtn.2022.04.004.

[4]These authors contributed equally

Correspondence: Zhaoyang Hu, PhD, Fengneng Pharmaceutical Technology (Shanghai) Co., Ltd., RM. A310, 115 Xinjunhuan Road, Minhang District, Shanghai 201100, China.
E-mail: zyhu310@fun-med.cn
Correspondence: Fengfeng Zhou, PhD, College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China.
E-mail: FengfengZhou@gmail.com
Correspondence: Wenwei Zhu, PhD, Department of Hematology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, 110 Ganhe Road, Hongkou District, Shanghai 200437, China.
E-mail: wwz3006@163.com

Table 1. Summary of the recruited cohort

	ITP		Control	
	Female	Male	Female	Male
Samples	46	13	31	24
Averaged age	55.69565	62.84615	48.19355	47.375

There are two groups of participants: the ITP patients and the controls. Each group consists of female and male samples. The numbers of samples are given in the row "Samples," and the averaged ages of these sample groups are in the row "Averaged age."

detection of ITP biomarkers. A procedure of feature selection, feature engineering, and classification is comprehensively evaluated. The best ITP detection model using only 40 transcriptome features is delivered, and it achieves an overall accuracy 0.974. Some interesting biological inferences are also discussed.

RESULTS

Sample summarization

This study recruited a cohort of 59 ITP patients and 55 controls, as shown in Table 1. There were 46 female and 13 male ITP patients. The control group consists of 31 female and 24 male samples. The control samples were recruited to match the age distribution of the ITP patients, with the t test p value 0.4752 (>0.05).

The experimental protocol of this study was approved by the ethics committee of the Yueyang Hospital. All participants in this study signed the informed-consent forms. This study was approved by the Ethical Committee of the Yueyang Hospital in accordance with the 1964 Helsinki Declaration. Written informed consent was obtained from all participants.

Comparison of different classification algorithms

Firstly, we fixed the feature-selection module with the following parameter-value choices for the RNA-seq transcriptome data. The transcriptome features were screened by the L1-regularization algorithm least absolute shrinkage and selection operator (Lasso), and only the features with non-zero weights were kept for further analysis. Then, the pairwise evaluation of the inter-feature Pearson correlation coefficient (PCC) was carried out. For a pair of features, F(i) and F(j), F(j) was removed if PCC(F(i), F(j)) > threshold and weight(i) > weight(j), where weight(i) and weight(j) were the Lasso weights of these two features.

We used a stratified method to split the samples into a training set (90%) and an independent test set (10%). Stratified k-fold cross validation is a variation of the k-fold cross validation that returns the stratified folds by preserving the same percentage of samples for each class in each fold. On the training set, we utilize the stratified 5-fold cross validation (S5FCV) strategy to train and tune the parameters and, finally, tested our model on the test set. As shown in Figure 1, all five classifiers achieved at least 0.884 in comparison of accuracy (ACC), suggesting that the biomarker-detection module in this study works effectively. The classifiers Logistic Regression (LR) and Support-Vector Machine (SVM) achieved the best receiver operating characteristic (ROC) (0.978) for the ITP diagnosis model, while another classifier, Random Forest (RF), achieved the best ACC (0.935). The classifier RF also performed the best in Sn (0.915) and Sp (0.962). LR only achieved an Sn of 0.881, which may not be a good choice with more mis-classified ITP patients (the positive samples). So, the following sections used RF as the default classifier.

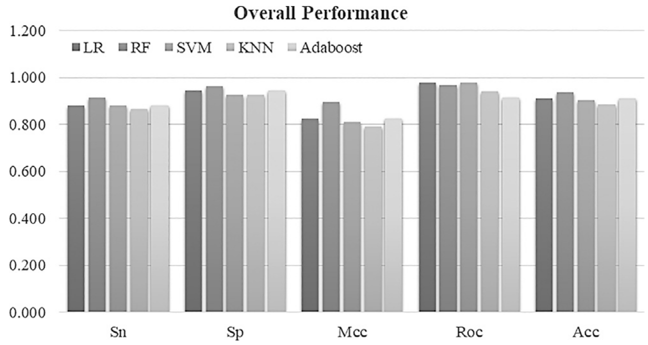


Figure 1. Evaluation of different classifiers on the ITP diagnosis problem

The five classifiers are evaluated, including LR, RF, SVM, KNN, and AdaBoost. The horizontal axis lists the classification performance metrics, i.e., Sn, Sp, MCC, ROC, and Acc. The vertical axis is the value of the performance metrics. Feature engineering and step 5 optimizes the diagnosis model. All boxed text is functional module names utilized in the pipeline.

Comparison of different feature-selection algorithms

Different feature-selection algorithms were evaluated for the S5FCV performances of the default classifier RF, as shown in Figure 2. The framework SelectFromModel of the Python package sklearn with default parameters was carried out to evaluate the six feature-selection algorithms in Figure 2, i.e., Ttest, Ftest, Ridge, Lasso, Variance Threshold (VThresh), and Laplacian score (LS). Lasso and Ridge achieved the best ACC (0.935) for the ITP diagnosis model, but Ridge had the worst performance on Sp (0.925). Ttest and Ridge performed similarly well in ACC and ROC. Ridge outperformed Ttest in Sn, and Ttest outperformed Ridge in Sp. Ttest performed the worst in Sn among the four supervised feature-selection algorithms, and the data suggested that this popular statistical-evaluation algorithm didn't work well on selecting a subset of features with the best classification performance. The performance of the two unsupervised feature-selection algorithms was lower than that of all the supervised learning feature-selection algorithms. Figure 3A gives the Ttest ranks of the 50 Lasso-selected features. Only 11 features are ranked as top 50 by Ttest, the last feature is ranked 23,696 out of the total 33,493 features, and 34 out of 50 (68.0%) Lasso-selected features are ranked more than 100. Figure 3B demonstrates that the performance of the prediction model keeps being improved by adding the 50 Lasso-selected features one by one in their Ttest ranks. This observation and Figure 2 suggest that Lasso may select a feature subset with better prediction performances than Ttest, Ftest, and Ridge.

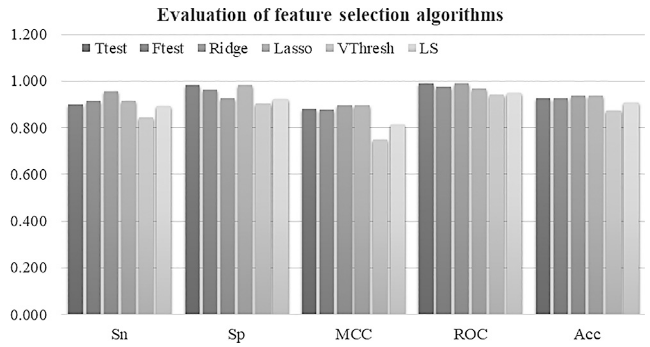


Figure 2. Performances of different feature-selection algorithms on detecting the ITP biomarkers

The four evaluated feature-selection algorithms are Ttest, Ftest, Ridge, Lasso, VThresh, and LS. The horizontal axis lists the performance metrics Sn, Sp, MCC, ROC, and Acc. The vertical axis is the value of these performance metrics.

Evaluation of the contributions of LDA and SVD

The 50 Lasso-selected transcriptome features (denoted as Lasso50) were further refined by two feature-engineering algorithms, linear discriminant analysis (LDA) and singular value decomposition (SVD), as shown in Figure 4. SVD calculated the first two components as the engineered features, while for the LDA, the number of components needs to be less than the number of the class minus one; therefore, we get one component for LDA. LDA + SVD denotes the union of the one LDA component and the two SVD components. The Lasso50-based prediction model achieved an ACC of 0.929. LDA achieved the same score to the Lasso50 in ACC, and LDA + SVD improved the LDA model by 0.017 in Sn. The engineered component features of LDA + SVD performed better (0.008 in Sn) than the Lasso50-based model, but both LDA (0.002) and LDA + SVD (0.019) models outperformed the Lasso50-based model in Matthews correlation coefficient (MCC). We looked into the detailed percentage of variance explained by each SVD component, as shown in Figure 4B. The data suggested that the first SVD component alone explained 38% of the total sample variances, and the first two components explained almost half of the sample variances, so it is reasonable to observe that only the first two components of LDA and SVD may simultaneously improve model performance and reduce feature dimensions.

Dot plots were generated to demonstrate the discriminative powers of the SVD components and the raw transcriptome features, as shown in Figure 5. Figure 5A gives an intuitive illustration that the first two SVD components separate well the two groups of samples, while the top-two Ttest-ranked features generate some mis-classifications, as shown in Figure 5B. It is interesting to see that both of the top-two Ttest-ranked features are from mitochondrion. The first ranked feature ENSG00000210082 encodes a ribosomal RNA gene MT-RNR2, and the second ranked feature ENSG00000198888 encodes the protein-coding gene MT-ND1. This observation supports the previous observation that ITP patients demonstrated various platelet mitochondrial abnormalities,[15,16] but the Ttest-ranked features may be further improved by more sophisticated feature selections and engineering algorithms for their ITP prediction performances.

Optimization the parameters of the prediction model

The pipeline in this study was further refined by evaluating different value choices for the following three parameters, as shown in Figures 6 and S1. All prediction-performance metrics were calculated using the S5FCV strategy. The parameter nFeatures is the number of features with the largest weights assigned and selected by Lasso. There are five values, {30, 40, 50, 60, 70}, for this parameter. The second parameter, nComponents, is the number of the first few components calculated by both LDA and SVD, and its value choices are {1, 2, 3, 4, 5}. The third parameter, nEstimators, is the number of decision trees used in the classifier RF. The five values {100, 200, 300, 400, 500} are evaluated for nEstimators. A grid-search strategy was carried out for all value combinations of these three parameters. Due to the limited space in this work, the detailed data are illustrated in Figures 6 and S1.

Figure 6 illustrates the value choice nFeatures = 40 when the best accuracy, ACC = 0.956, was achieved. This best model chose nFeatures = 40, nComponents = 2, and nEstimators = 100 and achieved Sn = 0.932, Sp = 0.982, and MCC = 0.914.

The five classifiers were stacked as a better classifier

A stacking prediction strategy was utilized to build a better ITP prediction model, and its prediction performance was compared with the previous best model in Figure 7. The prediction results of all the five classifiers LR/SVM/K-nearest neighbor (KNN)/adaptive boost (AdaBoost) were loaded as the input to an additional classifier to generate the final prediction result. The additional classifier was also selected in LR/SVM/KNN/AdaBoost/RF. The samples were randomly split into a training dataset (80%) and a test dataset (20%). Each classifier was exerted on the training dataset for 20 random runs, and its prediction result was averaged over the 20 runs. Overall, the final stacked RF model achieved ACC = 0.974. Figure 7 demonstrated that the stacking model (StackingModel) outperformed the best model in the above sections (PrevBestModel) on four out of five metrics (0.051 in Sn, 0.054 in MCC, 0.06 in ROC, and 0.039 in ACC).

In this study, in order to better verify the ability of the stacking model, except for the above-mentioned five machine-learning models, we used five additional machine-learning models, including Linear Regression (LiR),[17] Extremely Randomized Trees (ET),[18] Gaussian Process Classifier (GP),[19] Gradient Boosting Classifier (GB),[20] and Naive Bayes (NB).[19] We found that the prediction model only achieved 0.962 in ACC, which is lower than the previous best ACC = 0.974. The experimental results suggested that more meta-learners may not deliver better results.

Biological inferences from the detected biomarkers

The final model used 40 RNA-seq transcriptome features of which 19 features are annotated as protein-coding genes and 3 are as long intergenic non-coding RNAs (lincRNAs), as shown in Table 2. It is

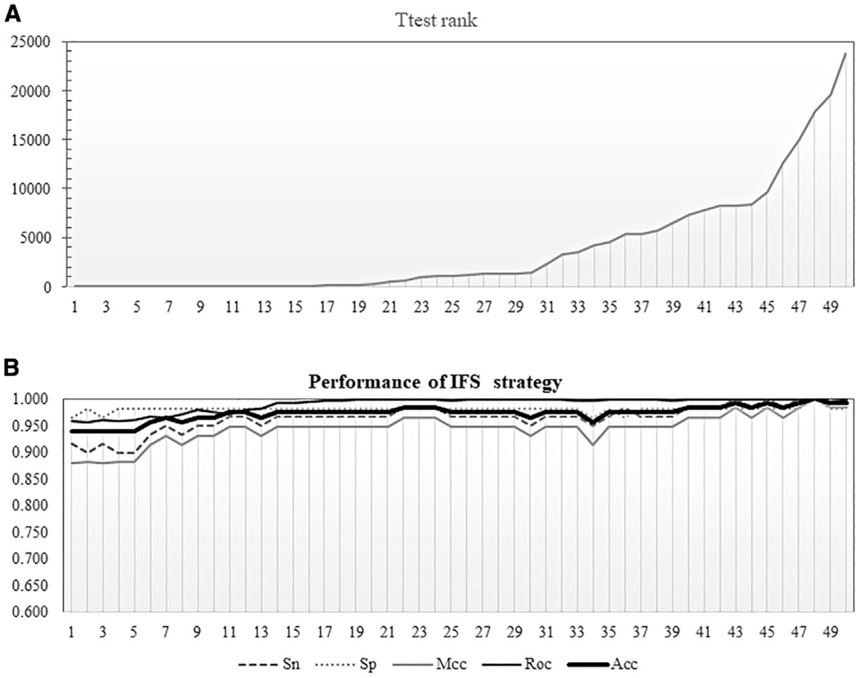


Figure 3. Ttest evaluation of the 50 Lasso-selected features

(A) The Ttest ranks of the 50 features selected by Lasso. The vertical axis is the Ttest rank of each of the 50 Lasso-selected features.

(B) The S5FCV prediction performances of the Incremental Feature Selection (IFS)[14] strategy of the 50 Lasso-selected features. The vertical axis gives the value of the prediction-performance metrics. The horizontal axis lists the 50 Lasso-selected features by their Ttest ranks in ascending order for both subfigures.

interesting to observe that 15 biomarkers are annotated as pseudogenes, but their expressions contribute to the ITP classification.

Firstly, we evaluated the enriched Gene Ontology (GO) categories of the 40 detected biomarkers using the online system DAVID.[21] The statistical significance p value was corrected by the Bonferroni method.[22] The biological processes GO: 0010729 (positive regulation of hydrogen peroxide biosynthetic process), 0042773 (ATP synthesis coupled electron transport), and 1900118 (negative regulation of execution phase of apoptosis) were enriched in the 40 biomarkers with the Bonferroni-corrected p values 8.63×10^{-4}, 1.21×10^{-3}, and 1.08×10^{-2}, respectively. ITP was known to be associated with the platelet apoptosis[23] but remained to be confirmed for its connections with hydrogen-peroxide synthesis and ATP synthesis. It is also interesting to observe that the molecular function GO: 0048019 (receptor antagonist activity) was also enriched in the 40 biomarkers, as confirmed by the elevated plasma levels of IL-1 receptor antagonist (Ra) in acute pediatric[24] and adult ITP patients.[25]

T cell dysfunction may stimulate auto-antibody productions in ITP.[26,27] The biomarker gene AANAT (aralkylamine N-acetyltransferase; transcriptome feature Ensembl: ENSG00000129673) is located on chromosome 17 and showed rhythmic expressions on the expression levels and phosphorylation levels in T cells of the bone marrow and spleen.[28] Another biomarker gene, TNFSF14 (tumor necrosis factor superfamily member 14; transcriptome feature Ensembl: ENSG00000125735), facilitates the increased production of CD8 central memory t cells *in vivo.*[29] Manresa et al. observed that the increased production of TNFSF14 by T cells induced the transcription of inflammatory genes in the esophageal fibroblasts in eosinophilic esophagitis.[30]

One of the symptoms of ITP is low platelet production in the blood, and ITP is mainly diagnosed by a low platelet count in a blood test.[31,32] The detected biomarker gene MAOB (monoamine oxidase; transcriptome feature Ensembl: ENSG00000069535) was not observed to be associated with ITP, but its low activities in platelets are associated with various impulsive behaviors[33] and alcoholism.[34] A case study demonstrated the connection between the gene monoamine oxidase inhibitor (MAOI) and the resolving of non-ITP.[35] So, the aberrant expression of MAOB in platelets of ITP patients may be worth further experimental validation.

The altered methylation of the biomarker mitochondrial gene MT-CO2 (mitochondrially encoded cytochrome-C-oxidase II; transcriptome feature Ensembl: ENSG00000198712) in platelets is associated with the prognosis of various complex diseases, e.g., cardiovascular disease [36,37] and Parkinson's disease.[38] This study demonstrated that the expression level of MT-CO2 is significantly associated with ITP by $p = 6.78 \times 10^{-26}$. MT-CO2 may serve as a candidate diagnosis biomarker of ITP on the transcriptome level.

Enriched functions in the close neighbors of the biomarkers

We further investigated the direct neighbors to the 40 detected biomarkers based on the STRING database.[39] There were 5,296 genes interacting with the 40 detected biomarkers. We assumed that these direct neighbors were closely associated with these ITP biomarkers and the functions enriched in these neighbors together with the biomarkers. DAVID does not allow for the analysis of more than 3,000 genes, so this section used the tool ShinyGO [40] to detect the enriched functions within these 5,296 + 40 = 5,336 genes. There were 79 molecular functions enriched (Bonferroni-corrected p value < 0.05) in the gene list, as shown in the Table S1. It is interesting to obsere that the molecular function GO: 0009055 (electron transfer activity) was enriched with the Bonferroni-corrected p value = 1.18×10^{-26}. The top-ranked GO terms included binding capabilities to small molecule (GO: 0036094, corrected $p = 2.91 \times 10^{-33}$), RNA (GO: 0003723, corrected $p = 1.47 \times 10^{-25}$) and nucleoside phosphage (GO: 1901265, corrected $p = 7.88 \times 10^{-24}$), etc.

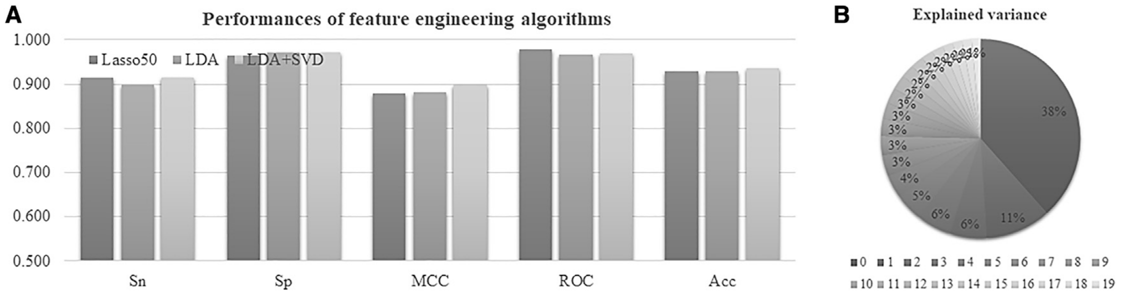


Figure 4. Prediction performances of the feature-engineering algorithms LDA and SVD

(A) The horizontal axis lists the prediction performance metrics Sn, Sp, MCC, ROC, and ACC. The vertical axis is the value of these performance metrics.

(B) The percentage of variance explained by each of the first 20 SVD components.

DISCUSSION

This study screened ITP biomarkers using RNA-seq platelet transcriptomes. A series of comprehensive machine-learning algorithms were utilized to find the best subset of biomarkers. The final ITP prediction model was based on 40 transcriptome features and two clinical variables (age and sex) and achieved an overall accuracy of 0.974. It is of interest to observe that both protein-coding genes and lincRNA genes contribute to the ITP prediction model. Some of the biomarker genes are closely associated with the two major symptoms of ITPs, i.e., T cell dysfunctions and aberrant platelet activities.

The gene-expression patterns of all 40 biomarker genes across different human tissues were obtained from the GTEx database,[41] as shown in Figure S2. Many biomarkers showed highly tissue-specific expression patterns, e.g., the genes DNAH7 and AANAT were only highly expressed in the testis, while the gene KLHDC8A was highly expressed only in the ovary. Multiple genes showed high expressions in different brain regions, including DNAH10OS, NORAD, MT-ATP8, HNRNPUL2, MT-RNR2, and MT-CO2, but most of the 40 biomarkers were expressed at relatively low levels in the whole blood. Combined with their ITP-specific expression patterns, the data suggested that the abnormal expressions of these tissue-specific expressed genes might have contributed to ITP's onset and progression, and it is important to investigate ITP's molecular mechanisms using the platelet cells.

As far as we know, this study presents the largest cohort of ITP RNA-seq platelet transcriptomes. This dataset and these experimental data may facilitate a better understanding of ITP onset and developmental mechanisms.

MATERIALS AND METHODS

Ethics approval and consent to participate

This study was approved by the Ethical Committee of the Yueyang Hospital in accordance with the 1964 Helsinki Declaration. Written informed consent was obtained from all participants.

Clinical sample collection

This cohort of ITP patients and healthy controls was recruited between August 10, 2017, and February 21, 2019, at the Shanghai Yueyang Integrated Traditional Chinese Medicine and Western Medicine Hospital (abbreviated as Yueyang Hospital), affiliated with the Shanghai University of Traditional Chinese Medicine. This study was approved by the Ethical Committee of the Yueyang Hospital in accordance with the 1964 Helsinki Declaration. Written informed consent was obtained from all participants.

The disease ITP was diagnosed based on the criteria of the American Society of Hematology.[42] Thrombocytopenia was defined as a platelet count <100 × 10^9 platelets per liter. All ITP patients enrolled in this

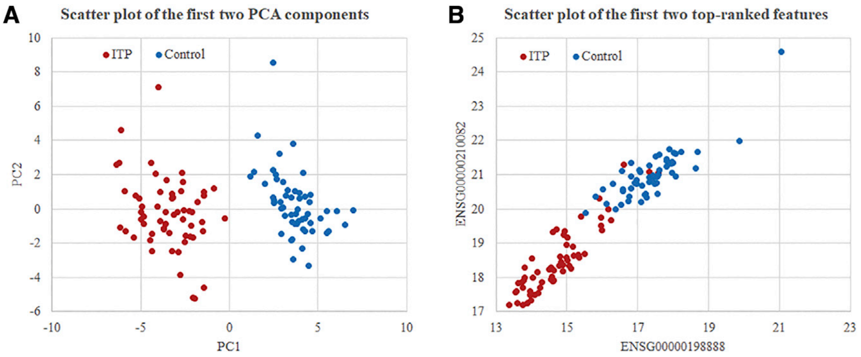
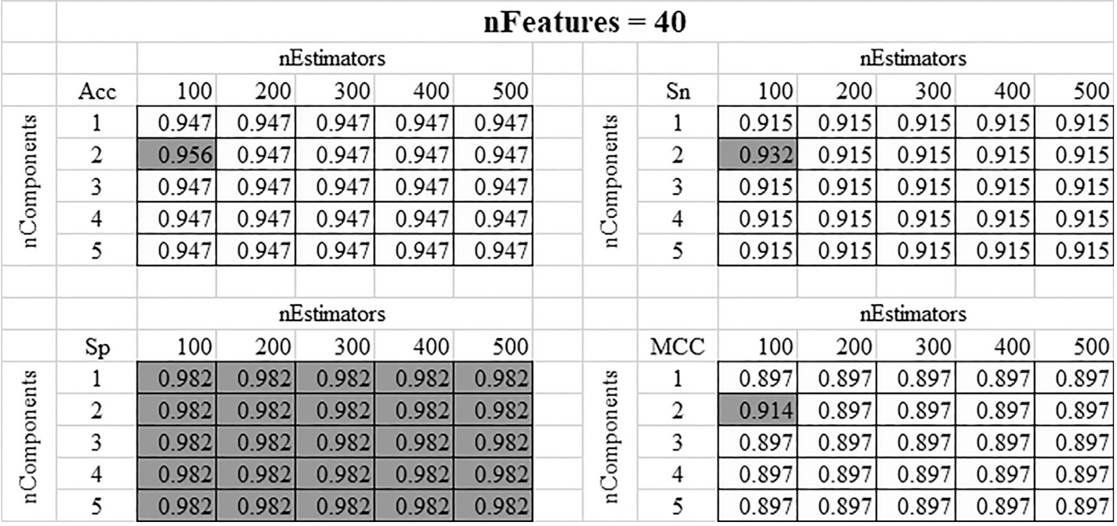


Figure 5. Dot-plot visualization of SVD and Ttest

(A) Dot plot of the first two components calculated by SVD. The horizontal and vertical axes are the first and second components calculated by SVD.

(B) Dot plot of the top two features ranked by Ttest. The horizontal and vertical axes are the top-ranked first and second features by Ttest. Dots represent the ITP samples and controls in red and blue, respectively.



nFeatures = 40

nComponents	Acc	100	200	300	400	500
	1	0.947	0.947	0.947	0.947	0.947
	2	0.956	0.947	0.947	0.947	0.947
	3	0.947	0.947	0.947	0.947	0.947
	4	0.947	0.947	0.947	0.947	0.947
	5	0.947	0.947	0.947	0.947	0.947

nComponents	Sn	100	200	300	400	500
	1	0.915	0.915	0.915	0.915	0.915
	2	0.932	0.915	0.915	0.915	0.915
	3	0.915	0.915	0.915	0.915	0.915
	4	0.915	0.915	0.915	0.915	0.915
	5	0.915	0.915	0.915	0.915	0.915

nComponents	Sp	100	200	300	400	500
	1	0.982	0.982	0.982	0.982	0.982
	2	0.982	0.982	0.982	0.982	0.982
	3	0.982	0.982	0.982	0.982	0.982
	4	0.982	0.982	0.982	0.982	0.982
	5	0.982	0.982	0.982	0.982	0.982

nComponents	MCC	100	200	300	400	500
	1	0.897	0.897	0.897	0.897	0.897
	2	0.914	0.897	0.897	0.897	0.897
	3	0.897	0.897	0.897	0.897	0.897
	4	0.897	0.897	0.897	0.897	0.897
	5	0.897	0.897	0.897	0.897	0.897

(Column headers 100–500 are under "nEstimators" in each panel.)

Figure 6. Optimizing the parameters of nComponents and nEstimators when nFeatures = 40

The matrices of the different parameter choices are colored in the grayscale background. A darker background suggests a better value of that metric.

study either had no treatment history or had not received glucocorticoids for at least 3 months. Patients were excluded from this cohort if they carried these complications, i.e., diabetes, hypertension, cardiovascular diseases, pregnancy, active infection, or autoimmune diseases other than ITP.

Platelet isolation and RNA extraction

Peripheral whole-blood samples of all recruited participants were drawn and kept in tubes with ethylenediamine tetraacetic acid (EDTA) at the Yueyang Hospital. To maintain platelet RNA quantity and quality, the samples were collected and processed for platelets within 24 h. The platelet-rich plasma was processed by a 20-min, 1-kilo RPM (kRPM) centrifugation at 4°C for platelet isolation. To avoid hemocyte contamination in the platelets, only 9/10th platelet-rich plasma was transferred to the 1.5-mL tubes. Then, the platelet samples were pelleted by a 20-min, 3-kRPM centrifugation. The platelets were then white precipitated by removing the supernatant. The platelets were washed with PBS and extracted with 15-min 3-kRPM centrifugation. The PBS was removed by a 10-μL pipette. The Thermo Scientific's stabilizer RNAlater was used to resuspend the precipitated platelets. The extracts were frozen at -20°C. A hematology analyzer was used to quantitatively measure the hemocyte contamination in the samples.

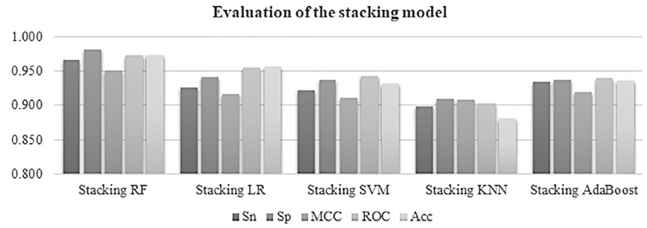


Figure 7. Performance of the stacking model

The horizontal axis lists the prediction-performance metrics. The vertical axis is the value of these performance metrics. The two series of histograms are for the stacking model (StackingModel) and the best prediction model in the above sections (PrevBestModel).

This study used Eppendorf tubes.

Sample purity

The sampling quality-control method was utilized to process sequencing samples. Quality control was performed on 10 separated platelet samples, and the average purity obtained by the test was 93.7%.

Library construction and mRNA sequencing

The cDNA library construction and mRNA sequencing were carried out following the procedures described in the previous study.[43]

Sequencing data processing and differential analysis

The software FASTP was used to preprocess raw sequencing data in the format fastq through quality control and adapter/primer removals.[44] Clean data were aligned to human Ensembl v.91.

FASTP was used to purify raw data of fastq format through removal of adapters, PCR primers, and low-quality reads. Clean data were aligned to human Ensembl 91 using the software STAR v.2.4.2a.[45] The generated alignment files in the format BAM were loaded to the software HTSeq v.0.6.1 to estimate the gene-expression levels.[46] Differential expression analysis was performed by the R Bioconductor package DESeq2.[47] The parameters of DESeq2 were set to base mean >100, |log2(fold change)| (|log2(FC)|) >1, and false discovery rate (FDR) <0.05.

Table 2. Gene annotations of the 40 detected biomarkers

ENSG_code	Chr	Gene	Gene type
ENSG00000259834	1	AL365361.1	lincRNA
ENSG00000213026	1	CFL1P4	processed_pseudogene
ENSG00000162873	1	KLHDC8A	protein_coding
ENSG00000229344	1	MTCO2P12	unprocessed_pseudogene
ENSG00000213110	2	AC019178.1	processed_pseudogene
ENSG00000226247	2	SUPT4H1P1	processed_pseudogene
ENSG00000229758	2	DYNLT3P2	processed_pseudogene
ENSG00000118997	2	DNAH7	protein_coding
ENSG00000269028	3	MTRNR2L12	protein_coding
ENSG00000248360	4	LINC00504	lincRNA
ENSG00000223908	5	AC068657.1	processed_pseudogene
ENSG00000271043	5	MTRNR2L2	processed_pseudogene
ENSG00000164576	5	SAP30L	protein_coding
ENSG00000196821	6	C6orf106	protein_coding
ENSG00000146587	7	RBAK	protein_coding
ENSG00000253276	7	CCDC71L	protein_coding
ENSG00000226824	7	AC006001.2	sense_intronic
ENSG00000229897	9	SEPT7P7	processed_pseudogene
ENSG00000175787	9	ZNF169	protein_coding
ENSG00000213260	10	YWHAZP5	processed_pseudogene
ENSG00000254616	11	AP001775.1	processed_pseudogene
ENSG00000188997	11	KCTD21	protein_coding
ENSG00000214753	11	HNRNPUL2	protein_coding
ENSG00000258359	12	PCNPP1	processed_pseudogene
ENSG00000250091	12	DNAH10OS	protein_coding
ENSG00000205240	13	OR7E36P	unprocessed_pseudogene
ENSG00000140254	15	DUOXA1	protein_coding
ENSG00000263177	16	MTND1P8	processed_pseudogene
ENSG00000235554	17	AC005822.1	processed_pseudogene
ENSG00000129673	17	AANAT	protein_coding
ENSG00000185262	17	UBALD2	protein_coding
ENSG00000267541	18	MTCO2P2	processed_pseudogene
ENSG00000125735	19	TNFSF14	protein_coding
ENSG00000260032	20	NORAD	lincRNA
ENSG00000210082	MT	MT-RNR2	Mt_rRNA
ENSG00000210049	MT	MT-TF	Mt_tRNA
ENSG00000198712	MT	MT-CO2	protein_coding
ENSG00000198786	MT	MT-ND5	protein_coding
ENSG00000228253	MT	MT-ATP8	protein_coding
ENSG00000069535	X	MAOB	protein_coding

The column “ENSG_code” lists the ensemble IDs of features generated in the RNA-seq transcriptome processing. The columns “Chr” and “Gene” give the chromosome this gene locates and the gene symbol. The column “Gene type” is which category this gene belongs to.

Feature-selection algorithms

Feature-selection algorithms were utilized to find a set of biomarkers with the best discrimination power of ITP samples from the controls.[48] The so-called Occam’s Razor principle suggests that a simpler model is preferred over a complicated one if these two models perform similarly.[49,50] A feature-selection algorithm may significantly reduce the dimensionality of the transcriptome datasets for a binary disease diagnosis model.[51–54] It is also anticipated that the training and prediction times of a disease diagnosis model will be shortened by selecting a subset of features from the transcriptome dataset. This study evaluated the following feature-selection algorithms.

The analysis of variance (ANOVA) algorithm measures the difference between the means of two groups of samples.[55] ANOVA uses F-test to calculate the statistical significance of rejecting the null hypothesis of the mean equality. The F-test assumes that the data follow the F-distribution.[56]

T test (Ttest) is a popular statistical test when a feature has different values in two groups of samples with the null hypothesis that the two variables have the same normal distribution.[57,58] The statistical p value measures how probable it is that the two variables will follow the same normal distribution. Ttest has been widely used in many different data types, including electrocardiograms (ECGs),[58] RNA-seq,[59] and metaproteomics.[60]

Ridge (Ridge) is a regression model that assigns balanced weights to the features,[61] and the features are ranked in descending order by their weights, while the Lasso algorithm exerts both regularization and feature selection.[62,63] Lasso assigns very sparse weights to the features, and only the features with non-zero weights are selected for the regression model. The features with non-zero weights are ranked in descending order by their weights.

The above feature-selection algorithms are all based on supervised learning. To avoid the bias of supervised signals, we also evaluated two unsupervised feature-selection algorithms. VThresh is an unsupervised feature-selection algorithm that removes all low-variance features. LS[64] utilizes the observation that samples from the same class are often similar to each other. It evaluates the power of locality preserved within the candidate feature subsets.

Binary-classification algorithms

The binary-classification problem was investigated between ITP patients and the control samples. The following binary classifiers are utilized for this task.

The statistical classifier LR uses the logistic function to describe the binary-classification task.[65,66] The logistic function evaluates the existence probability of ITP disease.

SVM is another popular binary classifier.[67] SVM attempts to maximize the margins of a hyperplane between the control samples and the patients. SVM may be used on various data types, e.g., one-hot encoded data,[68] imaging data,[69] etc.

KNN is another simple and effective supervised learner.[70,71] KNN evaluates the distances of the query samples against all training samples and collects the top k-closest neighbors to make the prediction. The query sample is assigned to the most frequently appearing class label of these top k-closest neighbors.

RF is a tree-based ensemble-learning algorithm.[72] This classifier RF assembles the predictions of multiple decision tree classifiers, and the final result of RF is determined by the majority voting strategy.

Another ensemble-classification framework, AdaBoost, is also evaluated for its prediction performance on the investigated ITP diagnosis problem. AdaBoost may be used with many supervised-learning algorithms, and this study chose the decision tree to work with AdaBoost. AdaBoost iteratively tunes the weight of multiple weakly boosted classifiers and may perform less susceptibility to the overfitting challenge.[73]

Feature-engineering algorithms

The final step of the experimental procedures is to calculate new features with enriched discrimination information from the original features.[74–77] Some datasets may be very sparse, and the individual features may not contribute significant discrimination powers to the final prediction models. This study hypothesizes that the feature-decomposition technique may generate new features with enriched discrimination powers.

Firstly, this study uses truncated SVD.[78,79] SVD calculates the singular value decomposition without centering the data matrix and performs very well with sparse data. Also, SVD may efficiently calculate the user-defined number of singular values.

The second feature-engineering algorithm is the LDA.[80] LDA doesn't calculate the overall covariance matrix and may calculate the singular values for large datasets.

Both SVD and LDA may be used to calculate new features, and the feature dimension may be reduced by choosing the first few engineered features.[81]

Deep learning has become a research hotspot in recent years for its powerful representation-learning capability. Representation-learning-based disease diagnosis models are also gaining increased attention in the literature. Su et al.[82] proposed the Siamese response deep factorization machine (SRDFM) algorithm, based on a Siamese network, to learn a feature vector of the drug property and gene expression for personalized anti-cancer drug recommendations. Peng et al.[83] proposed a novel subnetwork representation-learning method to uncover disease-disease relationships. Lv et al.[84] used deep-representation-learning features for the identification of sub-Golgi protein localization. It is also worth mentioning that deep-learning-based representation-learning models are often poor in the interpretability of learned features.

Performance evaluation metrics

A binary-classification model is usually evaluated by the following variables and performance metrics.[65,85,86] ITP patients are defined as positive samples, and their number is defined as P. The controls are negative samples, and their number is N. ITP patients are true positives (TPs) if they are predicted as the ITP class; otherwise, they are the false negatives (FNs). The control samples are defined as true negatives (TNs) if they are correctly predicted as controls; otherwise, they are defined as false positives (FP). The numbers of samples in these groups of samples are also denoted as TPs, FNs, TNs, and FPs.

A binary classifier is evaluated by its specificity, sensitivity, and accuracy. Specificity is the rate of correctly predicted controls, i.e., $Sp = TN/(TN + FP)$. Sensitivity is defined as the rate of correctly predicted ITP patients, i.e., $Sn = TP/(TP + FN)$. The overall prediction accuracy is defined as $ACC=(TN + TP)/(FP + TN + FN + TP)$. All three metrics are between 0 and 1. A better prediction model has larger values for Sp, Sn, and ACC.

The ROC curve is a 2-dimensional plot between Sn and (1-Sp).[87] The area under the ROC curve is defined as the area under the curve (AUC) value for a binary-classification model. The metric AUC is widely used to measure a binary-classification model independent of model thresholds.

Stacking weak classifiers

Stacking is an effective classification and regression model used in machine learning.[88] It builds multiple base models (meta-learner) in multiple layers so that the output of the previous model can be used as the input of the model of the later layer. In recent years, machine-learning models based on stacking technology have also been widely used in disease modeling. StackTADB is a stacking model for predicting the boundaries of topologically associating domains (TADs) accurately. Wu et al.[89] and Khoei et al.[90] propose a stacking-based ensemble-learning model with a genetic algorithm for detecting early stages of Alzheimer's disease. Rahman et al.[91] propose a coronavirus disease 2019 (COVID-19) detection system based on a stacking model. The main difference between our model and theirs is that feature selection is integrated into the model building process and forms an end-to-end learning process.

The overall pipeline of diagnosis model of ITP

This study processed two clinical features separately from the RNA-seq transcriptomes. The RNA-seq transcriptomes and the feature age were normalized by standardization. The feature sex was a category variable and was formatted by the one-hot encoding strategy.[92]

Due to the high dimensionality of the RNA-seq transcriptomes, an additional processing of feature selection was carried out on the transcriptomes. The RNA-seq transcriptome has 33,493 features, and

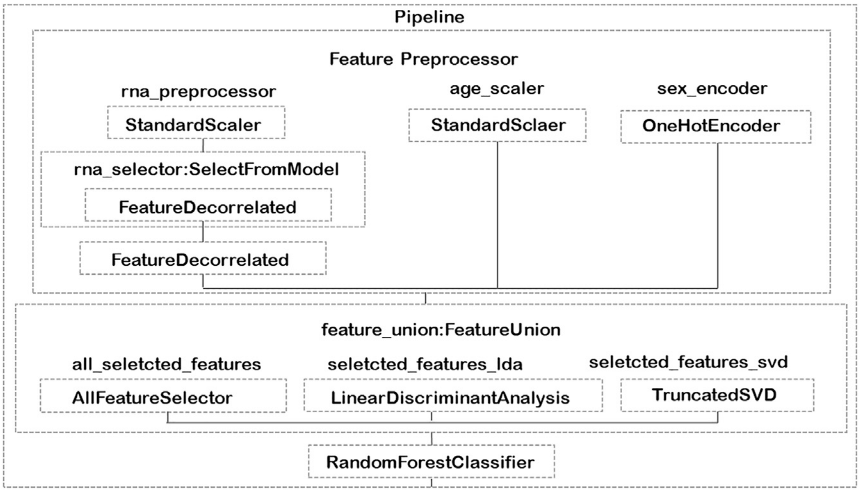


Figure 8. Workflow of ITP diagnosis modeling in this study

Step 1 normalizes the data, step 2 selects the best features, step 3 fuses the transcriptome feature with age and sex, step 4 features engineering, and step 5 optimizes the diagnosis model. All boxed text is the functional module names utilized in the pipeline.

there are only 114 samples in total. Most of these transcribed features are involved in various biological processes that are not associated with ITP, so the L1-regularization algorithm Lasso was used to remove those features with zero weights in the regression model with the class label. The redundant features were removed according to the inter-feature PCCs. Then, the normalized feature age and the encoded feature sex were integrated with the selected RNA-seq transcriptome features. The feature-engineering algorithms SVD and LDA were used to further enrich the information into fewer singular values.

An ensembled classifier, RF, was used to build the final ITP diagnosis model by integrating the results of multiple first-line classification models. Our experimental data suggest that the ensembled classifier demonstrated a better performance than the first-line classifiers. The workflow of the ITP diagnosis modeling in this study can been seen in Figure 8.

Availability of data and material

Detailed information on optimizing the parameters of nFeatures, nComponents, and nEstimators can been found in Figure S1. The dataset used in this study was archived in the NCBI SRA database. The full dataset of the platelet RNA-seq next-generation sequencing (NGS) data for the 59 ITP patients was archived in the NCBI SRA database with the project accession NCBI: PRJNA664615. The RNA-seq NGS data of the 55 health controls is available as the project accession NCBI: PRJNA668820. Researchers of interest may explore their scientific hypothesis in this dataset after the embargo period. The source code is released at the web site http://www.healthinformaticslab.org/supp/. Any future collaborations are welcome.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2022.04.004.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China (no. 81973798) and the Shanghai Clinical Key Department Development Program (shslczdzk05201). We really appreciate the insightful comments that helped improve the manuscript!

AUTHOR CONTRIBUTIONS

F.Z., Z.H., and W.Z. conceived the project and designed the experiments. C.X. and R.Z. carried out the experiments and data analysis. M.D. worked on the programming and helped with data analysis. Y.Z., J.B., H.L., J.W., and M.H. were involved in the collection and annotation of the data.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Rodeghiero, F., Stasi, R., Gernsheimer, T., Michel, M., Provan, D., Arnold, D.M., Bussel, J.B., Cines, D.B., Chong, B.H., Cooper, N., et al. (2009). Standardization of terminology, definitions and outcome criteria in immune thrombocytopenic purpura of adults and children: report from an international working group. Blood *113*, 2386–2393.
2. Fogarty, P.F. (2009). Chronic immune thrombocytopenia in adults: epidemiology and clinical presentation. Hematol. Oncol. Clin. North Am. *23*, 1213–1221.
3. Moulis, G., Palmaro, A., Montastruc, J.L., Godeau, B., Lapeyre-Mestre, M., and Sailler, L. (2014). Epidemiology of incident immune thrombocytopenia: a nationwide population-based study in France. Blood *124*, 3308–3315.
4. Schulze, H., and Gaedicke, G. (2011). Immune thrombocytopenia in children and adults: what's the same, what's different? Haematologica *96*, 1739–1741.
5. Liu, Q., Zhang, C., Yu, L., Shi, Y., Zhang, L., Peng, J., Ji, X., and Hou, M. (2016). Study of a humanized inhibitory anti-platelet glycoprotein VI phage antibody from a phage antibody library. Hematology *21*, 60–67.
6. Zhou, F., Olman, V., and Xu, Y. (2009). Large-scale analyses of glycosylation in cellulases. Genomics Proteomics Bioinformatics *7*, 194–199.
7. Liu, B., Zhou, F., Wu, S., Xu, Y., and Zhang, X. (2009). Genomic and proteomic characterization of a thermophilic Geobacillus bacteriophage GBSV1. Res. Microbiol. *160*, 166–171.
8. Frelinger, A.L., 3rd, Grace, R.F., Gerrits, A.J., Berny-Lang, M.A., Brown, T., Carmichael, S.L., Neufeld, E.J., and Michelson, A.D. (2015). Platelet function tests,

independent of platelet count, are associated with bleeding severity in ITP. Blood *126*, 873–879.

9. Cuker, A., Prak, E.T., and Cines, D.B. (2015). Can immune thrombocytopenia be cured with medical therapy? Semin. Thromb. Hemost. *41*, 395–404.

10. Hicks, S.M., Coupland, L.A., Jahangiri, A., Choi, P.Y., and Gardiner, E.E. (2020). Novel scientific approaches and future research directions in understanding ITP. Platelets *31*, 315–321.

11. Stimpson, M.L., Lait, P.J.P., Schewitz-Bowers, L.P., Williams, E.L., Thirlwall, K.F., Lee, R.W.J., and Bradbury, C.A. (2020). IL-10 and IL-17 expression by CD4(+) T cells is altered in corticosteroid refractory immune thrombocytopenia (ITP). J. Thromb. Haemost. *18*, 2712–2720.

12. Li, J.Q., Hu, S.Y., Wang, Z.Y., Lin, J., Jian, S., Dong, Y.C., Wu, X.F., Lan, D., and Cao, L.J. (2015). MicroRNA-125-5p targeted CXCL13: a potential biomarker associated with immune thrombocytopenia. Am. J. Transl Res. *7*, 772–780.

13. Zheng, C.X., Ji, Z.Q., Zhang, L.J., Wen, Q., Chen, L.H., Yu, J.F., and Zheng, D. (2012). Proteomics-based identification of haptoglobin as a favourable serum biomarker for predicting long-term response to splenectomy in patients with primary immune thrombocytopenia. J. Transl Med. *10*, 208.

14. Liu, H., and Setiono, R. (1998). Incremental feature selection. Applied Intelligence *9*, 217–230.

15. Deng, G., Yu, S., Li, Q., He, Y., Liang, W., Yu, L., Xu, D., Sun, T., Zhang, R., and Li, Q. (2017). Investigation of platelet apoptosis in adult patients with chronic immune thrombocytopenia. Hematology *22*, 155–161.

16. Qiao, J., Liu, Y., Li, D., Wu, Y., Li, X., Yao, Y., Niu, M., Fu, C., Li, H., Ma, P., et al. (2016). Imbalanced expression of Bcl-xL and Bax in platelets treated with plasma from immune thrombocytopenia. Immunol. Res. *64*, 604–609.

17. Zhang, T. (2004). Solving Large Scale Linear Prediction Problems Using Stochastic Gradient Descent Algorithms (Proceedings of the twenty-first international conference on Machine learning).

18. Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Machine Learn. *63*, 3–42, %@ 1573-0565.

19. Zhang, H. (2004). The optimality of naive Bayes. Aa *1*, 3.

20. Friedman, J.H. (2002). Stochastic gradient boosting. Comput. Stat. Data Anal. *38*, 367–378, %@ 0167-9473.

21. Jiao, X., Sherman, B.T., Huang da, W., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics *28*, 1805–1806.

22. Cho, H.H., Lee, H.Y., Kim, E., Lee, G., Kim, J., Kwon, J., and Park, H. (2021). Radiomics-guided deep neural networks stratify lung adenocarcinoma prognosis from CT scans. Commun. Biol. *4*, 1286.

23. Zheng, S.S., Ahmadi, Z., Leung, H.H.L., Wong, R., Yan, F., Perdomo, J.S., et al. (2022). Antiplatelet antibody predicts platelet desialylation and apoptosis in immune thrombocytopenia. Haematologica, 1592–8721.

24. Goelz, N., Bosch, A.M.S., Rand, M.L., Eekels, J.J.M., Franzoso, F.D., and Schmugge, M. (2020). Increased levels of IL-10 and IL-1Ra counterbalance the proinflammatory cytokine pattern in acute pediatric immune thrombocytopenia. Cytokine *130*, 155078.

25. Yadav, D.K., Tripathi, A.K., Gupta, D., Shukla, S., Singh, A.K., Kumar, A., Agarwal, J., and Prasad, K.N. (2017). Interleukin-1B (IL-1B-31 and IL-1B-511) and interleukin-1 receptor antagonist (IL-1Ra) gene polymorphisms in primary immune thrombocytopenia. Blood Res. *52*, 264–269.

26. Yu, J., Heck, S., Patel, V., Levan, J., Yu, Y., Bussel, J.B., and Yazdanbakhsh, K. (2008). Defective circulating CD25 regulatory T cells in patients with chronic immune thrombocytopenic purpura. Blood *112*, 1325–1328.

27. Semple, J.W., and Freedman, J. (1991). Increased antiplatelet T helper lymphocyte reactivity in patients with autoimmune thrombocytopenia. Blood *78*, 2619–2625.

28. Cordoba-Moreno, M.O., de Souza, E.D.S., Quiles, C.L., Dos Santos-Silva, D., Kinker, G.S., Muxel, S.M., Markus, R.P., and Fernandes, P.A. (2020). Rhythmic expression of the melatonergic biosynthetic pathway and its differential modulation in vitro by LPS and IL10 in bone marrow and spleen. Sci. Rep. *10*, 4799.

29. Fernandez, M.F., Qiao, G., Tulla, K., Prabhakar, B.S., and Maker, A.V. (2021). Combination immunotherapy with LIGHT and interleukin-2 increases CD8 central memory T-cells in vivo. J. Surg. Res. *263*, 44–52.

30. Manresa, M.C., Chiang, A.W.T., Kurten, R.C., Dohil, R., Brickner, H., Dohil, L., Herro, R., Akuthota, P., Lewis, N.E., Croft, M., et al. (2020). Increased production of LIGHT by T cells in eosinophilic esophagitis promotes differentiation of esophageal fibroblasts toward an inflammatory phenotype. Gastroenterology *159*, 1778–1792.e1713.

31. Woolley, P., Newton, R., Mc Guckin, S., Thomas, M., Westwood, J.P., and Scully, M.A. (2020). Immune thrombocytopenia in adults: a single-centre review of demographics, clinical features and treatment outcomes. Eur. J. Haematol. *105*, 344–351.

32. Zaninetti, C., and Greinacher, A. (2020). Diagnosis of inherited platelet disorders on a blood smear. J. Clin. Med. *9*, 539.

33. Lewitzka, U., Muller-Oerlinghausen, B., Felber, W., Brunner, J., Hawellek, B., Rujescu, D., Ising, M., Lauterbach, E., Broocks, A., Bondy, B., et al. (2008). Is MAO-B activity in platelets associated with the occurrence of suicidality and behavioural personality traits in depressed patients? Acta Psychiatr. Scand. *117*, 41–49.

34. Parsian, A., Suarez, B.K., Tabakoff, B., Hoffman, P., Ovchinnikova, L., Fisher, L., and Cloninger, C.R. (1995). Monoamine oxidases and alcoholism. I. Studies in unrelated alcoholics and normal controls. Am. J. Med. Genet. *60*, 409–416.

35. Brubacher, J.R., Hoffman, R.S., and Lurin, M.J. (1996). Serotonin syndrome from venlafaxine-tranylcypromine interaction. Vet. Hum. Toxicol. *38*, 358–361.

36. Baccarelli, A.A., and Byun, H.M. (2015). Platelet mitochondrial DNA methylation: a potential new marker of cardiovascular disease. Clin. Epigenetics *7*, 44.

37. Corsi, S., Iodice, S., Vigna, L., Cayir, A., Mathers, J.C., Bollati, V., and Byun, H.M. (2020). Platelet mitochondrial DNA methylation predicts future cardiovascular outcome in adults with overweight and obesity. Clin. Epigenetics *12*, 29.

38. Sharma, A., Schaefer, S.T., Sae-Lee, C., Byun, H.M., and Wullner, U. (2020). Elevated serum mitochondrial DNA in females and lack of altered platelet mitochondrial methylation in patients with Parkinson s disease. Int. J. Neurosci. 1–4.

39. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. *49*, D605–D612.

40. Ge, S.X., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics *36*, 2628–2629.

41. Consortium, G.T. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.

42. Thrombosis, and Hemostasis Group; HSCMA (2016). [Consensus of Chinese experts on diagnosis and treatment of adult primary immune thrombocytopenia (version 2016)]. Zhonghua Xue Ye Xue Za Zhi *37*, 89–93.

43. Best, M.G., Sol, N., 't Veld, S., Vancura, A., Muller, M., Niemeijer, A.N., Fejes, A.V., Tjon Kon Fat, L.A., Huis in 't Veld, A.E., Leurs, C., et al. (2017). Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. Cancer Cell *32*, 238–252.e239.

44. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

45. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

46. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

47. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

48. Byun, S., Kim, A.Y., Jang, E.H., Kim, S., Choi, K.W., Yu, H.Y., and Jeon, H.J. (2019). Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol. Comput. Biol. Med. *112*, 103381.

49. Alonso-Betanzos, A., Bolon-Canedo, V., Moran-Fernandez, L., and Sanchez-Marono, N. (2019). A review of microarray datasets: where to find them and specific characteristics. Methods Mol. Biol. *1986*, 65–85.

50. Bickel, D.R. (2019). Sharpen statistical significance: evidence thresholds and Bayes factors sharpened into Occam's razor. Stat *8*, e215.

51. Yang, S., Li, B., Zhang, Y., Duan, M., Liu, S., Zhang, Y., Feng, X., Tan, R., Huang, L., and Zhou, F. (2020). Selection of features for patient-independent detection of seizure events using scalp EEG signals. Comput. Biol. Med. *119*, 103671.

52. Gao, X., Liu, S., Song, H., Feng, X., Duan, M., Huang, L., and Zhou, F. (2020). AgeGuess, a methylomic prediction model for human ages. Front Bioeng. Biotechnol. *8*, 80.

53. Han, Y., Huang, L., and Zhou, F. (2021). A dynamic recursive feature elimination framework (dRFE) to further refine a set of OMIC biomarkers. Bioinformatics *37*, 2183–2189.

54. Wei, Z., Ding, S., Duan, M., Liu, S., Huang, L., and Zhou, F. (2020). FeSTwo, a two-step feature selection algorithm based on feature engineering and sampling for the chronological age regression problem. Comput. Biol. Med. *125*, 104008.

55. Giusca, S., Wolf, D., Hofmann, N., Hagstotz, S., Forschner, M., Schueler, M., Nunninger, P., Kelle, S., and Korosoglou, G. (2020). Splenic switch-off for determining the optimal dosage for adenosine stress cardiac MR in terms of stress effectiveness and patient safety. J. Magn. Reson. Imaging *52*, 1732–1742.

56. Dos Santos, R.A., de Lima, E.A., Mendonca, L.S., de Oliveira, J.E., Rizuto, A.V., de Araujo Silva Tavares, A.F., and Braz da Silva, R. (2019). Can universal adhesive systems bond to zirconia? J. Esthet Restor Dent *31*, 589–594.

57. Soh, D.C.K., Ng, E.Y.K., Jahmunah, V., Oh, S.L., San, T.R., and Acharya, U.R. (2020). A computational intelligence tool for the detection of hypertension using empirical mode decomposition. Comput. Biol. Med. *118*, 103630.

58. Peng, T., Trew, M.L., and Malik, A. (2019). Predictive modeling of drug effects on electrocardiograms. Comput. Biol. Med. *108*, 332–344.

59. Tran, S.S., Zhou, Q., and Xiao, X. (2020). Statistical inference of differential RNA-editing sites from RNA-sequencing data by hierarchical modeling. Bioinformatics *36*, 2796–2804.

60. Tang, J., Wang, Y., Fu, J., Zhou, Y., Luo, Y., Zhang, Y., Li, B., Yang, Q., Xue, W., Lou, Y., et al. (2019). A critical assessment of the feature selection methods used for biomarker discovery in current metaproteomics studies. Brief Bioinform *21*, 1378–1390.

61. Wang, M., Li, R., and Xu, S. (2020). Deshrinking ridge regression for genome-wide association studies. Bioinformatics *36*, 4154–4162.

62. Xu, W., Liu, X., Leng, F., and Li, W. (2020). Blood-based multi-tissue gene expression inference with Bayesian ridge regression. Bioinformatics *36*, 3788–3794.

63. Waldmann, P., Ferencakovic, M., Meszaros, G., Khayatzadeh, N., Curik, I., and Solkner, J. (2019). AUTALASSO: an automatic adaptive LASSO for genome-wide prediction. BMC Bioinformatics *20*, 167.

64. He, X., Cai, D., and Niyogi, P. (2005). Laplacian score for feature selection. Adv. Neural Inf. Process. Syst. *18*.

65. Meng, C., Hu, Y., Zhang, Y., and Guo, F. (2020). PSBP-SVM: a machine learning-based computational identifier for predicting polystyrene binding peptides. Front Bioeng. Biotechnol. *8*, 245.

66. Cuadrado-Godia, E., Jamthikar, A.D., Gupta, D., Khanna, N.N., Araki, T., Maniruzzaman, M., Saba, L., Nicolaides, A., Sharma, A., Omerzu, T., et al. (2019). Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: logistic regression approach. Comput. Biol. Med. *108*, 182–195.

67. Lin, E., Lin, C.H., Hung, C.C., and Lane, H.Y. (2020). An ensemble approach to predict schizophrenia using protein data in the N-methyl-D-Aspartate receptor (NMDAR) and tryptophan catabolic pathways. Front Bioeng. Biotechnol. *8*, 569.

68. ang, A., Wang, J., Lin, H., Zhang, J., Yang, Z., and Xu, K. (2017). A multiple distributed representation method based on neural network for biomedical event extraction. BMC Med. Inform. Decis. Mak *17*, 171.

69. Hosseini, M.P., Tran, T.X., Pompili, D., Elisevich, K., and Soltanian-Zadeh, H. (2020). Multimodal data analysis of epileptic EEG and rs-fMRI via deep learning and edge computing. Artif. Intell. Med. *104*, 101813.

70. Jia, C., Bi, Y., Chen, J., Leier, A., Li, F., and Song, J. (2020). PASSION: an ensemble neural network approach for identifying the binding sites of RBPs on circRNAs. Bioinformatics *36*, 4276–4282.

71. Zhang, S., Zhao, L., Zheng, C.H., and Xia, J. (2020). A feature-based approach to predict hot spots in protein-DNA binding interfaces. Brief Bioinform *21*, 1038–1046.

72. Vogel, I., Blanshard, R.C., and Hoffmann, E.R. (2019). SureTypeSC-a Random Forest and Gaussian mixture predictor of high confidence genotypes in single-cell data. Bioinformatics *35*, 5055–5062.

73. Hu, X., Cheng, Y., Ding, D., and Chu, D. (2018). Axis-Guided vessel segmentation using a self-constructing cascade-AdaBoost-SVM classifier. Biomed. Res. Int. *2018*, 3636180.

74. Alvarez-Machancoses, O., Fernandez-Martinez, J.L., and Kloczkowski, A. (2020). Prediction of protein tertiary structure via regularized template classification techniques. Molecules *25*, 2467.

75. Kornaropoulos, E.N., Niazi, M.K., Lozanski, G., and Gurcan, M.N. (2014). Histopathological image analysis for centroblasts classification through dimensionality reduction approaches. Cytometry A *85*, 242–255.

76. Cai, J., Xu, Y., Zhang, W., Ding, S., Sun, Y., Lyu, J., et al. (2020). A comprehensive comparison of residue-level methylation levels with the regression-based gene-level methylation estimations by ReGear. Brief Bioinform. *22*, bbaa253.

77. Duan, M., Song, H., Wang, C., Zheng, J., Xie, H., He, Y., Huang, L., and Zhou, F. (2020). Detection and independent validation of model-based quantitative transcriptional regulation relationships altered in lung cancers. Front Bioeng. Biotechnol. *8*, 582.

78. Chen, C.H., Hsieh, J.G., Cheng, S.L., Lin, Y.L., Lin, P.H., and Jeng, J.H. (2020). Early short-term prediction of emergency department length of stay using natural language processing for low-acuity outpatients. Am. J. Emerg. Med. *38*, 2368–2373.

79. Gupta, Y., Lama, R.K., Kwon, G.R., and Alzheimer's Disease Neuroimaging, I. (2019). Prediction and classification of alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers. Front Comput. Neurosci. *13*, 72.

80. Reddy, G.T., Reddy, M.P.K., Lakshmanna, K., Kaluri, R., Rajput, D.S., Srivastava, G., and Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. IEEE Access *8*, 54776–54788.

81. Howland, P., Wang, J., and Park, H. (2006). Solving the small sample size problem in face recognition using generalized discriminant analysis. Pattern Recognition *39*, 277–287.

82. Su, R., Huang, Y., Zhang, D.G., Xiao, G., and Wei, L. (2022). SRDFM: siamese response deep factorization machine to improve anti-cancer drug recommendation. Brief. Bioinform. *23*, bbab534.

83. Peng, J., Guan, J., Hui, W., and Shang, X. (2021). A novel subnetwork representation learning method for uncovering disease-disease relationships. Methods *192*, 77–84, %@ 1046-2023.

84. Lv, Z., Wang, P., Zou, Q., and Jiang, Q. (2020). Identification of sub-Golgi protein localization by use of deep representation learning features. Bioinformatics *36*, 5600–5609, %@ 1367-4803.

85. Lv, Z., Jin, S., Ding, H., and Zou, Q. (2019). A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. Front Bioeng. Biotechnol. *7*, 215.

86. Li, T., Song, R., Yin, Q., Gao, M., and Chen, Y. (2019). Identification of S-nitrosylation sites based on multiple features combination. Sci. Rep. *9*, 3098.

87. Meng, C., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. Front Bioeng. Biotechnol. *7*, 224.

88. Pavlyshenko, B. (2018). Using Stacking Approaches for Machine Learning Models (IEEE), pp. 255–258, %@ 1538628740.

89. Wu, H., Zhang, P., Ai, Z., Wei, L., Zhang, H., Yang, F., and Cui, L. (2022). StackTADB: a stacking-based ensemble learning model for predicting the boundaries of topologically associating domains (TADs) accurately in fruit flies. Brief. Bioinform. *23*, bbac023.

90. Khoei, T.T., Labuhn, M.C., Caleb, T.D., Hu, W.C., and Kaabouch, N. (2021). A Stacking-Based Ensemble Learning Model with Genetic Algorithm for Detecting Early Stages of Alzheimer's Disease (IEEE), pp. 215–222, %@ 166541846X.

91. Rahman, T., Khandakar, A., Abir, F.F., Faisal, M.A.A., Hossain, M.S., Podder, K.K., Abbas, T.O., Alam, M.F., Kashem, S.B., and Islam, M.T. (2022). QCovSML: a reliable COVID-19 detection system using CBC biomarkers by a stacking machine learning model. Comput. Biol. Med. *143*, 105284, %@ 100010-104825.

92. Wang, J., Deng, F., Zeng, F., Shanahan, A.J., Li, W.V., and Zhang, L. (2020). Predicting long-term multicategory cause of death in patients with prostate cancer: random forest versus multinomial model. Am. J. Cancer Res. *10*, 1344–1355.