

Accurate diagnosis of atopic dermatitis by applying random forest and neural networks with transcriptomic data

Weixin Zhou¹, Aimin Li², Caiyun Zhang¹, Yongtao Chen¹, Zifan Li¹, and Ying Lin^{1,3,4,✉}

¹The Second Clinical Medical College, Guangzhou University of Traditional Chinese Medicine, Guangzhou, 510006, China

²Department of Electronic and Information Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, China

³Department of Dermatology, The Second Affiliated Hospital of Guangzhou University of Chinese Medicine (Guangdong Provincial Hospital of Chinese Medicine), Guangzhou, 510120, China

⁴Guangdong Provincial Key Laboratory of Chinese Medicine for Prevention and Treatment of Refractory Chronic Diseases, Guangzhou, 510120, China

✉: Corresponding author. Email: lin19791226@gzuocm.edu.cn

ABSTRACT

Atopic dermatitis (AD) is one of the most common inflammatory skin diseases. But the great heterogeneity of AD makes it difficult to design an accurate diagnostic pipeline based on traditional diagnostic methods. In other words, the AD diagnosis has suffered from an inaccurate bottleneck. Thus, it is necessary to develop a novel and accurate diagnostic model to supplement existing methods. The recent development of advanced gene sequencing technologies enables potential in accurate AD diagnosis. Inspired by this, we developed an accurate AD diagnosis based on transcriptomic data in skin tissue. Using these data of 149 subjects, including AD patients and healthy controls, from Gene Expression Omnibus (GEO) database, we screened differentially expressed genes (DEGs) of AD and identified six critical genes (PPP4R1, SERPINB4, S100A7, S100A9, BTC, and GALNT6) by random forest classifier. In a follow-up study of these genes, we constructed a neural network model (average AUC=0.943) to automatically distinguish subjects with AD from healthy controls. Among these critical genes, we found that PPP4R1 and GALNT6 had never been reported to be associated with AD. Although further replications in other cohorts are needed, our findings suggest that these genes may be developed into useful biomarkers of AD diagnosis and may provide invaluable clues or perspectives for further researches on the pathogenesis of AD.

Subject terms: Transcriptome, Diagnostic markers, Atopic dermatitis, Random forest, Neural network.

Introduction

Atopic dermatitis (AD) is one of the most prevalent inflammatory skin diseases, affecting about 15% of children and 7% of adults in the United States^{1,2}. As reported by the WHO Burden of Disease, AD ranks the first with respect to the Global Burden of Skin Diseases (GBSD)³. Specifically, in addition to suffering AD-associated discomfort in the skin, when compared with healthy individuals, patients with AD have shown to have a higher risk of psychosocial comorbidities, such as depression, anxiety, sleep disorders and suicidal ideation, which brings about colossal economics cost and health threats to human society⁴. Formally, AD is a type of highly heterogeneous disease with different clinical phenotypes in terms of clinical features, severity, and course⁵. Elevation in total serum IgE level, eosinophilia, and allergen-specific IgE are the primary laboratory abnormalities of AD⁶. The pathophysiology is complex and involves a strong genetic predisposition, skin barrier disruption, T-cell driven inflammation, and microbial dysbiosis⁷. Currently, the therapeutic approaches remain limited, leaving difficulties in completely controlling AD in most patients⁸.

Due to the great heterogeneity mentioned above, it is difficult to define AD without its determinant laboratory biomarkers. The current AD clinical diagnostic rate is only 34.5%⁹. This phenomenon is due to the fact that there are no uniform standard diagnostic criteria that cover the entire spectrum of AD patients, and an experienced dermatologist's assessment is considered the gold standard¹⁰. Traditional diagnostic criteria have been proposed to aid diagnosis mainly based on clinical manifestations and medical history. Among such criteria, the sensitivities of Hanifin and Rajka criteria¹¹, Williams criteria¹²⁻¹⁴ and JDA criteria¹⁵ are 48.2%, 32.7%, 79.4% respectively in a hospital-based study⁹. The absence of uniform and accurate diagnostic

criteria for AD limits the interpretability and comparability of clinical trials¹⁶. Moreover, it is often challenging to diagnose AD in adults who did not develop this disease in their childhood¹⁷.

Recently, the rapid development of second-generation sequencing technologies has facilitated the identification of many disease-related marker genes, including AD-related genes; for instance, Filaggrin (FLG) shows potential for screening and prognosis in AD¹⁸; CCL17 shows the highest correlation with AD severity¹⁹; GZMB, CXCL1, and CD274 show diagnostic value in skin tissue²⁰. Furthermore, the gene expression data are increasingly used to establish diagnostic models of diseases with machine learning, especially random forest and artificial neural networks; for instance, heart failure, ulcerative colitis and polycystic ovary syndrome diagnostic models show excellent performance in accurate diagnosis^{21–23}. However, there are currently few studies on prediction analysis using random forest and artificial neural networks based on the skin transcriptome in AD.

In conclusion, it is a challenge to diagnose AD with certainty because subjective clinical observations may not fully appreciate the variable clinical features. Therefore, it will have a potentially high impact on improving the accuracy of AD diagnosis to develop an objective and reliable biomarker for accurate AD diagnosis, especially in adults. In this study, based on the transcriptome of skin tissue, we build a classifier for accurate diagnosis of AD by integrating random forest and artificial neural networks. Firstly, we screen the Gene Expression Omnibus(GEO) database for differentially expressed genes (DEGs) between AD skin lesion samples and normal skin samples from healthy individuals. Then, based on these DEGs, we employ random forest and artificial neural network to select critical genes and calculate the weights of these genes, thus construct a genetic diagnostic model of AD. Finally, we apply this diagnostic model in other test sets to verify the efficiency of the designed neural-networks-based diagnostic pipeline. The flow chart is shown in Fig. 1.

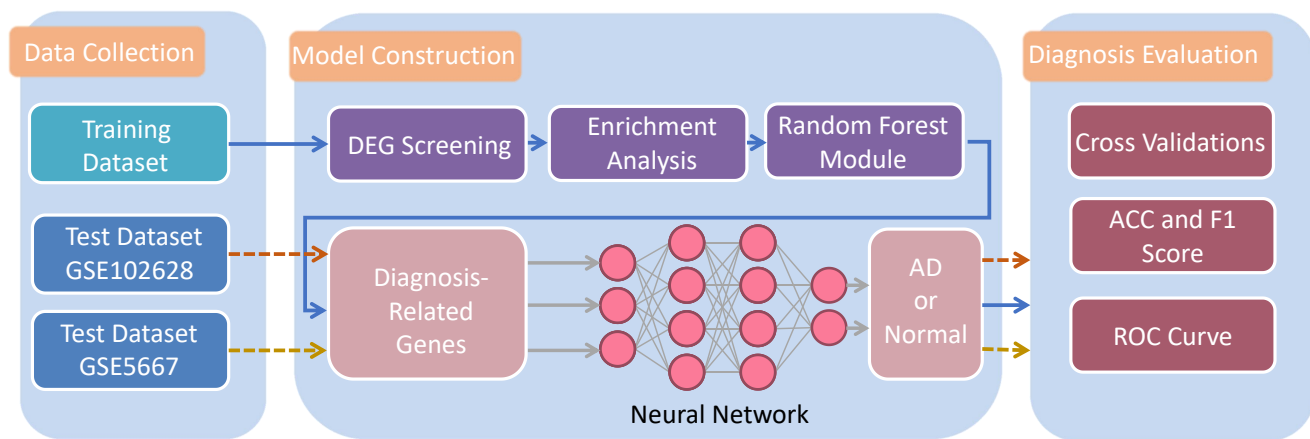


Figure 1. The flow chart in the study.

Materials and Methods

Public data collection

Five microarray datasets and two RNA-seq datasets were collected from the public Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>). The detailed information of the collected datasets has been shown in Table 1. In order to obtain one training dataset with a relatively large sample size, five datasets (GSE6012, GSE32924, GSE60709, GSE121212, GSE153007) were combined and the batch effects were removed with R package "sva"²⁴. Meanwhile, GSE102628 and GSE5667 were used as test datasets respectively. Only AD lesional tissue and normal tissue samples were chosen from the training dataset for further analysis. One test dataset (GSE5667) includes AD lesional tissue and normal tissue samples, while the other test dataset (GSE102628) includes AD non-lesional tissue and normal tissue samples. The patient information in each dataset has been shown in Supplementary Table S1.

DEGs Screening

The training dataset containing 75 cases of AD and 74 cases of healthy control was used for differentially expressed genes (DEGs) analysis. The R package "limma" was used to screen DEGs between the AD and control samples by the classical Bayesian method with adjusted $P < 0.05$ and $|\log_2\text{FoldChange}| > 1$ ²⁵. DEGs screening was visualized by volcano plot and heat map^{26,27}.

Dataset	Total samples	Normal Samples	AD Samples	Data Type	Country	Platform
GSE6012	20	10	10	Microarray	Sweden	GPL96
GSE32924	21	8	13	Microarray	USA	GPL570
GSE60709	26	14	12	Microarray	Germany	GPL6974
GSE121212	65	38	27	RNA-seq	USA	GPL16791
GSE153007	17	5	12	Microarray	USA	GPL6480
GSE102628	19	9	10	RNA-seq	Estonia	GPL20148
GSE5667	11	5	6	Microarray	USA	GPL96

Table 1. Detailed information of the collected GEO datasets.

Enrichment Analyses of DEGs

To determine functions of the identified AD-related DEGs, we performed Disease Ontology (DO), Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses, using R packages "*clusterProfiler*" and "*DOSE*" with $Q < 0.05$ ^{28,29}. The GO analysis included biological process (BP), cellular component (CC) and molecular function (MF) categories. GObubble plots and GOChord plots were conducted with R package "*GOplot*" to illustrate the functional analysis data³⁰.

Diagnosis-Related Gene Identification with Random Forest

As it is unlikely that all the identified DEGs are strongly associated with the occurrence of AD, it is necessary to further extract critical genes from these DEGs, with benefits from both accurate diagnosis and lower classification calculation complexity. Since the Mean Decrease Gini (MDG) method³¹ has shown superiority in screening important genes^{21-23,32}, we leveraged it in the random forest classifier (RFC) to identify critical genes from DEGs.

- **Parameters Selection.** There are two important parameters that will significantly affect the performance of RFC implementation; one of them is called *mytry*, which represents the number of variables for the decision trees in the nodes; the other is denoted as *ntree*, which is the number of decision trees constructing the random forest model. For the optimal parameters setup, we first implemented a recurrent random forest classification for each possible value of *mytry* and calculated the average error rate. The value that enabled the minimal average error rate was selected as the optimal *mytry*. Then, we fixed *mytry* to this optimal value and plotted a figure regarding the relationship between *ntree* and the model error rate. The best *ntree* was finally selected based on a comprehensive consideration of model proficiency, model complexity, and model robustness.
- **MDG Score Ranking.** Based on the above-chosen parameters, we formulated the RFC model. During this process, the MDG score was also output as a default option. The MDG score is a metric that measures how much each input variable contributes to the homogeneity of the nodes and leaves in the generated RFC. Specifically, the higher its MDG score, the more influential the variable is in the model. In our study, we obtained the MDG score of each DEG by adopting R package "*random forest*"³³. We accomplished feature extraction by ranking the DEGs in decreasing order of MDG scores and selecting the top n_{gene} genes as our candidate diagnosis-related genes.

neural-networks-based AD Classifier

Once we screened out the most important genes, we next sought to construct a neural network model to automatically distinguish AD patients from normal people. Particularly, the construction of our designed neural-networks-based AD diagnostic model consists of several critical steps, which will be elaborated in the following.

- **Normalizing.** The inputs of the designed pipeline are the expressions with regard to the selected critical genes of each subject in the collected datasets. Denote the j^{th} critical gene expressions of the i^{th} sample in the collected dataset as $x_{i,j}$ and define the average expression of the j^{th} diagnosis-related gene as $\bar{x}_j \triangleq \sum_{i=1}^{n_{total}} x_{i,j}$, the normalization process is given as

$$y_{i,j} = \begin{cases} \text{sign}(x_{i,j} - \bar{x}_j), & \text{for } \log_2 FC_j \geq 0 \\ -\text{sign}(x_{i,j} - \bar{x}_j), & \text{for } \log_2 FC_j < 0 \end{cases}, \quad 1 \leq i \leq n_{total}, \quad 1 \leq j \leq n_{gene}, \quad (1)$$

where $y_{i,j}$ is the normalized data, n_{total} represents the number of samples, and $\log_2 FC_j$ denotes the $\log_2 \text{FoldChange}$ value of the j^{th} critical gene that has been calculated in the previous subsection **DEGs Screening**, and the function $\text{sign}(\cdot)$ is defined as

$$\text{sign}(x) = \begin{cases} 1, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases}. \quad (2)$$

The above normalization process forms a threshold-segmentation structure. In particular, if the $\log_2\text{FoldChange}$ value of a gene is greater than 0, the gene expression will exert a positive impact on the occurrence of AD, a value greater than the average gene expression is normalized to 1, and the one lower than the average gene expression is normalized to 0; otherwise, the gene expression will exert a negative impact on the occurrence of AD, and the value is normalized in an opposite way. The implementation of thresholding and normalizing mainly brings about two advantages: *i*) the scaled gene expressions are transformed to positive numbers; *ii*) the threshold-segmentation amplifies the effect of up-regulated genes on the pathogenesis, and thus enables improvement in the diagnosis performance of the classifier. Here both the training and test datasets are normalized together.

- **Neural Networks Constructing.** We constructed the neural network by utilizing R packages "neuralnet" and "Neu-

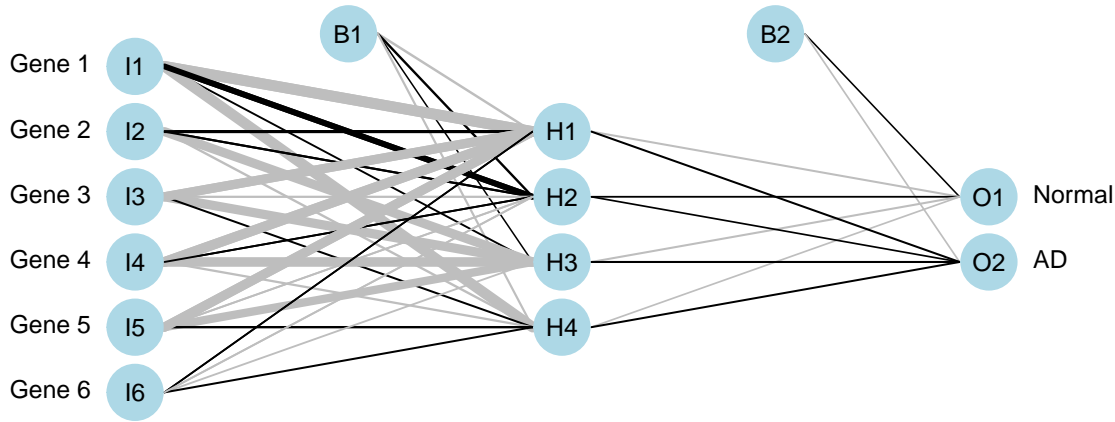


Figure 2. The construction of the neural network.

*ralNetTools*³⁴. The constructed neural network consists of an input layer with $n_{gene} = 6$ neurons, a hidden layer with four neurons, and an output layer with two neurons. The details of the neural network structure were demonstrated in Figure 6 (a). The training process was conducted according to the normalized training data matrix denoted by $\mathbf{Y}_{train} \in \{0, 1\}^{n_{train} \times n_{gene}}$, where n_{train} is the size of the training dataset, n_{gene} is the number of selected diagnosis-related genes. The training dataset is marked with a label vector $\mathbf{V}_{train} = (v_1, v_2, \dots, v_{n_{train}}) \in \{0, 1\}^{1 \times n_{train}}$ with $v_i = 1$ if the i^{th} sample is from AD lesional tissue and $v_i = 0$ if the i^{th} sample is from normal tissue. The goal of the training process is to obtain the weighted matrix \mathbf{W} and the bias vector \mathbf{B} at each layer, so that the output of the neural network turns closer to the label vector \mathbf{V}_{train} . Details of weighted parameters are shown in Supplementary Table S6.

- **Performance Metric.** The results of the binary classification can be directly visualized by a 2×2 matrix, which is known as the confusion matrix (Table 2).

Diagnosis Experiments		Predicted Class	
		AD	Normal
Actual Class	AD	TP (True Positive)	FN (False Negative)
	Normal	FP (False Positive)	TN (True Negative)

Table 2. An example of the confusion matrix.

From a confusion matrix, several critical performance metrics can be obtained straightforwardly, which are True Positive Percentage (TPP), False Positive percentage (FPP), Accuracy (ACC), and F1 Score. We elaborate on these performance metrics as follows.

$$\begin{aligned}
 \text{TPP} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, & \text{ACC} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\
 \text{FPP} &= \frac{\text{FP}}{\text{FP} + \text{TN}}, & \text{F1 Score} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.
 \end{aligned} \tag{3}$$

We mainly focused on ACC and F1 score in this paper to evaluate the performance of the neural-networks-based AD classifier. The ACC is defined as the proportion of correctly classified samples to total samples. In contrast, the F1 score

can provide a more integrated classifier performance evaluation, especially when the dataset is imbalanced. Both ACC and F1 score are real numbers ranging from 0 to 1. The closer the ACC (or F1 score) is to 1, the better the diagnostic model performs.

In addition to the aforementioned metrics, another integrated evaluation method is the Receiver Operating Characteristic (ROC) curve. The ROC curve sketches the relationship between TPP and FPP at various threshold settings. In such cases, the ROC curve not only focuses on a single classifier, but also broadens its evaluations to a series of classifiers that can be generated by different classifying threshold setups. Once we obtain the ROC curve, we can then calculate the area under the ROC curve (AUC). The AUC value also ranges from 0.5 to 1, where the value 1 represents the perfect and efficient diagnosis while 0.5 represents the inapplicable random diagnosis.

- **Cross-Validations on the Training Dataset.** To verify the robustness of our proposed model, we also conducted 6-fold cross-validations (CV) in this work. Specifically, the normalized training dataset \mathbf{Y}_{train} is randomly divided into six uniform sub-matrices, one of which is regarded as a test set while the others are input to the neural network for the training purpose. This 6-fold CV process will be conducted six times, so that each of the six sub-matrices will serve as a test dataset once, and evaluation metrics such as ACC, F1 Score, ROC curve, and AUC will be obtained correspondingly.
- **Validations on Test Dataset.** Up to this point, we have acquired the weighted matrix and the bias vector in the neural network through the training process. The next step is to apply this diagnostic model in other test datasets, so as to verify the efficiency of the designed neural-networks-based diagnostic pipeline. During the validation process, the test matrix \mathbf{Y}_{test} is input to the neural network to generate the predicted results. According to the predicted results, the ACC, F1 Score, ROC curve, and AUC can also be obtained.

Results

Identification of DEGs

A total of 288 DEGs were identified in the training dataset, consisting of 188 up-regulated and 100 down-regulated genes (Fig.3(a) and Supplementary Table S2). The heat map (Fig.3(b)) displays the expression level of each gene in different samples, showing that the overall expression level of DEGs differed greatly between AD lesional skin tissues and normal skin tissues.

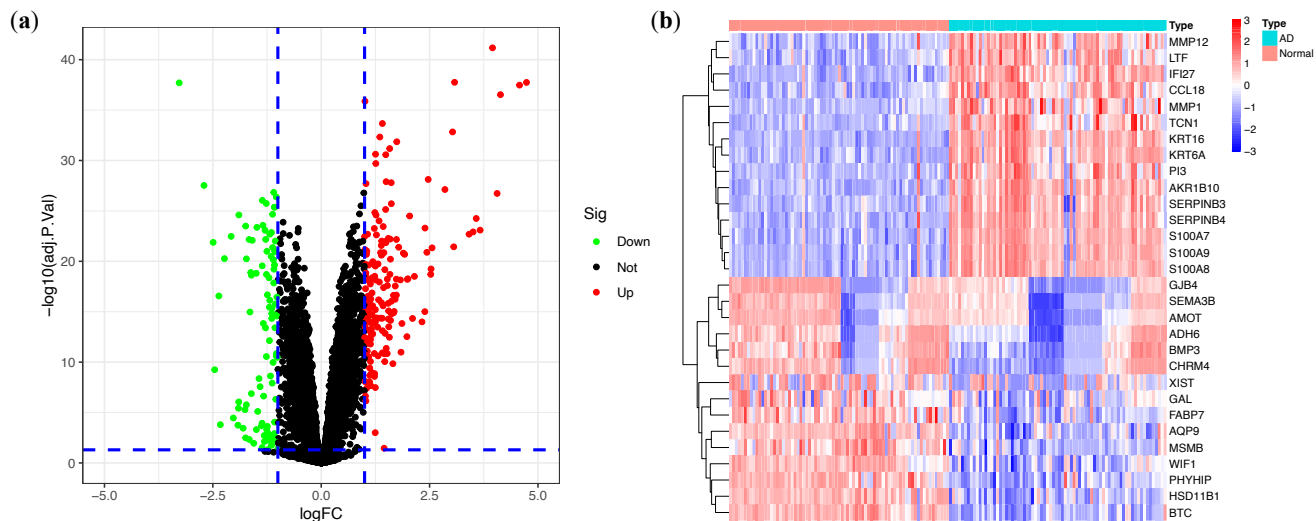


Figure 3. Analyses of DEGs in the training dataset. (a) Volcano plot of DEGs in lesional vs. normal tissues. Red dots represent up-regulated genes and green dots represent down-regulated genes. The black dots indicate the remaining stable genes. DEGs are defined as the cut-off criteria of $|\log_2 \text{FoldChange}| > 1$ and adjusted $P < 0.05$. (b) Heat map of the top 15 up-regulated and down-regulated DEGs in lesional vs. normal tissues. Red represents DEGs with high expression, while blue represents DEGs with low expression.

Functional Enrichment Analysis

To explore the potential related diseases and underlying mechanisms of the 288 DEGs, DO, GO, and KEGG enrichment analyses were conducted. GO enrichment analysis for BP category indicated that the DEGs were enriched in cytokine-mediated signaling pathway, response to virus and granulocyte migration; for CC category, tertiary granule lumen, cornified envelope and external side of plasma membrane were significantly enriched; for MF category, chemokine activity, serine-type endopeptidase activity and serine-type peptidase activity were significantly enriched (Fig.4 (a) and Supplementary Table S3). The KEGG analysis suggested that the DEGs were significantly enriched in Viral protein interaction with cytokine and cytokine receptor, cytokine-cytokine receptor interaction and IL-17 signaling pathway (Fig.4 (c) and Supplementary Table S4). The DO enrichment analysis showed that the DEGs were enriched in integumentary system disease, bacterial infectious disease, atopic dermatitis, etc. (Fig.4 (e) and Supplementary Table S5) Part of GO, KEGG and DO enriched terms and the significant DEGs involved were showed respectively in Fig.4 (b), Fig.4 (d) and Fig.4 (f).

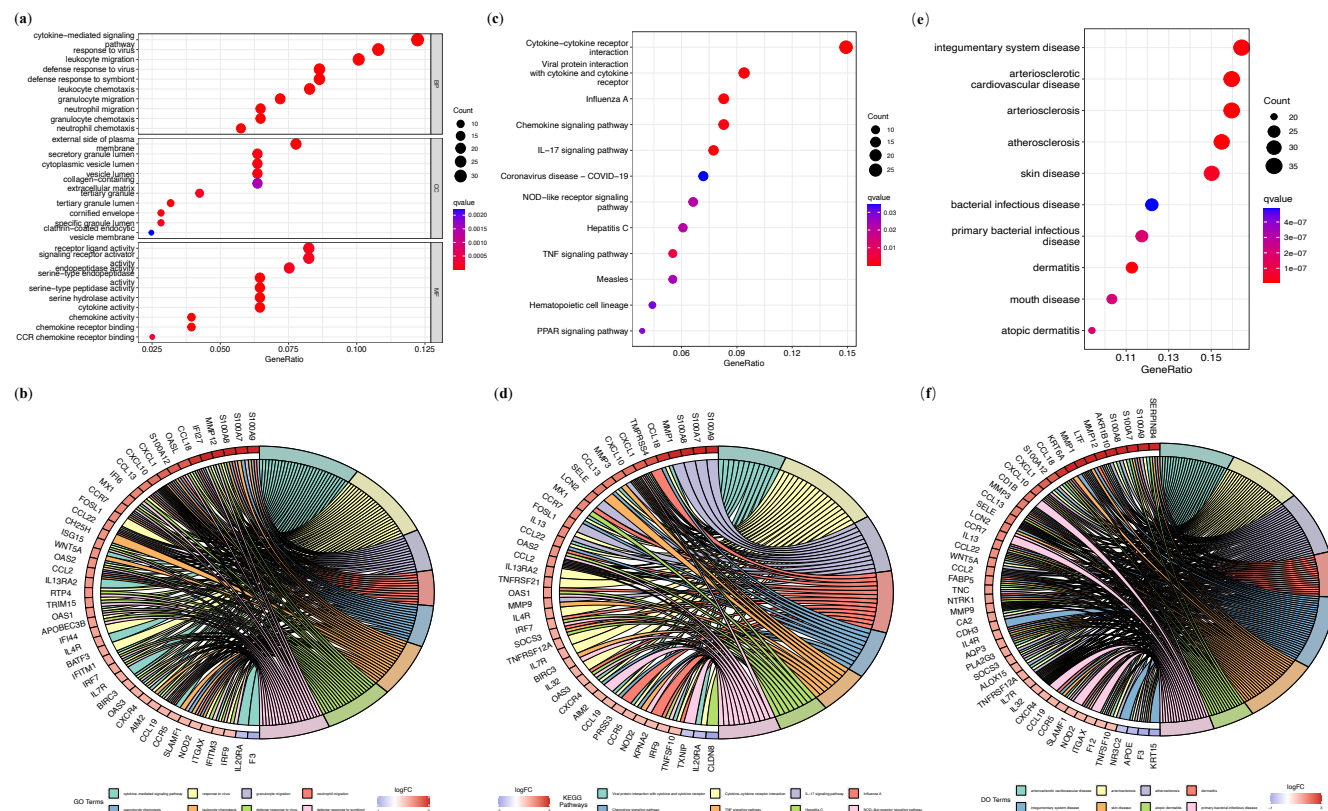


Figure 4. Enrichment Analyses of DEGs in AD. (a) Bubble plot showing significantly enriched GO terms included biological process (BP), cellular component (CC) and molecular function (MF) categories. (b) Chord plot of the top 8 enriched GO terms with minimum Q values. (c) Bubble plot of 12 enriched KEGG pathways. (d) Chord plot of the top 8 enriched KEGG pathways with minimum Q value. (e) Bubble plot of the top 10 enriched DO terms with minimum Q value. (f) Chord plot of the top 8 enriched DO terms with minimum Q value.

The Diagnosis-Related Genes

Fig. 5 (a) and Fig. 5 (b) demonstrate the detailed variables selection results while constructing the random forest. In Fig. 5 (a), the scatter plot is carried out to show the relationship between the out-of-band error rate and $mytry$. As the red point in the figure exhibits the lowest error rate, we finally choose $mytry = 17$ for the best $mytry$ setup. Fig. 5 (b) shows the relationship between n_{tree} and the error rate. We set $n_{tree} = 400$ as the parameter of the RFC model because such a setting demonstrates superiority in both low error rate and stable system performance.

Fig. 5 (c) shows the top 15 genes in MDG scores, wherein the top 6 genes (i.e., PPP4R1, SERPINB4, S100A7, S100A9, BTC, and GALNT6) are selected as the diagnosis-related genes. In Figure. 5 (d), the heatmap is drawn to visually show the expression levels of each selected genes in different samples. From Fig. 5 (d), we can observe that the overall expression levels of the selected genes differed greatly between AD samples and normal samples, which demonstrates that the selected genes

enable potential to be new biomarkers of AD, thus bringing about new insights in the diagnosis of AD.

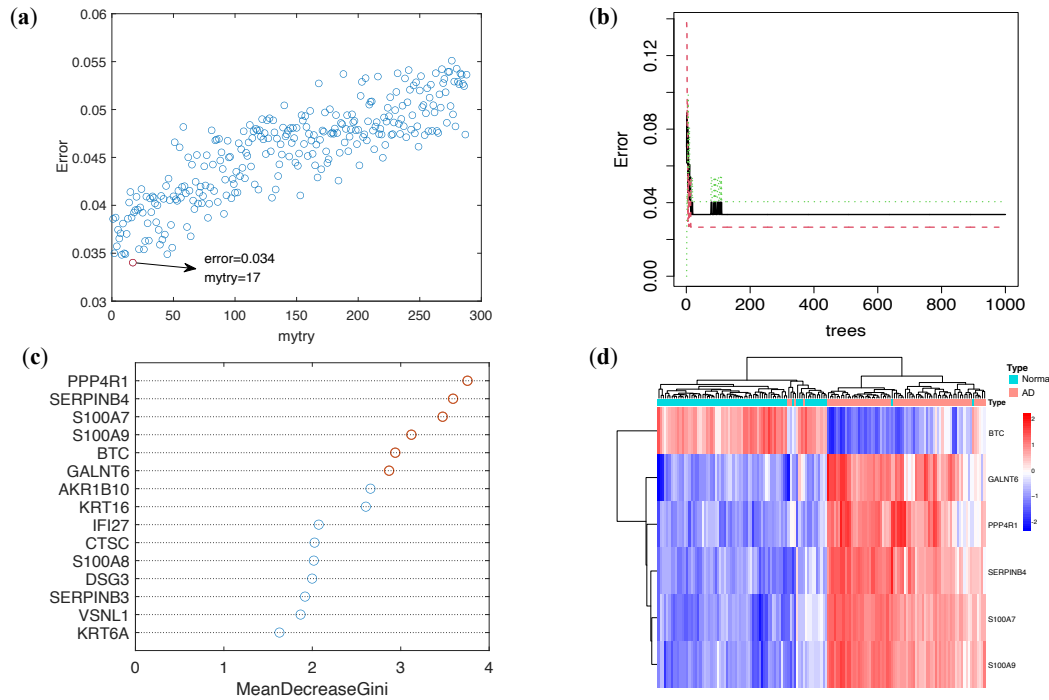


Figure 5. Diagnosis-related genes selection process. (a) Scatter plot regarding the relationship between *mytry* and the out-of-band error rate. (b) Error rate vs. *ntree*. (c) Results of the top 15 MDG scores. The red points indicate MDG scores of the selected diagnosis-related genes. (d) Heatmap of the selected diagnosis-related genes.

Six-Fold Cross-Validations

Cross-validation (CV) is a general-purpose method to evaluate whether a neural network is overfitted³⁵. Therefore, we performed the six-fold CV to evaluate this particular characteristic of our AD-diagnosis-oriented model. To this end, the ROC curves together with the AUC, ACC and F1 Score are shown in this subsection.

Figure. 6 aims to display the robustness of our model by giving the ROC charts of each cross-validation. In addition, Table 3 demonstrates the specific ACC, F1 Score and AUC of each cross-validation. All these performance metrics are close to 1, verifying that the constructed neural network is not overfitted, offering superior robustness and generalization.

Cross-Validations	ACC	F1 Score	AUC
CV1	1.000	1.000	1.000
CV2	1.000	1.000	1.000
CV3	0.960	0.962	0.987
CV4	0.960	0.923	0.945
CV5	0.958	0.960	0.990
CV6	1.000	1.000	1.000

Table 3. ACC, F1 Score and AUC of the 6-fold cross-validations.

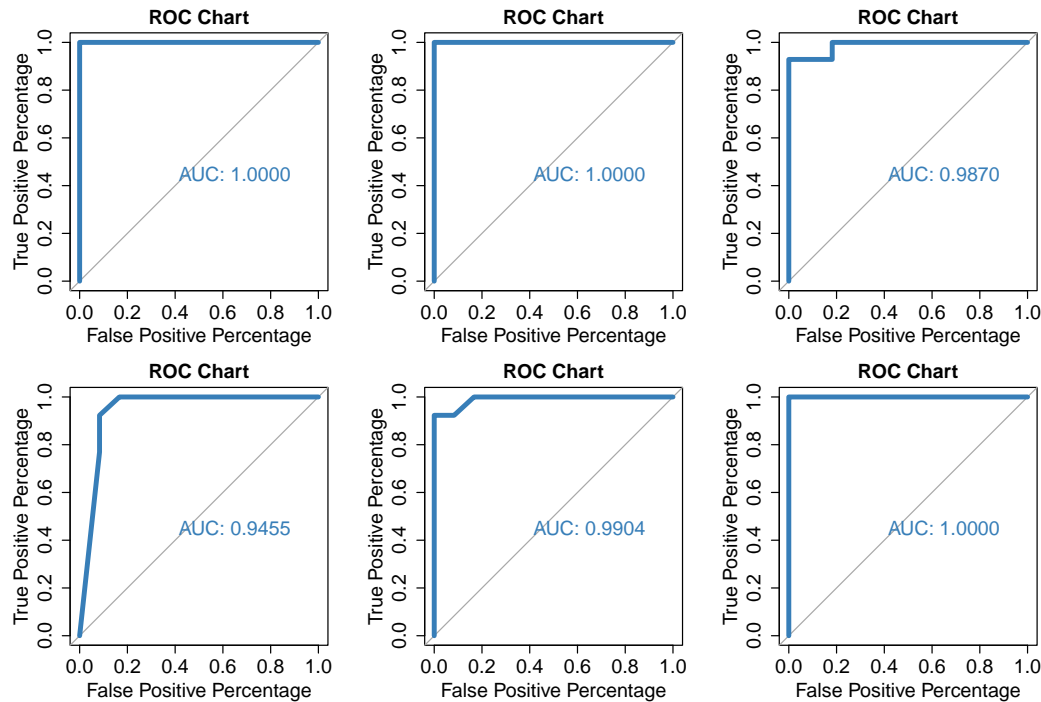


Figure 6. ROC curves of 6-fold cross-validations.

Performance Evaluation

In an available research, CXCL1 and GZMB are reported as critical biomarkers in prediction of the incidence of AD³⁶. As such, we conducted performance comparisons among the neural-networks-based diagnostic model, the CXCL1-based diagnostic model, and the GZMB-based diagnostic model. Diagnosis experiments were carried out on the training dataset and the independent test datasets (GSE5667 and GSE102628) respectively.

Figure 7 shows the comparison of the ROC charts of the three diagnosis baselines. Evaluations were conducted on the training dataset, the GSE5667, and the GSE102628, respectively. In addition, Table 4 demonstrates the specific ACC, F1 Score and AUC to verify the strength of the neural-networks-based model. It shows that the proposed neural-networks-based diagnosis draws its strength from excellent accuracy and higher robustness, compared with the single-biomarker-based diagnosis.

Diagnosis Experiments	ACC	F1 Score	AUC
neural-networks-based Diagnosis on the Training Dataset	0.973	0.972	0.980
CXCL1-Based Diagnosis on the Training Dataset	0.785	0.780	0.852
GZMB-Based Diagnosis on the Training Dataset	0.819	0.794	0.850
neural-networks-based Diagnosis on Test Dataset GSE5667	1.000	1.000	1.000
CXCL1-Based Diagnosis on Test Dataset GSE5667	0.878	0.833	0.833
GZMB-Based Diagnosis on Test Dataset GSE5667	0.878	0.833	0.900
neural-networks-based Diagnosis on Test Dataset GSE102628	0.842	0.823	0.85
CXCL1-Based Diagnosis on Test Dataset GSE102628	0.684	0.727	0.544
GZMB-Based Diagnosis on Test Dataset GSE102628	0.631	0.744	0.544

Table 4. Evaluation results with regard to the ACC, F1 Score and AUC.

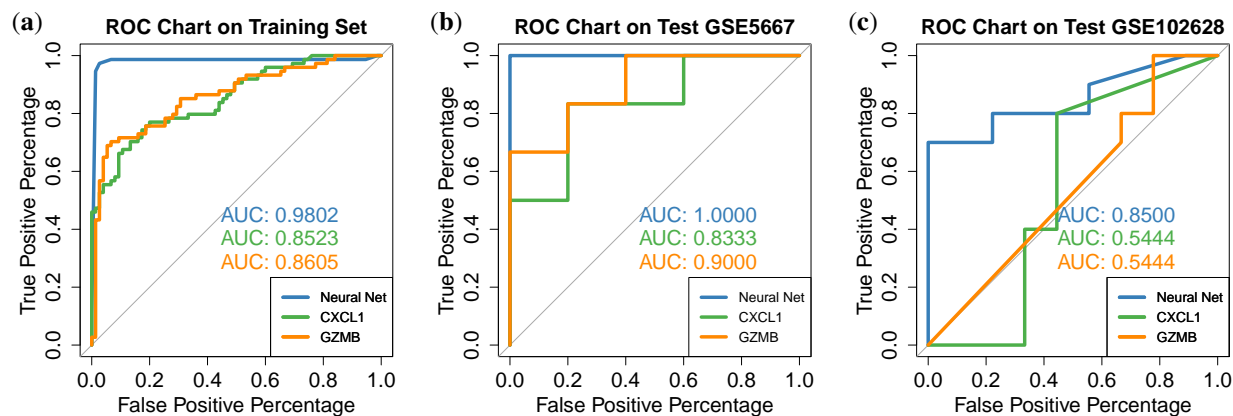


Figure 7. ROC curves on training dataset, test dataset GSE5667 and test dataset GSE102628.

Discussion

AD is a common but complex disorder characterized by increasingly recognized diverse clinical and molecular phenotypes which may be generated by gene-gene and gene-environment interactions³⁷. Due to this diversity, clinical observations may not entirely appreciate skin lesions, posing enormous diagnostic and therapeutic challenges. Thus, it is imperative to explore reliable and objective biomarkers for AD diagnosis. The purpose of this study was to predict phenotypes of AD with gene expression data. We constructed an AD diagnostic classifier solely based on gene expression data obtained from GEO database. We integrated and took advantage of random forest and artificial neural networks to infer important candidate classification genes and calculate the weights of these genes. Our diagnostic model with 6 selected critical genes automatically classified AD with excellent AUC efficiency (up to 1.00). It accurately distinguished subjects with AD from healthy controls. Among the critical genes, we found that PPP4R1 and GALNT6 had never been reported to be associated with AD. In summary, our classifier represents better prediction performance than utilizing previously reported marker genes and provides important biological insights into the development of the biomarkers of this disease.

In GO enrichment analysis based on the AD-related DEGs, the following aspects were significantly enriched: 1) inflammation-related processes: cytokine-mediated, granulocyte migration, neutrophil migration, etc.; 2) immune response processes: lymphocyte chemotaxis, regulation of innate immune response, T cell migration, etc.; 3) bacterial response processes: defense response to bacterium, antimicrobial humoral immune response mediated by antimicrobial peptide, response to molecule of bacterial origin, etc.; 4) skin-related processes: skin development, epidermal cell differentiation, epidermal cell differentiation, etc.; these results represented that the over activation of inflammatory and immune responses and the dysfunction of epidermis development processes occupied an essential position in the development of AD. In addition, several immune-related pathways were significantly enriched in the KEGG pathway, such as cytokine-cytokine receptor interaction, IL-17 signaling pathway, chemokine signaling pathway. As primary mediators of the adaptive immune system, cytokines including IL-4, IL-5, IL-13, IL-31 and IL-33 are important players in the Th2 immune response of AD³⁸. IL-17, a key cytokine for activating macrophages and fibroblasts, can induce antimicrobial peptide expression, which may contribute to bacterial infection frequently seen in AD³⁹. Interestingly, by executing the DO analysis, we observed that the DO analysis significantly enriched in not only atopic dermatitis but also artery diseases. This observation indicates that it may have some potential pathogenesis relationships between these two diseases. In practice, vascular inflammation in AD is linked to Th2 response and clinical severity that has been suggested previously⁴⁰. Among the six critical genes, three up-regulated genes (S100A7, S100A9, SERPINB4) are involved in multiple immune processes, while one down-regulated gene (BTC) plays a prominent role in positive regulation of epidermal growth. S100A7 and S100A9, members of the S100 family, can exhibit antimicrobial activities and induce immunomodulatory activities in hyperproliferative skin diseases⁴¹. SERPINB4 has crucial effects on initiating early inflammatory response and can lead to chronic skin inflammation like AD⁴². BTC is one of the epidermal growth factors, playing important roles in skin morphogenesis, homeostasis, and repair⁴³. It is worth noting that the function of PPP4R1 and GALNT6 on AD remains unclear, but these two genes are among the top 15 DEGs with minimum adjusted *P* value in AD (Supplementary Table S2). Therefore, these six genes can be used as promising biomarkers to establish a diagnostic model of AD.

To the best of our knowledge, this is the first attempt to apply random forest and neural network algorithms with gene expression profiles in AD diagnosis. Compared with the other two potential gene biomarkers reported previously for AD diagnosis³⁶, the proposed neural-networks-based diagnostic model in our study has drawn its strength from more excellent accuracy, as well as higher robustness in the training set (AUC=0.9802) and two validation sets (AUC=1.0000 in GSE5667

and AUC=0.8500 in GSE102628 respectively). Our diagnostic model covered patients with AD of different races, severities, and disease courses, greatly increasing the universality of the diagnosis. It should be noted that GSE102628 is a special dataset different from other collected datasets. Specifically, the transcriptomic data of the AD samples in GSE102628 were obtained from their non-lesional skin tissues. In such a case, we can witness an evident performance loss if CXCL1 or GZMB was adopted solely for the diagnosis of AD, i.e., the AUC of such a diagnostic model is close to 0.5. Nevertheless, the proposed neural-networks-based diagnostic model remains efficient on GSE102628, validating the generalization of the neural-networks-based model. Previously, several sets of diagnostic models of AD using omics data were established to explore the potential diagnostic biomarkers. Based on the transcriptomic data for host gene expression, 16S rRNA microbiome and metagenome shotgun data of the intestinal microbiota, Park, J et al. employed machine learning to construct a diagnostic model, the AUC of which was 0.715⁴⁴. In a similar manner, AUC of another machine learning diagnostic model based on the transcriptome of gut epithelial colonocytes and gut microbiota data was 0.75⁴⁵. Besides, a diagnostic model developed based on microbiome in serum extracellular vesicle analysis showed high accuracy (AUC=1.00), but without external validation. As we know, the accessibility of skin makes it the perfect tissue for the investigation of skin disease. With the emergence of minimally invasive methods, such as tape strips, access to skin sampling from both adults and children may become more convenient and easier in the near future⁴⁶. Meanwhile, it might be more cost-effective to obtain transcriptomic data with the rapid development of second-generation sequencing technologies. Thus, with a high degree of predictive accuracy, our classifier is promising to assist clinicians to perform accurate AD diagnosis and inform clinical decision-making in the future.

There are also limitations in our study. We elaborate some of them here, which may facilitate future works. First, the neural network training process in this paper was conducted on a dataset with a sample size of 149, which may not be sufficient for the data-driven learning method, thus might cause overfitting and biases when training the neural-networks-based model. To verify the robustness of our proposed model, we applied 6-fold cross-validations and two independent validation datasets, whose performance showed that we successfully controlled the biases and overfitting of our diagnostic model. Second, the number of the validation datasets in this paper was also limited. In this regard, it may be valuable to generalize our diagnostic model into more AD-diagnosis scenarios, so as to further test the effectiveness of the diagnostic model. Third, since the ages of our subjects range from 16 to 81, there is a possibility that the results may differ in children. Therefore, in order to validate and improve the neural-networks-based model and evaluate its performance more accurately, further studies in larger sample sizes and independent cohorts are needed. Although this is a compromising strategy in the situation of limited sample size, our model has shown an excellent classification performance. A diagnostic model for AD patients of all ages still needs to be constructed with more convincing datasets in the future.

In this study, we offered pieces of evidence that the transcriptome in skin tissue might be a critical factor in disease diagnosis. Through transcriptome utilization, objective and reliable diagnostic biomarkers could be attained. The 6 critical genes (PPP4R1, SERPINB4, S100A7, S100A9, BTC, and GALNT6) in our diagnostic model may be developed into valuable biomarkers of AD diagnosis and may provide invaluable clues or perspectives for further researches on the pathogenesis of AD.

Author contributions statement

W.Z., Y.L. and Y.C. conceived the study and designed the experiments, W.Z. and A.L. conducted the experiments and analyzed the results, W.Z., A.L., C.Z. and Z.L. drafted the manuscript. All authors reviewed the manuscript.

Funding

This study was supported by National Natural Science Foundation of China (No.31972856), Guangdong Provincial Key Laboratory of Chinese Medicine for Prevention and Treatment of Refractory Chronic Diseases (No. 2018B030322012) and Top Talents Project of Guangdong Provincial Hospital of Traditional Chinese Medicine(No.BJ2022YL08).

Data availability

The data supporting the findings of this study are included within the article.

Competing interests

The authors declare no competing interests.

References

1. McKenzie, C. & Silverberg, J. I. The prevalence and persistence of atopic dermatitis in urban united states children. *Annals Allergy, Asthma & Immunol.* **123**, 173–178 (2019).

2. Fuxench, Z. C. C. *et al.* Atopic dermatitis in america study: a cross-sectional study examining the prevalence and disease burden of atopic dermatitis in the us adult population. *J. Investig. Dermatol.* **139**, 583–590 (2019).
3. Laughter, M. *et al.* The global burden of atopic dermatitis: lessons from the global burden of disease study 1990–2017. *Br. J. Dermatol.* **184**, 304–309 (2021).
4. Kage, P., Simon, J.-C. & Treudler, R. Atopic dermatitis and psychosocial comorbidities. *JDDG: J. der Deutschen Dermatol. Gesellschaft* **18**, 93–102 (2020).
5. Langan, S. M., Irvine, A. D. & Weidinger, S. Atopic dermatitis. *Lancet (London, England)* **396**, 345–360, DOI: [10.1016/S0140-6736\(20\)31286-1](https://doi.org/10.1016/S0140-6736(20)31286-1) (2020). [Online; accessed 2022-03-23].
6. Bakker, D. S. *et al.* Confirmation of multiple endotypes in atopic dermatitis based on serum biomarkers. *The J. allergy clinical immunology* **147**, 189–198, DOI: [10.1016/j.jaci.2020.04.062](https://doi.org/10.1016/j.jaci.2020.04.062) (2021). [Online; accessed 2022-03-23].
7. Patrick, G. J., Archer, N. K. & Miller, L. S. Which way do we go? complex interactions in atopic dermatitis pathogenesis. *J. Investig. Dermatol.* **141**, 274–284 (2021).
8. Bieber, T. Atopic dermatitis: an expanding therapeutic pipeline for a complex disease. *Nat. reviews. Drug discovery* **21**, 21–40, DOI: [10.1038/s41573-021-00266-6](https://doi.org/10.1038/s41573-021-00266-6) (2022). [Online; accessed 2022-03-23].
9. Liu, P. *et al.* Clinical features of adult/adolescent atopic dermatitis and chinese criteria for atopic dermatitis. *Chin. medical journal* **129**, 757–62, DOI: [10.4103/0366-6999.178960](https://doi.org/10.4103/0366-6999.178960) (2016). [Online; accessed 2022-03-23].
10. Dizon, M. P. *et al.* Systematic review of atopic dermatitis disease definition in studies using routinely collected health data. *The Br. journal dermatology* **178**, 1280–1287, DOI: [10.1111/bjd.16340](https://doi.org/10.1111/bjd.16340) (2018). [Online; accessed 2022-03-23].
11. HanifIn, J. M. Diagnostic features of atopic dermatitis. *Acta Derm. Venereol.* **92**, 44–47 (1980).
12. Williams, H. *et al.* The uk working party's diagnostic criteria for atopic dermatitis. i. derivation of a minimum set of discriminators for atopic dermatitis. *Br. journal dermatology* **131**, 383–396 (1994).
13. Williams, H., Jburney, P., Strachan, D., Hay, R. & Party, A. D. D. C. W. The uk working party's diagnostic criteria for atopic dermatitis ii. observer variation of clinical diagnosis and signs of atopic dermatitis. *Br. journal dermatology* **131**, 397–405 (1994).
14. Williams, H., Jburney, P., Pembroke, A., Hay, R. & Party, A. D. D. C. W. The uk working party's diagnostic criteria for atopic dermatitis. iii. independent hospital validation. *Br. journal dermatology* **131**, 406–416 (1994).
15. Saeki, H. *et al.* Guidelines for management of atopic dermatitis. *The J. dermatology* **36**, 563–77, DOI: [10.1111/j.1346-8138.2009.00706.x](https://doi.org/10.1111/j.1346-8138.2009.00706.x) (2009). [Online; accessed 2022-03-23].
16. Vakharia, P. P., Chopra, R. & Silverberg, J. I. Systematic review of diagnostic criteria used in atopic dermatitis randomized controlled trials. *Am. journal clinical dermatology* **19**, 15–22, DOI: [10.1007/s40257-017-0299-4](https://doi.org/10.1007/s40257-017-0299-4) (2018). [Online; accessed 2022-03-23].
17. Silvestre Salvador, J. F., Romero-Pérez, D. & Encabo-Durán, B. Atopic dermatitis in adults: A diagnostic challenge. *J. investigational allergology & clinical immunology* **27**, 78–88, DOI: [10.18176/jiaci.0138](https://doi.org/10.18176/jiaci.0138) (2017). [Online; accessed 2022-03-23].
18. Thijs, J. L., de Bruin-Weller, M. S. & Hijnen, D. Current and future biomarkers in atopic dermatitis. *Immunol. allergy clinics North Am.* **37**, 51–61, DOI: [10.1016/j.iac.2016.08.008](https://doi.org/10.1016/j.iac.2016.08.008) (2017). [Online; accessed 2022-03-23].
19. Puzzovio, P. G. & Levi-Schaffer, F. Latest progresses in allergic diseases biomarkers: Asthma and atopic dermatitis. *Front. pharmacology* **12**, 747364, DOI: [10.3389/fphar.2021.747364](https://doi.org/10.3389/fphar.2021.747364) (2021). [Online; accessed 2022-03-23].
20. Zhong, Y. *et al.* Identification of immunological biomarkers of atopic dermatitis by integrated analysis to determine molecular targets for diagnosis and therapy. *Int. journal general medicine* **14**, 8193–8209, DOI: [10.2147/IJGM.S331119](https://doi.org/10.2147/IJGM.S331119) (2021). [Online; accessed 2022-03-23].
21. Tian, Y., Yang, J., Lan, M. & Zou, T. Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure. *Aging (Albany NY)* **12**, 26221 (2020).
22. Li, H., Lai, L. & Shen, J. Development of a susceptibility gene based novel predictive model for the diagnosis of ulcerative colitis using random forest and artificial neural network. *Aging (Albany NY)* **12**, 20471 (2020).
23. Xie, N.-N., Wang, F.-F., Zhou, J., Liu, C. & Qu, F. Establishment and analysis of a combined diagnostic model of polycystic ovary syndrome with random forest and artificial neural network. *BioMed Res. Int.* **2020** (2020).
24. Leek, J. T. *et al.* sva: Surrogate variable analysis. *R package version* **3**, 882–883 (2019).

25. Ritchie, M. E. *et al.* limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research* **43**, e47–e47 (2015).
26. Kolde, R. pheatmap: Pretty heatmaps. r package version 1.0. 12. *R Packag. version 1.0* **8** (2019).
27. Villanueva, R. A. M. & Chen, Z. J. ggplot2: elegant graphics for data analysis (2019).
28. Wu, T. *et al.* clusterprofiler 4.0: A universal enrichment tool for interpreting omics data. *The Innov.* **2**, 100141 (2021).
29. Yu, G., Wang, L.-G., Yan, G.-R. & He, Q.-Y. Dose: an r/bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* **31**, 608–609 (2015).
30. Walter, W., Sánchez-Cabo, F. & Ricote, M. Goplot: an r package for visually combining expression data with functional analysis. *Bioinformatics* **31**, 2912–2914 (2015).
31. Han, H., Guo, X. & Yu, H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 219–224, DOI: [10.1109/ICSESS.2016.7883053](https://doi.org/10.1109/ICSESS.2016.7883053) (2016).
32. Jiang, Z. *et al.* Accurate diagnosis of atopic dermatitis by combining transcriptome and microbiota data with supervised machine learning. *Sci. reports* **12**, 1–13 (2022).
33. Cutler, A., Cutler, D. R. & Stevens, J. R. Random forests. In *Ensemble machine learning*, 157–175 (Springer, 2012).
34. Günther, F. & Fritsch, S. Neuralnet: training of neural networks. *R J.* **2**, 30 (2010).
35. Jia, Z. Controlling the overfitting of heritability in genomic selection through cross validation. *Sci. reports* **7**, 1–9 (2017).
36. Zhong, Y. *et al.* Identification of immunological biomarkers of atopic dermatitis by integrated analysis to determine molecular targets for diagnosis and therapy. *Int. J. Gen. Medicine* **14**, 8193 (2021).
37. Bieber, T. *et al.* Unraveling the complexity of atopic dermatitis: The ck-care approach towards precision medicine. *Allergy* **75**, 2936–2938 (2020).
38. Gavrilova, T. Immune dysregulation in the pathogenesis of atopic dermatitis. *Dermatitis* **29**, 57–62 (2018).
39. Sugaya, M. The role of th17-related cytokines in atopic dermatitis. *Int. journal molecular sciences* **21**, 1314 (2020).
40. Villani, A. P. *et al.* Vascular inflammation in moderate-to-severe atopic dermatitis is associated with enhanced th2 response. *Allergy* **76**, 3107–3121, DOI: [10.1111/all.14859](https://doi.org/10.1111/all.14859) (2021). [Online; accessed 2022-03-24].
41. Hattori, F. *et al.* The antimicrobial protein s100a7/psoriasin enhances the expression of keratinocyte differentiation markers and strengthens the skin's tight junction barrier. *Br. J. Dermatol.* **171**, 742–753 (2014).
42. Sivaprasad, U. *et al.* Serpinb3/b4 contributes to early inflammation and barrier dysfunction in an experimental murine model of atopic dermatitis. *J. Investig. Dermatol.* **135**, 160–169 (2015).
43. Shirakata, Y., Tokumaru, S., Sayama, K. & Hashimoto, K. Auto-and cross-induction by betacellulin in epidermal keratinocytes. *J. dermatological science* **58**, 162–164 (2010).
44. Park, J. *et al.* Multi-omics analyses implicate ears2 in the pathogenesis of atopic dermatitis. *Allergy* **76**, 2602–2604 (2021).
45. Jiang, Z. *et al.* Accurate diagnosis of atopic dermatitis by combining transcriptome and microbiota data with supervised machine learning. *Sci. reports* **12**, 1–13 (2022).
46. Hulshof, L. *et al.* A minimally invasive tool to study immune response and skin barrier in children with atopic dermatitis. *Br. journal dermatology* **180**, 621–630 (2019).