Hindawi
BioMed Research International
Volume 2020, Article ID 2613091, 13 pages
https://doi.org/10.1155/2020/2613091

Research Article

Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network

Ning-Ning Xie,[1] **Fang-Fang Wang**,[1] **Jue Zhou**,[2] **Chang Liu**,[3] **and Fan Qu**[1]

[1]*Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310006, China*
[2]*College of Food Science and Biotechnology, Zhejiang Gongshang University, Hangzhou 310018, China*
[3]*Zhejiang Chinese Medical University, Hangzhou 310053, China*

Correspondence should be addressed to Fan Qu; syqufan@zju.edu.cn

Received 30 April 2020; Revised 27 July 2020; Accepted 3 August 2020; Published 20 August 2020

Academic Editor: Stefano Pascarella

Polycystic ovary syndrome (PCOS) is one of the most common metabolic and reproductive endocrinopathies. However, few studies have tried to develop a diagnostic model based on gene biomarkers. In this study, we applied a computational method by combining two machine learning algorithms, including random forest (RF) and artificial neural network (ANN), to identify gene biomarkers and construct diagnostic model. We collected gene expression data from Gene Expression Omnibus (GEO) database containing 76 PCOS samples and 57 normal samples; five datasets were utilized, including one dataset for screening differentially expressed genes (DEGs), two training datasets, and two validation datasets. Firstly, based on RF, 12 key genes in 264 DEGs were identified to be vital for classification of PCOS and normal samples. Moreover, the weights of these key genes were calculated using ANN with microarray and RNA-seq training dataset, respectively. Furthermore, the diagnostic models for two types of datasets were developed and named neuralPCOS. Finally, two validation datasets were used to test and compare the performance of neuralPCOS with other two set of marker genes by area under curve (AUC). Our model achieved an AUC of 0.7273 in microarray dataset, and 0.6488 in RNA-seq dataset. To conclude, we uncovered gene biomarkers and developed a novel diagnostic model of PCOS, which would be helpful for diagnosis.

1. Introduction

Polycystic ovary syndrome (PCOS), as a heterogeneous endocrine disorder, is closely associated with menstrual dysfunction, infertility, hirsutism, acne, obesity, and metabolic syndrome [1]. The three major diagnostic criteria of PCOS widely followed are criteria raised by National Institutes of Health (NIH) [2], 2003 Rotterdam Consensus raised by European Society of Human Reproduction and Embryology (ESHRE) and American Society for Reproductive Medicine (ASRM) [3, 4], and criteria raised by Androgen Excess Society (AES) [5]. However, these criteria have created some controversy in the field [6]. The multifactorial etiology of PCOS is underpinned by a complex genetic architecture [7]. Ethnicity is eminently related to PCOS phenotype because of the different genetic and environmental propensity to metabolic disorders [8–10].

Although the identified genetic risk markers can be used as predictive and diagnostic tools for PCOS, they may not possess the strong power due to the complicated genetic architecture [6]. Combination of various markers in diagnostic panels may significantly improve the success [11]. Many studies have successfully used genetic risk scores to explain increasing amounts of variance in diseases [12].

In recent years, the wide application of microarray technology and more advanced, accurate RNA-sequencing technology made the study of disease mechanism more convenient. In view of the differences between the two platforms, it is necessary to analyze the data of the two platforms separately.

The main difficulty arisen in establishing a classification model using gene expression data was how to find the most meaningful index or feature for classification. To address this, various machine learning approaches such as random forest (RF) [13, 14] and artificial neural network (ANN) [15] were utilized. The single or combined use of these algorithms has contributed much in gene expression data classification [16], disease diagnosis [17], cell migration [18], and microbiome research [19]. Given their high classification accuracy and convenience, they have become powerful tools to learn feature representations.

In this work, we established a diagnosis model of PCOS using microarray and RNA-seq data from Gene Expression Omnibus (GEO) database with the combined utilization of RF and ANN. Firstly, the RF classifier was used to identify the key genes for classification, and then, the ANN was performed to calculate the weights of the key genes in microarray and RNA-seq data, respectively. Finally, a scoring model named neuralPCOS was developed with the integration of RF and ANN. To validate the accuracy and superiority of the diagnosis model we established, we evaluated the performance with microarray and RNA-seq data and compared them to other marker genes obtained in previous studies [20, 21].

2. Materials and Methods

2.1. Study Design. For establishment of the diagnostic model of PCOS, RF and ANN were adopted in this study. The study overview was schematically depicted in Figure 1. GSE6798 dataset ($n = 29$) was used for the differentially expressed genes (DEGs) screening (step 1). Gene ontology (GO) enrichment analysis (step 2) and the acquisition of key genes for classification by RF (step 3) were further performed. After computing the gene weight using ANN in two kinds of expression data (microarray and RNA-seq) (step 4), a classification model was developed (step 5). Finally, we used two independent dataset (the microarray ComBat dataset2 and the RNA-seq–based dataset GSE84958) for further validation (step 6).

2.2. Data Selection and Preprocessing. In the present study, a wide search through the National Center for Biotechnology Information Gene Expression Omnibus database (NCBI-GEO) platform was conducted with the key words "PCOS, human". As shown in Table 1, 6 sets of microarray data and 1 set of RNA-seq data were downloaded from GEO database. In order to obtain one training dataset (microarray ComBat dataset1) with large sample size, three microarray datasets with small sample size (GSE137684, GSE137354, and GSE34526) were combined. Meanwhile, GSE43264 and GSE124226 were combined to form one validation dataset (microarray ComBat dataset2). These datasets were converted to logarithmic form after standardization, and the R package ComBat was used to remove the batch effects [22]. Two microarray datasets with 28 and 23 samples were obtained using classical and Bayesian correction methods.

2.3. Differentially Expressed Genes (DEGs) Screening. The dataset GSE6798, based on Affymetrix Human Genome U133 Plus 2.0 Array (Affymetrix Inc., Santa Clara, California, USA) contained 16 cases of PCOS and 13 cases of control, was used for DEGs analysis. The boxplot was performed using R package stats (v 3.5.0). The R package limma was used to calculate the DEGs between the PCOS and control samples by the classical Bayesian method with $P < 0.01$ and |logFoldchange| >0.26 [23] and was visualized by volcano plot [24].

2.4. Gene Ontology (GO) Enrichment Analysis. To further reveal the biofunction of selected DEGs, GO enrichment analysis, including biological process (BP), cellular component (CC), and molecular function (MF), was performed using R package clusterProfiler [25]. Significant enrichment terms were screened with the threshold adjusted $P < 0.01$ after adjusted by the Benjamini and Hochberg method. To eliminate some redundancies, GO terms that intersects more than 75% of the genes contained in term were removed. GObubble and GOChord were performed with R package GOplot to illustrate the functional analysis data [26].

2.5. Random Forest (RF) Classification. We used random forest to classify the DEGs with the R package randomforest [27]. Firstly, the optimal number of variables (mtry parameter, the optimal number of variables used in the binary tree in the specified node) was identified. All possible variables (1~2000) were looped into the random forest classifier. Each error rate was calculated, and the optimal number of variables was selected. Next, each error rate of 1~3000 trees was calculated, and the optimal tree number was determined by the lowest error rate and best stability. Based on the above-selected parameters, the random forest classifier was used to calculate the results, and the important genes were selected as the candidate PCOS-specific genes according to the Gini coefficient method.

2.6. Calculation of DEGs Weight by Artificial Neural Network (ANN). The GSE84958 dataset was randomly divided into training data ($n = 26$) and validation data ($n = 27$). The RNA-seq training data GSE84958 ($n = 26$) and microarray ComBat dataset1 ($n = 28$) were used to construct the neural network model. The R package neuralnet was used for neural network analysis [28]. First of all, the integration data were filtered and normalized by min-max normalization. Secondly, the processed training data was inputted into the neural network model. Eleven genes were inputted and 3 hidden layers, and 2 outputs (normal and PCOS) were set in both microarray data and RNA-seq data. Finally, the output of the first hidden layer (input of the last output layer) in the network results were considered as the results of gene weight.

2.7. Neural-PCOS. We constructed an equation named neuralPCOS that could estimate the classification score of each gene in microarray data or RNA-seq data.

$$\text{neuralPCOS} = \sum \left(GeneExpression \times NeuralNetworkWeight \right). \tag{1}$$

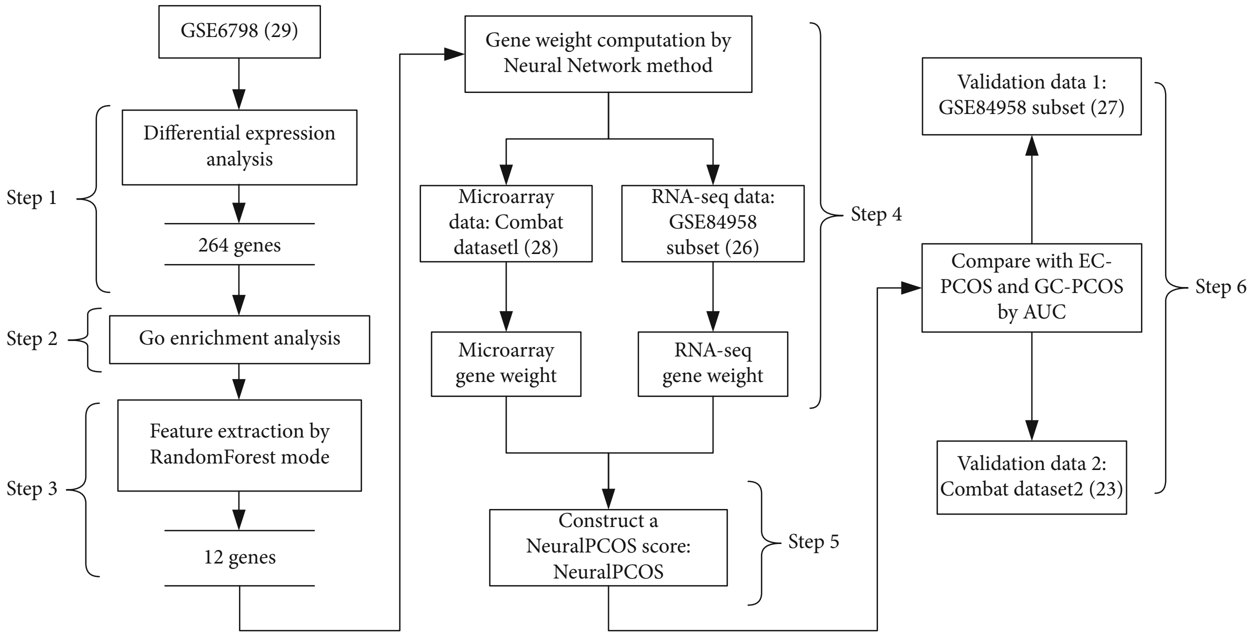

GSE6798 (29)
Differential expression analysis
264 genes
Go enrichment analysis
Feature extraction by RandomForest mode
12 genes
Step 1
Step 2
Step 3
Gene weight computation by Neural Network method
Microarray data: Combat datasetl (28)
RNA-seq data: GSE84958 subset (26)
Microarray gene weight
RNA-seq gene weight
Step 4
Construct a NeuralPCOS score: NeuralPCOS
Step 5
Validation data 1: GSE84958 subset (27)
Compare with EC-PCOS and GC-PCOS by AUC
Validation data 2: Combat dataset2 (23)
Step 6

Figure 1: Schematic illustration of study design. A total of 264 differentially expressed genes (DEGs) were obtained in differential expression analysis with GSE6798 dataset (skeletal muscle, $n = 29$) (step 1), and functional enrichment analysis were also performed (step 2). All the 264 DEGs were tested for their potential as classification-related genes with random forest model, and 12 key genes were identified (step 3). Artificial neural network (ANN), another machine learning algorithm, was used to calculate the weight of genes (step 4). Therefore, a versatile classification model, designated as neuralPCOS, was established with the use of RF and ANN (step 5). Finally, the utility of neuralPCOS was validated in microarray data and RNA-seq data (step 6).

Table 1: Gene expression data from Gene Expression Omnibus (GEO) database.

Dataset ID	Total samples	Control	PCOS	Data type	Tissue type	Country
GSE6798	29	13	16	Microarray	Skeletal muscle	Denmark
GSE43264	15	7	8	Microarray	Adipose	Ireland
GSE34526	10	3	7	Microarray	Granulosa cells	India
GSE137684	12	4	8	Microarray	Granulosa cells	China
GSE137354	6	3	3	Microarray	Endometrium	China
GSE124226	8	4	4	Microarray	Adipose stem cells	USA
GSE84958	53	23	30	RNA-seq	Adipose	UK

The gene expression value was multiplied by the weight of gene, and the results of all genes were added. (Note: before calculating the score, the expression data after log2 processing needs to be normalized by min-max normalization.)

2.8. Evaluation of Performance by Area under Curve (AUC). The AUCs of three kinds of scores (neuralPCOS, EC-PCOS, GC-PCOS) were calculated in GSE84958 RNA-seq validation data ($n = 27$) and microarray validation data ($n = 23$) with R package pROC, respectively [29].

Three kinds of score:

(1) neuralPCOS

(2) EC-PCOS: three upregulated genes including insulin-like growth factor 1 (*IGF1*), phosphatase and tensin homolog (*PTEN*), and insulin-like growth factor-binding protein 1 (*IGFBP1*) in endometrial cells (ECs) of PCOS [20].

(3) GC-PCOS: upegulated genes including hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2 (*HSD3B2*), steroidogenic acute regulatory protein (*STAR*), inhibin subunit beta A (*INHBA*), and cytochrome P450 family 19 subfamily A member 1 (*CYP19A1*) in granulosa cells (GCs) of PCOS [21].

3. Results

3.1. Identification of DEGs. Firstly, the boxplot presented RNA expression level in GSE6798 ($n = 29$) (Figure S1). A total of 20174 gene symbols were identified after annotation, and the distribution of DEGs ($P < 0.01$, |logFC| >0.26) were

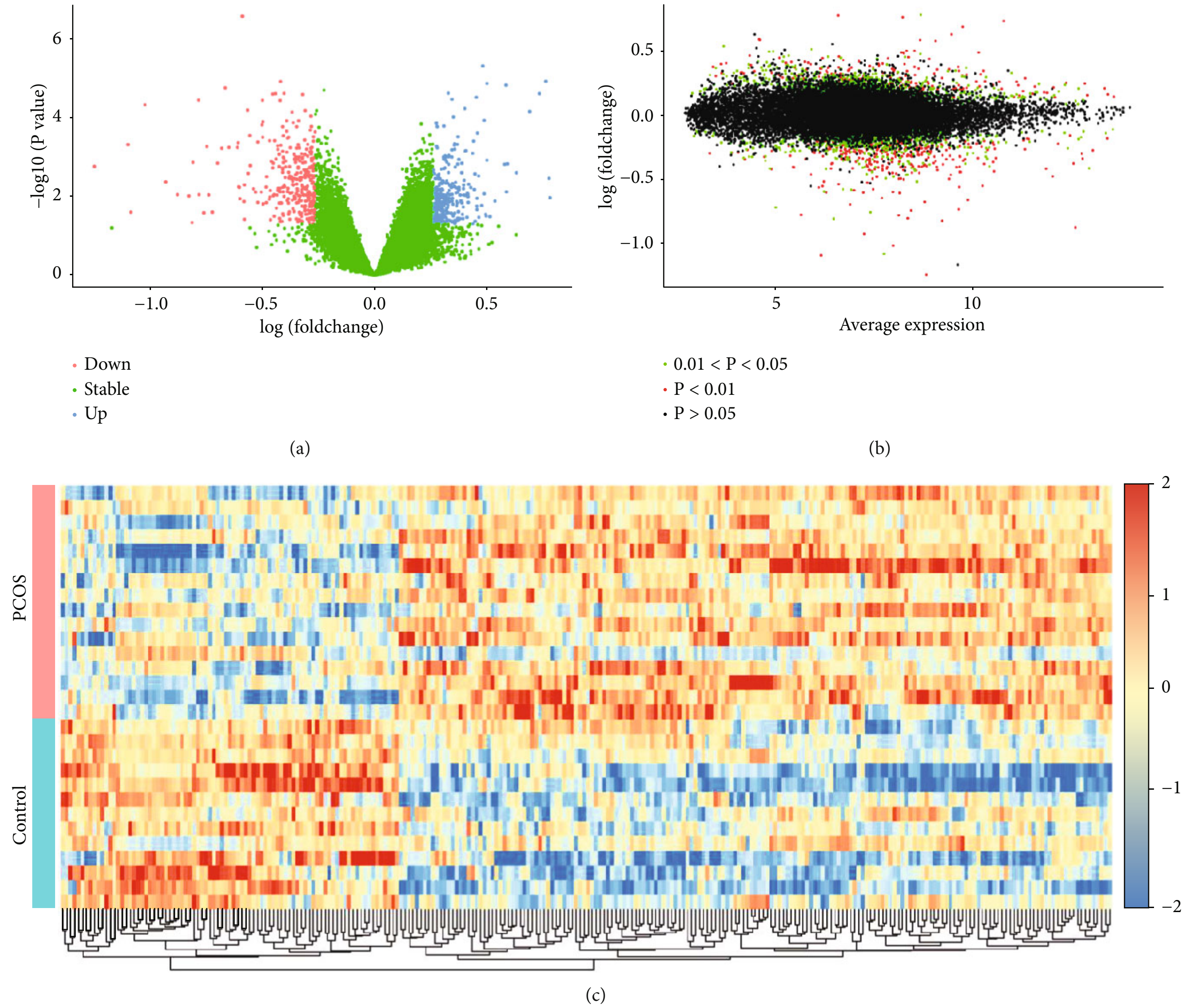
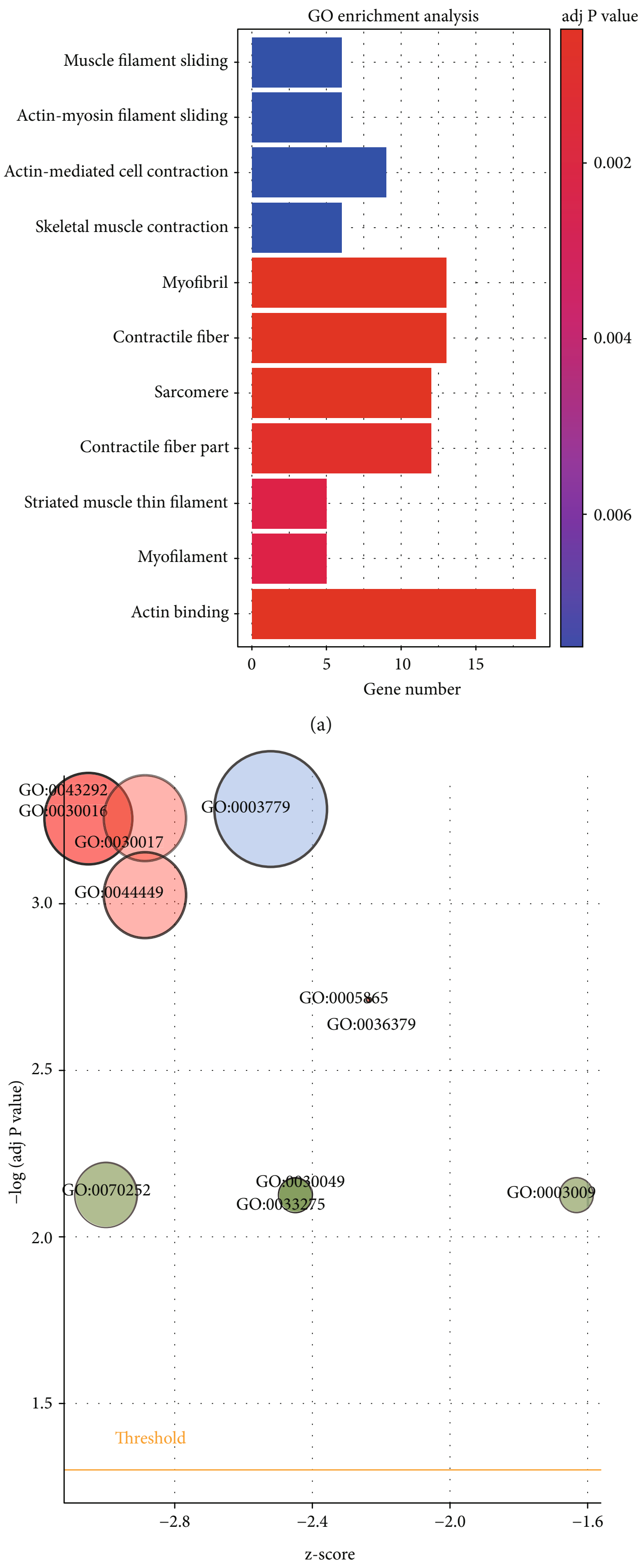


Figure 2: Analyses of DEGs in GSE6798 dataset (skeletal muscle, $n = 29$). (a) Volcano plot of differential gene expression with log (foldchange) as the abscissa and -log10 (P value) as the ordinate. Blue and red splashes represent the genes that were significantly up- or downregulated in PCOS, respectively. Green splashes mean genes without significantly different expression. $P < 0.01$, $|\text{logFC}| > 0.26$. (b) Volcano plot of gene average expression level. The x-axis represents the average expression levels of genes in all samples. The y-axis indicates logFC. The red spots are DEGs with $P < 0.01$, the green spots, DEGs with $0.01 < P < 0.05$; and the black spots, stable genes ($P > 0.05$). (c) Heatmap of the 264 DEGs in GSE6798 dataset. Each row represents a sample and each column represents a gene. Red color means a higher expression level; blue color means a lower expression level.

represented by volcano plot, including 134 upregulated genes and 210 downregulated genes (Figure 2(a)). The volcano plot of gene average expression level was shown in Figure 2(b). Moreover, the genes with low expression level (last 25%) were removed, and 264 genes ($P < 0.01$, $|\text{logFC}| > 0.26$) were obtained (Table S1). The heat map of the screened 264 DEGs in GSE6798 dataset was shown in Figure 2(c).

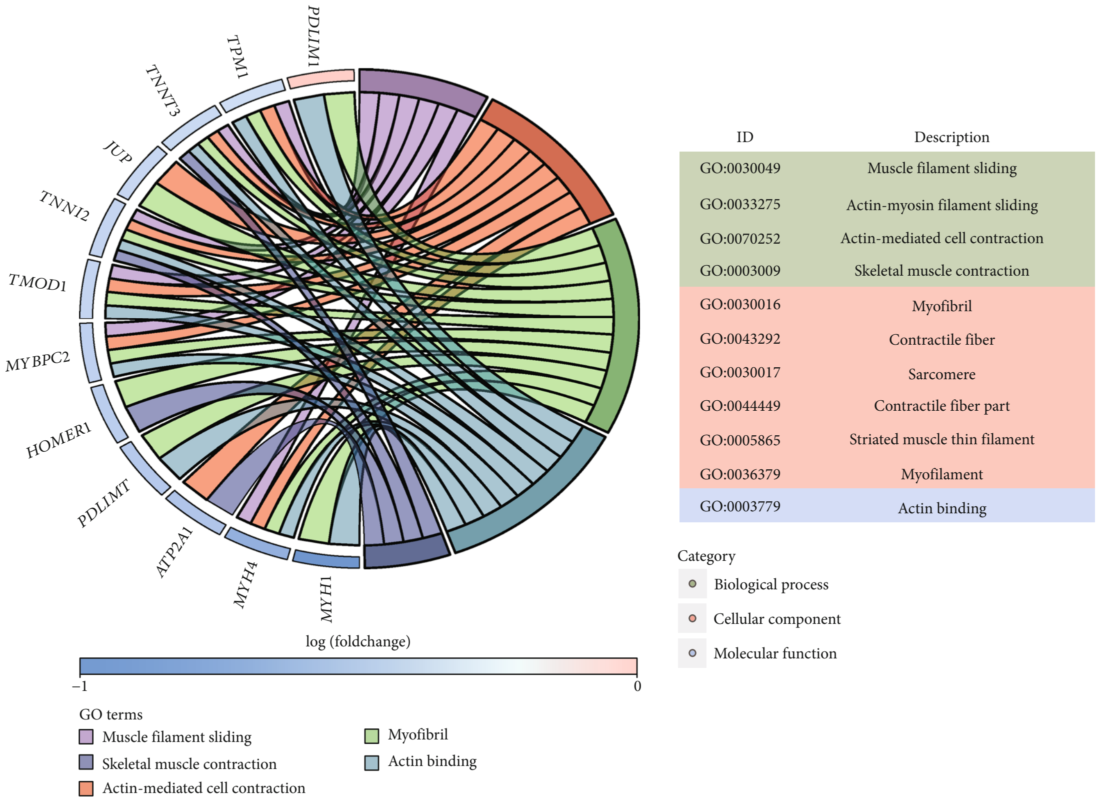
3.2. Functional Characterization of Selected DEGs. GO enrichment analysis for the selected 264 DEGs was carried out to identify the significantly enriched GO terms. The GObar showed the predominant significantly enriched GO terms (adjusted $P < 0.01$) (Figure 3(a)). Muscle filament sliding (adjusted $P = 7.49E - 03$), myofibril (adjusted $P = 5.55E - 04$), and actin binding (adjusted $P = 5.19E - 04$) were the most significantly enriched GO terms in BP, CC, and MF, respectively (Table S2). The 11 enriched terms were displayed in bubble plot (Figure 3(b)). The analysis revealed that skeletal muscle contraction was the most upregulated term; contractile fiber was the most downregulated one. After de-redundant the resulting GO terms, 5 enriched terms were obtained. To add quantitative molecular data in the GO terms of interest, GOChord was performed. It indicated that 12 DEGs were enriched into 5 Go terms, among which myofibril contained the most DEGs (Figure 3(c)).



(a)

(b)

Figure 3: Continued.



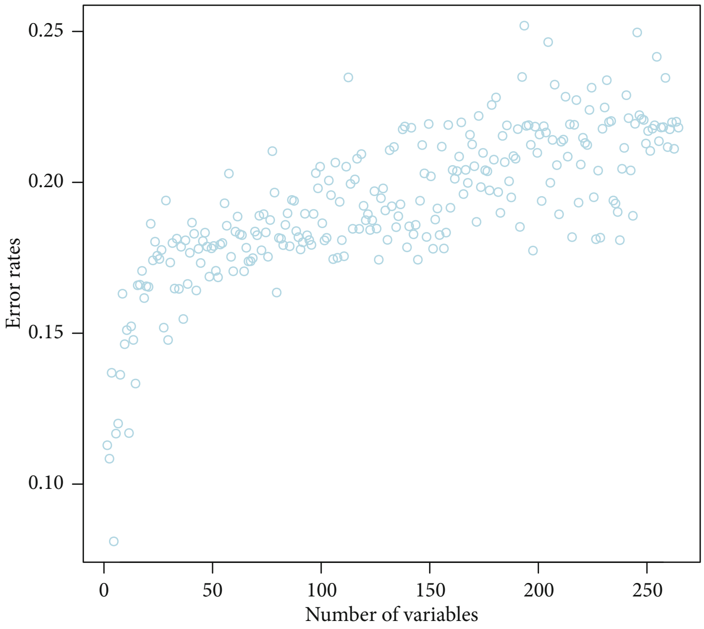
(c)

Figure 3: Gene ontology (GO) enrichment analysis. (a) The bar plot of enriched GO terms in biological process (BP), cell components (CC), and molecular function (MF). The x-axis represents the GOid, and y-axis represents the significance of terms. The terms are placed according to their z-score (which indicates that the term is more likely to increase or decrease). (b) The bubble plot of GO analysis of 264 DEGs. The z-score is assigned to x-axis and the negative logarithm of the adjusted P value to y-axis. Bubble size is proportional to the number of genes in GO terms, and the color represents three categories (green: BP; red: CC; blue: MF). (c) GOChord plot: a plot indicates the relationship between DEGs and their associated terms. The color represents upregulation (red) or downregulation (blue).

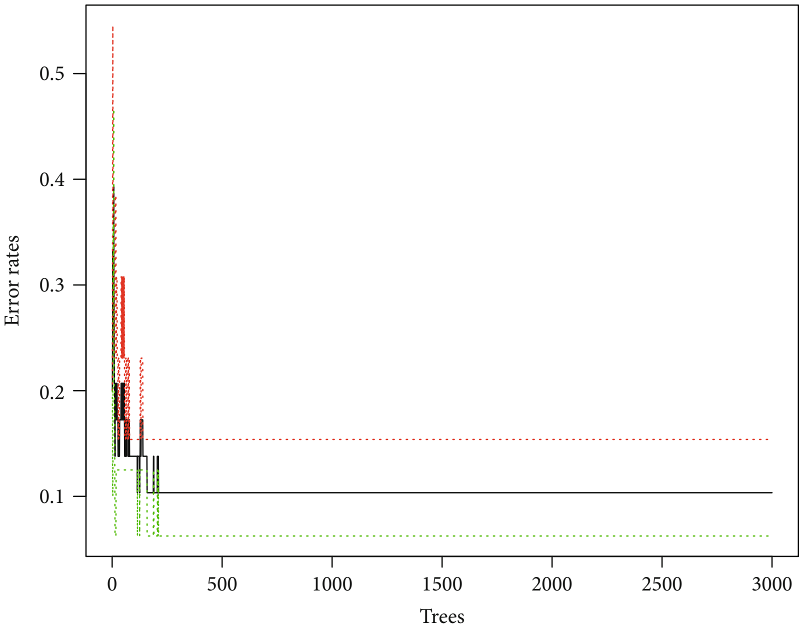
3.3. Screening Candidate PCOS-Specific Genes by Random Forest. In order to obtain more reliable PCOS-specific genes, we inputted the above 264 DEGs into the RF classifier. The lowest error rate occurred when the number of variables was 4 (Figure 4(a)); meanwhile, the optimal number of trees in RF classifier was set to 1000 due to the low error rate and stability (Figure 4(b)). Therefore, we finally choose 4 and 1000 trees as the final parameter in RF classifier to obtain the dimensional importance of all variables. Top 12 genes in the results of MeanDecreaseAccuracy and MeanDecrease-Gini were shown in Figure 4(c). Finally, we selected 0.15 as the screening threshold of importance in MeanDecreaseGini result, and a set of 12 PCOS-specific DEGs was identified.

3.4. ANN Analysis. RF classifier identified the key genes, which optimally differentiated between PCOS and controls. To further construct a PCOS-specific scoring model, ANN analysis was performed to calculate the weight of 12 genes. Here, two parallel training processes were carried out according to format of the training data, including RNA-seq training data GSE84958 ($n = 26$) and microarray ComBat dataset1 ($n = 28$). ANN topology of microarray ComBat dataset1 and RNA-seq data indicated 11 input layer, 3 hidden layer, and 2 output layer (Figure 5). The weight of each gene was detailed in Table S3 for microarray ComBat dataset1 and Table S4 for RNA-seq data. Based above, we constructed a model for classifying the gene expression data between PCOS and control samples.

3.5. The Validation of neuralPCOS. Microarray ComBat dataset2 ($n = 23$) and GSE84958 RNA-seq verification data ($n = 27$) were used to test the ability for classifying the samples in 3 classification models, including neuralPCOS constructed in this study and EC-PCOS and GC-PCOS from other researches. The performance of these models was examined using area under the receiver operating characteristic curve (ROC) (Figure 6). First, we estimated differences in the AUC values among 3 models in microarray data (Figure 6(a)). The results showed that neuralPCOS had a high-level classification performance with an AUC of 0.7273, compared with the AUC of EC-PCOS (0.5985) and GC-PCOS (0.5227). The optimal threshold values for 3 models were 1.2, 0.4, and 0.3, respectively. NeuralPCOS and EC-PCOS achieved the highest level of specificity (75.0%), and GC-PCOS had 100% sensitivity at optimal threshold value. The result of RNA-seq validation data suggested that the AUC score of neuralPCOS (0.6488) was higher than EC-PCOS (0.5770), but lower than GC-PCOS

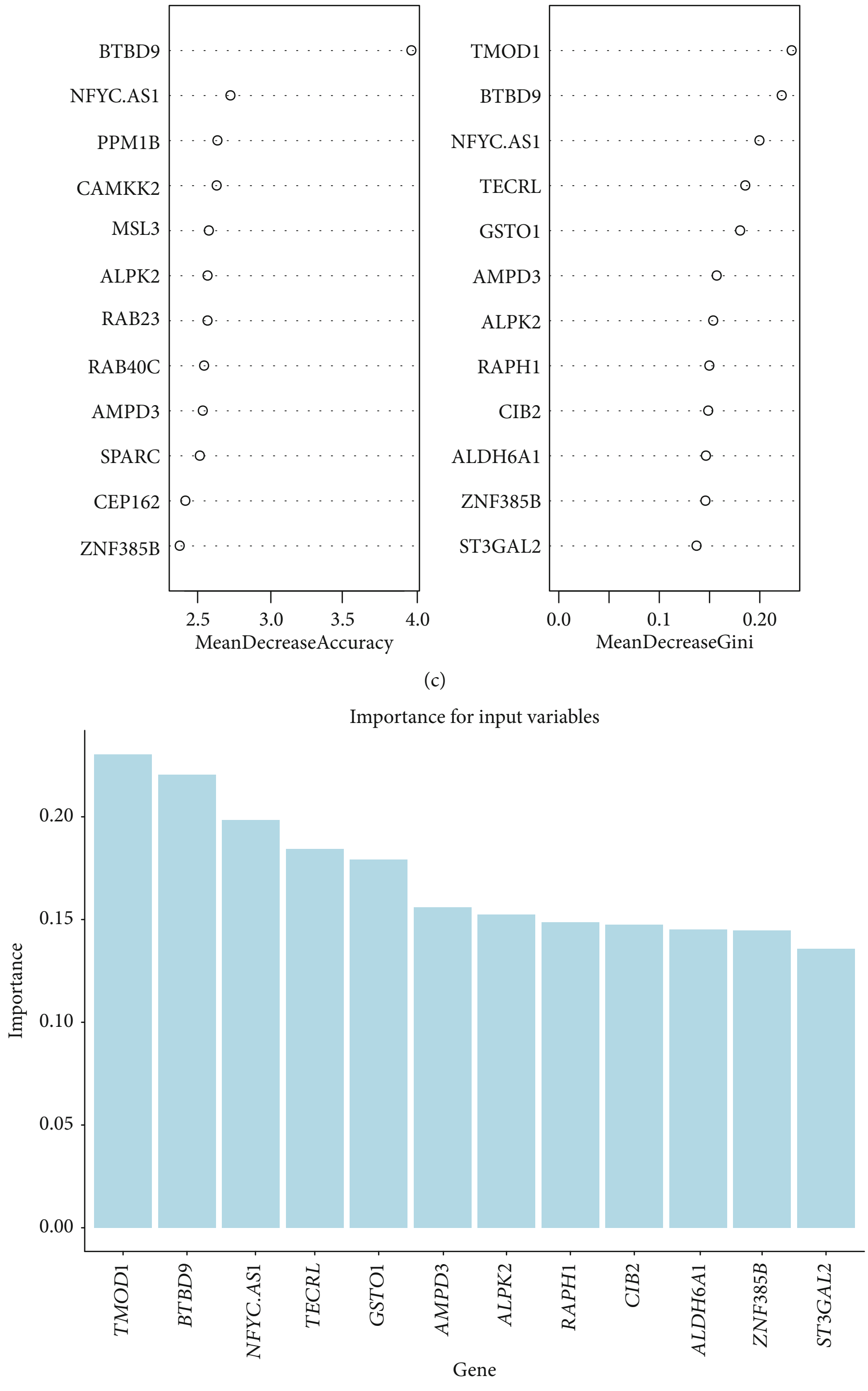


(a)



(b)

FIGURE 4: Continued.

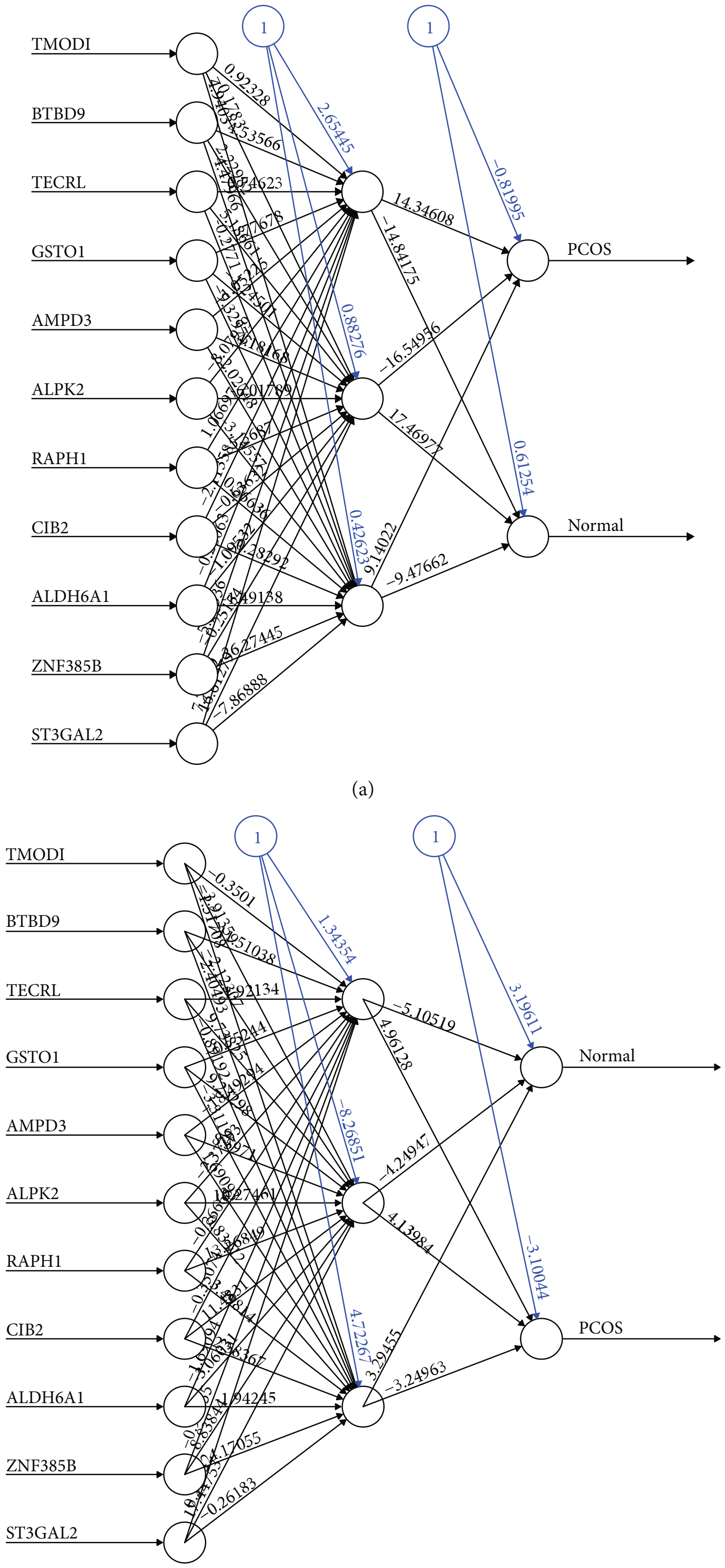


(c)

(d)

Figure 4: Screening candidate PCOS-specific genes by random forest. (a) Parameter selection for random forest classifier. The scatter plot of the variables and corresponding error rate. The x-axis is the number of variables, and the y-axis is the error rate. (b) The influence of the number of decision trees on the error rate. The x-axis is the number of decision trees and the y-axis is the error rate. (c) Ranking of input variables in the random forest model to classify PCOS and normal samples. Top 12 key genes were listed from the most important ones to the least ones based on MeanDecreaseAccuracy and MeanDecreaseGini. (d) Top 12 key genes in MeanDecreaseGini. The x-axis represents the genes, and the y-axis is the importance index.

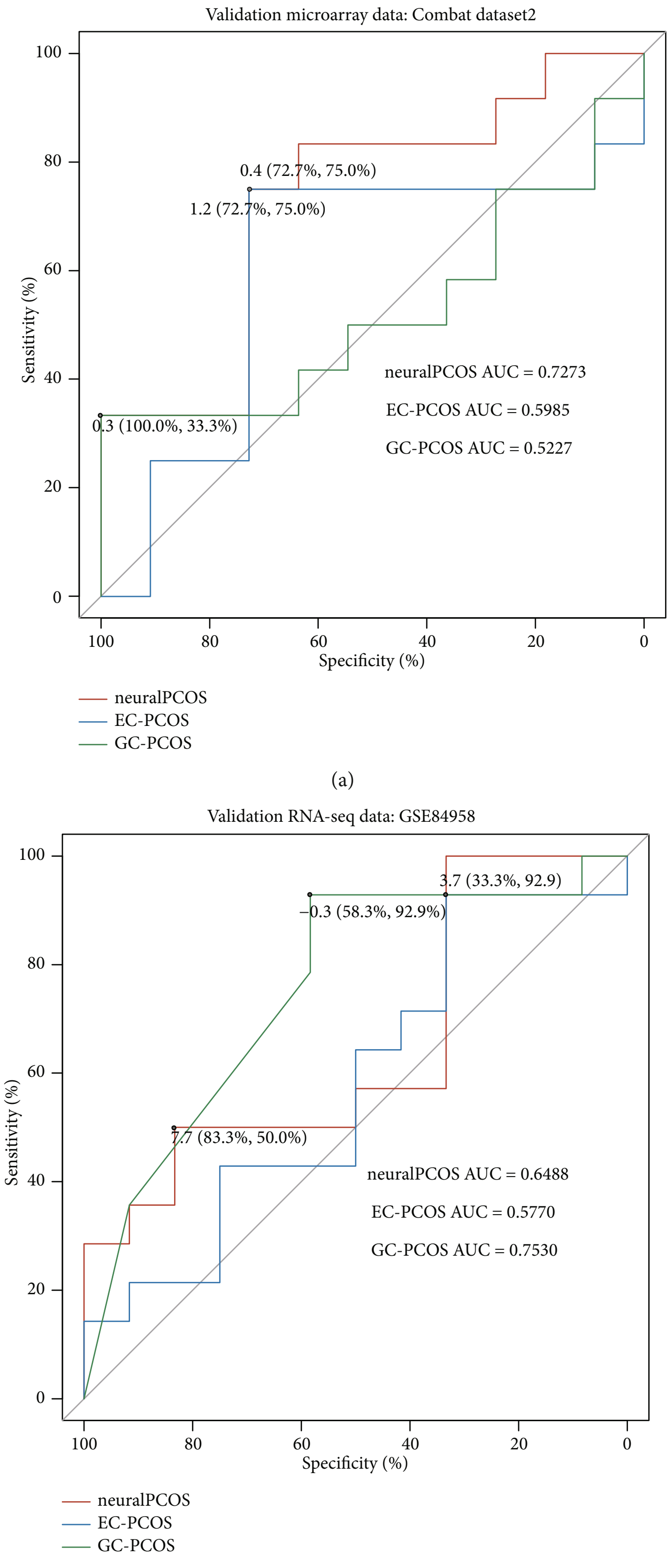
(0.7530). The optimal threshold values for 3 models were 7.7, 3.7, and -0.3, respectively. NeuralPCOS had the highest level of sensitivity than EC-PCOS and GC-PCOS (Figure 6(b)). From the above results, it can be concluded that the classification model established in this study was more suitable in microarray data than in RNA-seq data.



(a)

(b)

Figure 5: Neural network topology of two kinds of training data. (a) Neural network topology of microarray ComBat dataset1 (GSE137684, GSE137354, and GSE34526, $n = 28$) with 11 input layer, 3 hidden layer, and 2 output layer. (b) Neural network topology of RNA-seq data GSE84958 (adipose, $n = 26$) with 11 input layer, 3 hidden layer, and 2 output layer.



(a)

(b)

Figure 6: Performance evaluation of different classification models by the area under the receiver operating characteristic (ROC) curves and their AUC values. (a) In microarray ComBat dataset2 (GSE43264 and GSE124226, $n = 23$), neuralPCOS achieved superior performance (AUC: 0.7273), compared to the other two methods: EC-PCOS (AUC: 0.5985) and GC-PCOS (AUC: 0.5227). (b) In GSE84958 RNA-seq validation data (adipose, $n = 27$), neuralPCOS achieved an AUC of 0.6488, EC-PCOS (AUC: 0.5770) and GC-PCOS (AUC: 0.7530). The optimal threshold values were labeled at inflection points, and the sensitivities and specificities were listed in the bracket.

4. Discussion

In recent years, the development of machine learning algorithms and the availability of gene expression data in public databases provide approaches to infer biomarkers for disease diagnosis or prognosis in a wide range of fields [30–33]. In the field of PCOS, some attempts have been made to explore a better way for PCOS diagnosis by using various machine learning algorithms [34–38], among which, suitable algorithms using some clinical data, such as survey data [35] or pelvic ultrasound data, were used [37]. An algorithm was ever constructed to predict new PCOS candidates using the data from Polycystic Ovary Syndrome Database (PCOSDB; http://www.pcosdb.net/) [39] and the KnowledgeBase on Polycystic Ovary Syndrome (PCOSKB; http://pcoskb.bicnirrh.res.in) [36, 40]. Another study converted the ovary microarray data of GEO database to the gene set regularity (GSR) indices, and the GSR indices were then computed by the modified differential rank conversion algorithm [38]. Comparing with these studies, we aimed to develop a diagnostic model based on gene expression data using as many samples as possible from GEO database. We finally integrated RF and ANN algorithms to infer the key classification genes and calculate the weights of these genes.

In the present study, when identifying DEGs with GSE6798 dataset, we removed the DEGs with low expression level, which can obtain more authentic genes. GO enrichment analysis was performed and displayed by bar plot and bubble plot. Among the 11 enriched GO terms, 4 terms including actin binding [41], myofibril [42], sarcomere [42], and contractile fiber part [42] were also identified in other PCOS researches. We listed the top 12 core genes screened by the RF model for classification in DEGs based on MeanDecreaseGini. Moreover, 10 of the 12 genes were also regarded as PCOS candidate genes in other studies: tropomodulin 1 (*TMOD1*) [43]; BTB domain containing 9 (*BTBD9*) [44]; trans-2,3-enoyl-CoA reductase like (*TECRL*) [44, 45]; glutathione S-transferase omega 1 (*GSTO1*) [44, 46, 47]; adenosine monophosphate deaminase 3 (*AMPD3*) [45]; alpha kinase 2 (*ALPK2*) [48]; Ras association (RalGDS/AF-6) and pleckstrin homology domains 1 (*RAPH1*) [44, 45, 48, 49]; aldehyde dehydrogenase 6 family member A1 (*ALDH6A1*) [44, 45, 50–52]; zinc finger protein 385B (*ZNF385B*) [53]; ST3 Beta-galactoside alpha-2,3-sialyltransferase 2 (*ST3GAL2*) [44]. Given that RNA-seq technology has the superiorities to detect novel transcripts with wider dynamic range, higher specificity, and higher sensitivity than microarray technology [54], the gene expression data obtained by these two technologies may have some differences. In the study, we calculated the weights of core genes by ANN using each type of data separately. Although the weights of only 11 genes in both microarray data and RNA-seq data were calculated, 10 genes were verified in previous studies in both platforms. The novelty of our diagnostic model was that the scoring model was obtained by comprehensively considering the genes those are vital to classification and their weights. In order to validate the applicability and superiority of this model in different types of data, AUC analysis was performed in microarray ComBat dataset2 ($n = 23$) and RNA-seq validation dataset (GSE84958, $n = 27$). In the meanwhile, two sets of marker genes in other researches were also evaluated. One set of genes was the upregulated genes that involved in the insulin signaling pathway (*IGF1*, *PTEN*, and *IGFBP1*) [20]; the other was the upregulated genes including *HSD3B2*, *STAR*, *INHBA*, and *CYP19A1* [21]. The results of AUC scores indicated that our model achieved a superior performance compared with the other two sets of genes in microarray data, and moderate performance but highest level of sensitivity in RNA-seq data. Our model got high AUC scores, indicating it could separate PCOS samples from normal samples with a good probability in microarray data.

Even so, our study still has some limitations. Although our total sample size is not too small (PCOS: $n = 76$; normal: $n = 57$), the number of sample size in each dataset is small, and the individuals in integrated microarray training dataset are from different countries. To get microarray training and validation datasets with larger sample size, 3 and 2 small sample size datasets were combined, respectively. Although the batch effect was removed, it was still not the most suitable datasets. Another drawback of our study is that the expression data are from diverse tissues containing skeletal muscle, adipose, endometrium, and granulosa cells. Last but not least, we did not perform 10 fold cross-validation in ANN analyse due to the limited sample size. Although this is a compromising strategy in the case of limited sample size, our model has an excellent classification performance, a diagnostic model for single tissue type still needs to be constructed with more convincing datasets and machine learning algorithms in the future.

5. Conclusions

A novel diagnostic model for PCOS was established based on machine learning algorithms using microarray and RNA-seq datasets, which showed better prediction performance in microarray data than using existing marker genes.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Ning-Ning Xie and Fang-Fang Wang contributed equally to this work.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (81874480; 81873837).

Supplementary Materials

Supplementary 1. Figure S1: boxplot for gene expression data in GSE6798 dataset. The abscissa axis indicates 29 samples in GSE6798 dataset. Axis of ordinates represents gene expression level.

Supplementary 2. Table S1: the selected 264 differentially expressed genes (DEGs) in GSE6798 dataset.

Supplementary 3. Table S2: significantly enriched gene ontology (GO) terms in biological process (BP), cell components (CC), and molecular function (MF).

Supplementary 4. Table S3: the weight of each gene for microarray ComBat dataset1.

Supplementary 5. Table S4: the weight of each gene for RNA-seq data.

References

[1] R. J. Norman, D. Dewailly, R. S. Legro, and T. E. Hickey, "Polycystic ovary syndrome," *Lancet*, vol. 370, no. 9588, pp. 685–697, 2007.

[2] J. K. Zawadzki and A. Dunaif, "Diagnostic criteria for polycystic ovary syndrome: towards a rational approach," *Polycystic Ovary Syndrome*, pp. 377–384, 1992.

[3] The Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome," *Fertility and Sterility*, vol. 81, no. 1, pp. 19–25, 2004.

[4] The Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS)," *Human Reproduction*, vol. 19, no. 1, pp. 41–47, 2004.

[5] R. Azziz, E. Carmina, D. Dewailly et al., "The Androgen Excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report," *Fertility and Sterility*, vol. 91, no. 2, pp. 456–488, 2009.

[6] B. C. J. M. Fauser, B. C. Tarlatzis, R. W. Rebar et al., "Consensus on women's health aspects of polycystic ovary syndrome (PCOS): the Amsterdam ESHRE/ASRM-Sponsored 3rd PCOS Consensus Workshop Group," *Fertility and Sterility*, vol. 97, no. 1, pp. 28–38.e25, 2012, e25.

[7] M. R. Jones and M. O. Goodarzi, "Genetic determinants of polycystic ovary syndrome: progress and future directions," *Fertility and Sterility*, vol. 106, no. 1, pp. 25–32, 2016.

[8] I. Sirota, D. E. Stein, M. Vega, and M. D. Keltz, "Increased insulin-resistance and beta-cell function in polycystic ovary syndrome women-does ethnicity play a role?," *Reproductive Sciences*, vol. 20, no. S3, pp. 180a-181a, 2013.

[9] Y. Louwers, O. Lao, and M. Kayser, "Inferred genetic ancestry versus reported ethnicity in polycystic ovary syndrome (PCOS)," *Human Reproduction*, vol. 28, pp. 349–349, 2013.

[10] R. Azziz, U. Ezeh, M. Pall, D. A. Dumesic, and M. O. Goodarzi, "Effect of race on the metabolic dysfunction of Polycystic Ovary Syndrome (PCOS): comparing African-American (AA) and Non-Hispanic White (NHW) patients.," *Endocrine Reviews*, vol. 31, no. 3, 2010.

[11] B. J. Vilhjálmsson, J. Yang, H. K. Finucane et al., "Modeling linkage disequilibrium increases accuracy of polygenic risk scores," *The American Journal of Human Genetics*, vol. 97, no. 4, pp. 576–592, 2015.

[12] P. J. Talmud, J. A. Cooper, R. W. Morris et al., "Sixty-five common genetic variants and prediction of type 2 diabetes," *Diabetes*, vol. 64, no. 5, pp. 1830–1840, 2015.

[13] M. B. Kursa, "Robustness of Random Forest-based gene selection methods," *BMC Bioinformatics*, vol. 15, no. 1, p. 8, 2014.

[14] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S.-M. Ngai, and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods," *Molecular BioSystems*, vol. 11, no. 3, pp. 791–800, 2015.

[15] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," *Computers in Biology and Medicine*, vol. 48, pp. 1–7, 2014.

[16] Y. Kong and T. Yu, "A deep neural network model using random forest to extract feature representation for gene expression data classification," *Scientific Reports*, vol. 8, no. 1, article 16477, 2018.

[17] C.-H. Hsieh, R.-H. Lu, N.-H. Lee, W.-T. Chiu, M.-H. Hsu, and Y.-C. (. J.). Li, "Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks," *Surgery*, vol. 149, no. 1, pp. 87–93, 2011.

[18] Z. Zhang, L. Chen, B. Humphries et al., "Morphology-based prediction of cancer cell migration using an artificial neural network and a random decision forest," *Integrative Biology*, vol. 10, no. 12, pp. 758–767, 2018.

[19] R. Janßen, J. Zabel, U. von Lukas, and M. Labrenz, "An artificial neural network and Random Forest identify glyphosate-impacted brackish communities based on 16S rRNA amplicon MiSeq read counts," *Marine Pollution Bulletin*, vol. 149, p. 110530, 2019.

[20] M. N. Shafiee, C. Seedhouse, N. Mongan et al., "Up-regulation of genes involved in the insulin signalling pathway (*IGF1*, *PTEN* and *IGFBP1*) in the endometrium may link polycystic ovarian syndrome and endometrial cancer," *Molecular and Cellular Endocrinology*, vol. 424, pp. 94–101, 2016.

[21] L. A. Owens, S. G. Kristensen, A. Lerner et al., "Gene expression in granulosa cells from small antral follicles from women with or without polycystic ovaries," *The Journal of Clinical Endocrinology and Metabolism*, vol. 104, no. 12, pp. 6182–6192, 2019.

[22] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[23] M. E. Ritchie, B. Phipson, D. Wu et al., "*limma* powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, article e47, 2015.

[24] W. Li, "Volcano plots in analyzing differential expressions with mRNA microarrays," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 6, p. 1231003, 2012.

[25] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS: A Journal of Integrative Biology*, vol. 16, no. 5, pp. 284–287, 2012.

[26] W. Walter, F. Sánchez-Cabo, and M. Ricote, "GOplot: an R package for visually combining expression data with functional analysis," *Bioinformatics*, vol. 31, no. 17, pp. 2912–2914, 2015.

[27] L. Breiman, “Machine learning, volume 45, number 1- springer link,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] F. Günther and S. Fritsch, “neuralnet: training of neural networks,” *The R Journal*, vol. 2, no. 1, pp. 30–38, 2010.

[29] X. Robin, N. Turck, A. Hainard et al., “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[30] A. A. Tabl, A. Alkhateeb, W. ElMaraghy, L. Rueda, and A. Ngom, “A machine learning approach for identifying gene biomarkers guiding the treatment of breast Cancer,” *Frontiers in Genetics*, vol. 10, p. 256, 2019.

[31] D. Wang, J. R. Li, Y. H. Zhang, L. Chen, T. Huang, and Y. D. Cai, “Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms,” *Genes*, vol. 9, no. 3, p. 155, 2018.

[32] C. Wang, W. Pu, D. Zhao et al., “Identification of hyper-methylated tumor suppressor genes-based diagnostic panel for esophageal squamous cell carcinoma (ESCC) in a Chinese Han population,” *Frontiers in Genetics*, vol. 9, p. 356, 2018.

[33] Y. Zhang, J. T. C. Tseng, I. C. Lien, F. Li, W. Wu, and H. Li, “mRNAsi index: machine learning in mining lung adenocarcinoma stem cell biomarkers,” *Genes*, vol. 11, no. 3, p. 257, 2020.

[34] D. K. Meena, D. M. Manimekalai, and S. Rethinavalli, “A novel framework for filtering the PCOS attributes using data mining techniques,” *International Journal of Engineering Research & Technology*, vol. 4, no. 1, pp. 702–706, 2015.

[35] B. Vikas, B. Anuhya, K. S. Bhargav, S. Sarangi, and M. Chilla, “Application of the apriori algorithm for prediction of Polycystic Ovarian Syndrome (PCOS),” in *Information Systems Design and Intelligent Applications*, pp. 934–944, Springer, 2018.

[36] X. Z. Zhang, Y. L. Pang, X. Wang, and Y. H. Li, “Computational characterization and identification of human polycystic ovary syndrome genes,” *Scientific Reports*, vol. 8, no. 1, article 12949, 2018.

[37] J. J. Cheng and S. Mahalingaiah, “Data mining polycystic ovary morphology in electronic medical record ultrasound reports,” *Fertility Research and Practice*, vol. 5, no. 1, p. 13, 2019.

[38] C.-H. Ho, C.-M. Chang, H.-Y. Li, H.-Y. Shen, F.-K. Lieu, and P. S.-G. Wang, “Dysregulated immunological and metabolic functions discovered by a polygenic integrative analysis for PCOS,” *Reproductive BioMedicine Online*, vol. 40, no. 1, pp. 160–167, 2020.

[39] M. Jesintha Mary, U. Vetrivel, D. Munuswamy, and V. Melanathuru, “PCOSDB: PolyCystic Ovary Syndrome Database for manually curated disease associated genes,” *Bioinformation*, vol. 12, no. 1, pp. 4–8, 2016.

[40] S. Joseph, R. S. Barai, R. Bhujbalrao, and S. Idicula-Thomas, “PCOSKB: A KnowledgeBase on genes, diseases, ontology terms and biochemical pathways associated with PolyCystic Ovary Syndrome,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D1032–D1035, 2016.

[41] T. S. Domingues, T. C. Bonetti, D. C. Pimenta et al., “Proteomic profile of follicular fluid from patients with polycystic ovary syndrome (PCOS) submitted to in vitro fertilization (IVF) compared to oocyte donors,” *JBRA Assisted Reproduction*, vol. 23, no. 4, pp. 367–391, 2019.

[42] C. Lu, X. Liu, L. Wang et al., “Integrated analyses for genetic markers of polycystic ovary syndrome with 9 case-control studies of gene expression profiles,” *Oncotarget*, vol. 8, no. 2, pp. 3170–3180, 2017.

[43] E. Jansen, J. S. E. Laven, H. B. R. Dommerholt et al., “Abnormal gene expression profiles in human ovaries from polycystic ovary syndrome patients,” *Molecular Endocrinology*, vol. 18, no. 12, pp. 3050–3063, 2004.

[44] D. Haouzi, S. Assou, C. Monzo, C. Vincens, H. Dechaud, and S. Hamamah, “Altered gene expression profile in cumulus cells of mature MII oocytes from patients with polycystic ovary syndrome,” *Human Reproduction*, vol. 27, no. 12, pp. 3523–3530, 2012.

[45] Z. G. Ouandaogo, N. Frydman, L. Hesters et al., “Differences in transcriptomic profiles of human cumulus cells isolated from oocytes at GV, MI and MII stages after in vivo and in vitro oocyte maturation,” *Human Reproduction*, vol. 27, no. 8, pp. 2438–2447, 2012.

[46] H. Liu, L. Zeng, K. Yang, and G. Zhang, “A network pharmacology approach to explore the pharmacological mechanism of xiaoyao powder on anovulatory infertility,” *Evidence-based Complementary and Alternative Medicine: Ecam*, vol. 2016, article 2960372, 13 pages, 2016.

[47] A. S. Ambekar, D. S. Kelkar, S. M. Pinto et al., “Proteomics of follicular fluid from women with polycystic ovary syndrome suggests molecular defects in follicular development,” *The Journal of Clinical Endocrinology & Metabolism*, vol. 100, no. 2, pp. 744–753, 2015.

[48] V. Skov, D. Glintborg, S. Knudsen et al., “Reduced expression of nuclear-encoded genes involved in mitochondrial oxidative metabolism in skeletal muscle of insulin-resistant women with polycystic ovary syndrome,” *Diabetes*, vol. 56, no. 9, pp. 2349–2355, 2007.

[49] E. Nilsson, A. Benrick, M. Kokosar et al., “Transcriptional and epigenetic changes influencing skeletal muscle metabolism in women with polycystic ovary syndrome,” *The Journal of Clinical Endocrinology & Metabolism*, vol. 103, no. 12, pp. 4465–4477, 2018.

[50] J. Qiao, L. Wang, R. Li, and X. Zhang, “Microarray evaluation of endometrial receptivity in Chinese women with polycystic ovary syndrome,” *Reproductive Biomedicine Online*, vol. 17, no. 3, pp. 425–435, 2008.

[51] H. Xu, Y. Han, J. Lou et al., “PDGFRA, HSD17B4 and HMGB2 are potential therapeutic targets in polycystic ovarian syndrome and breast cancer,” *Oncotarget*, vol. 8, no. 41, pp. 69520–69526, 2017.

[52] J. R. Wood, V. L. Nelson-Degrave, E. Jansen, J. M. McAllister, S. Mosselman, and J. F. Strauss III, “Valproate-induced alterations in human theca cell gene expression: clues to the association between valproate use and metabolic side effects,” *Physiological Genomics*, vol. 20, no. 3, pp. 233–243, 2005.

[53] S. Kenigsberg, Y. Bentov, V. Chalifa-Caspi, G. Potashnik, R. Ofir, and O. S. Birk, “Gene expression microarray profiles of cumulus cells in lean and overweight-obese polycystic ovary syndrome patients,” *Molecular Human Reproduction*, vol. 15, no. 2, pp. 89–103, 2009.

[54] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.