MEDICAL SCIENCE MONITOR

CLINICAL RESEARCH

e-ISSN 1643-3750
© Med Sci Monit, 2020; 26: e919035
DOI: 10.12659/MSM.919035

Received: 2019.07.27
Accepted: 2019.10.22
Available online: 2020.01.21
Published: 2020.02.07

Combined Use of Three Machine Learning Modeling Methods to Develop a Ten-Gene Signature for the Diagnosis of Ventilator-Associated Pneumonia

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

AE Yunfang Cai
B Wen Zhang
CD Runze Zhang
DF Xiaoying Cui
G Jun Fang

Department of Anesthesia, Zhejiang Cancer Hospital, Hangzhou, Zhejiang, P.R. China

Corresponding Author: Jun Fang, e-mail: fangjun477@zjcc.org.cn
Source of support: Departmental surces

Background: This study aimed to use three modeling methods, logistic regression analysis, random forest analysis, and fully-connected neural network analysis, to develop a diagnostic gene signature for the diagnosis of ventilator-associated pneumonia (VAP).

Material/Methods: GSE30385 from the Gene Expression Omnibus (GEO) database identified differentially expressed genes (DEGs) associated with patients with VAP. Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment identified the molecular functions of the DEGs. The least absolute shrinkage and selection operator (LASSO) regression analysis algorithm was used to select key genes. Three modeling methods, including logistic regression analysis, random forest analysis, and fully-connected neural network analysis, also known as also known as the feed-forward multi-layer perceptron (MLP), were used to identify the diagnostic gene signature for patients with VAP.

Results: Sixty-six DEGs were identified for patients who had VAP (VAP+) and who did not have VAP (VAP–). Ten essential or feature genes were identified. Upregulated genes included matrix metallopeptidase 8 (MMP8), arginase 1 (ARG1), haptoglobin (HP), interleukin 18 receptor 1 (IL18R1), and NLR family apoptosis inhibitory protein (NAIP). Down-regulated genes included complement factor D (CFD), pleckstrin homology-like domain family A member 2 (PHLDA2), plasminogen activator, urokinase (PLAU), laminin subunit beta 3 (LAMB3), and dual-specificity phosphatase 2 (DUSP2). Logistic regression, random forest, and MLP analysis showed receiver operating characteristic (ROC) curve area under the curve (AUC) values of 0.85, 0.86, and 0.87, respectively.

Conclusions: Logistic regression analysis, random forest analysis, and MLP analysis identified a ten-gene signature for the diagnosis of VAP.

MeSH Keywords: Diagnosis • Pneumonia, Ventilator-Associated • Transcriptome

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/919035

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Background

Ventilator-associated pneumonia (VAP) is defined as pneumonia that occurs 48 hours or more following mechanical ventilation and extubation [1]. VAP is a hospital-acquired pneumonia that occurs in a large proportion of mechanically ventilated patients (8–28%). Although national surveillance data indicate a decline in the incidence of VAP, worldwide, it remains a common hospital-acquired infection [2]. The mortality rate for patients with VAP is between 24–50%, and can reach 76% when associated with certain pathogens [1]. The mortality associated with VAP remains high, partly because there are no guidelines for prediction of patient susceptibility or risk for VAP [3]. The use of antibiotics for suspected VAP in patients is recommended in the 2005 American Thoracic Society (ATS) guidelines [4]. Prevention measures include modifying known risk factors, but the prediction, prevention, and diagnosis of VAP remain challenging [4].

Currently available bioinformatics databases, including the Gene Expression Omnibus (GEO) database, allow gene expression profiles of human diseases to be studied [5,6]. Differentially expressed genes (DEGs) for disease based on data from the Gene Expression Omnibus (GEO) database have been increasingly reported. In a previous study on gene expression profiling in VAP, Xu et al. [7], used the expression profile GSE30385 to identify 69 DEGs that included 36 down-regulated and 33 upregulated genes in patients with VAP patients. Upregulated genes were mainly associated with pathways and functions related to the mitogen-activated protein kinase (MAPK) signaling pathway and immune response [7]. However, this previous study used traditional bioinformatics analysis and showed that genes, including ELANE, LTF, and MAPK14 [7]. In 2012, a previously published study on VAP by Swanson et al. used a cross-validated logistic regression model to identify five predictive genes, including HCN4, ADAM8, PI3, ATP2A1, and PIK3R3 [8]. However, there was only one algorithm used in establishing the model in this previous study [8].

Therefore, this study aimed to use three modeling methods, logistic regression analysis, random forest analysis, and fully-connected neural network analysis, also known as the feed-forward multi-layer perceptron (MLP), to develop a diagnostic gene signature for the diagnosis and prediction of VAP.

Material and Methods

Gene Expression Omnibus (GEO) database selection

Gene expression profiles were downloaded as raw data (CEL files) from the GSE30385 dataset [8] in the GEO database (*http://www.ncbi.nlm.nih.gov/geo*) [9]. The GPL201 [HG-Focus] Affymetrix Human HG-Focus Target Array served as the annotation platform. In this dataset, whole blood from 20 patient samples was obtained from patients with serious trauma, including ten patients with ventilator-associated pneumonia (VAP) (VAP+) and ten without VAP (VAP–). A total of 40 mL of whole blood was collected and immediately stimulated with 1,000 ng/mL of lipopolysaccharide (LPS) solution.

Data processing

The processing of raw downloaded data, including background correction, quintile normalization as well as probe summarization by robust multi-array average (RMA) algorithm [10], the affy R package [11] in Bioconductor was used (*http://bioconductor.org/packages/release/bioc/html/affy.html*). Then, probe serials were transformed into gene symbols. Mapping multiple probes to the same gene helped to calculate the median probe expression value as the ultimate gene expression value.

Screening for differentially expressed genes (DEGs) for the VAP+ and VAP– patient groups

The Linear Models for Microarray Data (limma) package in R [12] was used to screen the DEGs of the VAP+ and the VAP– groups. DEGs were screened with cutoff values of $p<0.05$ and $|\text{fold change (FC)}| \geq 1.5$ [7]. The eligible DEGs were classified into down-regulated and upregulated DEGs. To ensure two specimen types were in the identified DEGs, a three-dimensional principal component analysis (PCA) was performed using the ggord R package. An expression heatmap was used with the pheatmap R package.

Functional annotation and pathway enrichment analysis

Gene Ontology (GO) term enrichment analysis was performed on DEGs using the clusterProfiler R package to identify the molecular function [13]. The cellular component (CC), molecular function (MF), and biological process (BP) were selected with a cutoff false discovery rate (FDR) of <0.05. For the pathway analysis, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment was analyzed using KOBAS (version 3.0) (*http://kobas.cbi.pku.edu.cn/*) [14]. An FDR <0.05 was considered to be statistically significant.

Protein–protein interaction (PPI) network construction and module analysis

The STRING (version 10.5) (*http://string-db.org/*) [15] database was used to establish a PPI network. The parameter of protein interactions was set at a medium confidence level. Cytoscape (version 3.6.1) software (*http://www.cytoscape.org/*) [16] was used for the visualization and analysis of the PPI network. The key modules of the entire network were screened using

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

the MCODE plug-in. The ClueGO [17], and CluePedia [18] plug-ins of the Cytoscape software were used to perform GO enrichment analysis of the module, with the parameters set to default.

Data preprocessing and manifold learning before building a predictive signature

The range of expression data of all genes was evaluated in the VAP data and was 0–13 without notable outliers. The min-max scaling method was suitable for this type of data. For all patient samples, min-max scaling was used to transform the expression data of a given gene (i) to the range (0, 1), using the following formula:

$$genei\_scaled=(genei–min(genei))/(max(genei)–min\ (genei))$$

The intrinsic geometry of the data structure of VAP data was investigated and was easily visualized in the lower dimensions, and a robust projection method was used to extract essential data. For this type of task, manifold learning algorithms, which is a subfield of machine learning, were developed. In this study, the Isomap nonlinear dimensionality reduction method was chosen [19,20] to project the 66 dimensions of the data into two dimensions, which helped visualize the data geometry and primarily determine the machine learning algorithms that should be used in the modeling stage.

Screening for feature genes

The least absolute shrinkage and selection operator (LASSO) regression analysis algorithm was used to select key genes to build a linear model between target variables and genes with L1 norm constraints. This analysis method was more effective at selecting important features when compared with the traditional least-squared method. LASSO was used to improve the accuracy of the linear model and avoid over-fitting by penalizing coefficients with large values. The linear model derived from LASSO was reduced most of the coefficients to zero, and the features with non-zero coefficients were essential for predicting the target variables or patient labels.

Establishment of a gene signature with diagnostic value for VAP using logistic regression analysis, random forest analysis, and fully-connected neural network analysis

In this study, widely used and validated algorithms, including logistic regression and random forest algorithms, were applied to construct classification models. In particular, a type of deep neural network was applied, namely a fully connected or dense-layer network, to construct a generalized model from the data.

Logistic regression has been used in many machine learning and medical fields and has previously provided good results. As a widely used statistical model in binary classification tasks, the algorithm identifies correlations between features (f) and binary dependent target variables (zero and one tag) on a given dataset and fit a multivariate linear equation (L). Then, the output of the linear equation was passed to the logistic function, and the final probability (P) of data belonging to class one, or VAP–related patients, was obtained. As a result, the patient with an output probability >0.5 was identified as a VAP–related patient, or was identified as a normal patient. In this study, the penalty coefficient of logistic regression was set to 0.02 to obtain good model generalization.

The logistic regression formula used was as follows:

$$L=a_1*f_1+a_2*f_2+....+a_n*f_n \text{ and } P=1/(1+e^{-L})$$

The random forest algorithm introduced randomization to the algorithm to reduce overfitting of a single decision tree and to promote model accuracy by building many related decision trees from one training set. In particular, deep neural networks have been widely applied in a variety of fields, including natural language processing [21], and have achieved performances comparable with those of traditional machine learning algorithms. Therefore, to take full advantage of this type of power model, a fully connected three-layer deep neural network was constructed, also known as the feed-forward multi-layer perceptron (MLP), to investigate the intrinsic relationship between patient types and gene expression data. The MLP model utilized the backward propagation method as a supervised learning algorithm to minimize the error or loss function, a statistical measuring distance between predicted labels and true labels across the neural network, to converge on the network.

Cross-validation and metrics

Given that the number of patient samples in the dataset was relatively small, a reliable leave-one-out (LOO) cross-validation procedure was used to evaluate the model's generalization abilities derived from the three algorithms used in this study, logistic regression analysis, random forest analysis, and fully-connected neural network analysis. Also, the area under the receiver operating characteristic (ROC) curve (AUC) and the metrics of accuracy was applied as quantitative measurements to assess the predicted abilities of the constructed models.

Results

Identification of ventilator-associated pneumonia (VAP) gene expression dataset

A multistep bioinformatic analysis was performed in this study to develop a ten-gene signature to predict VAP (Figure 1).

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

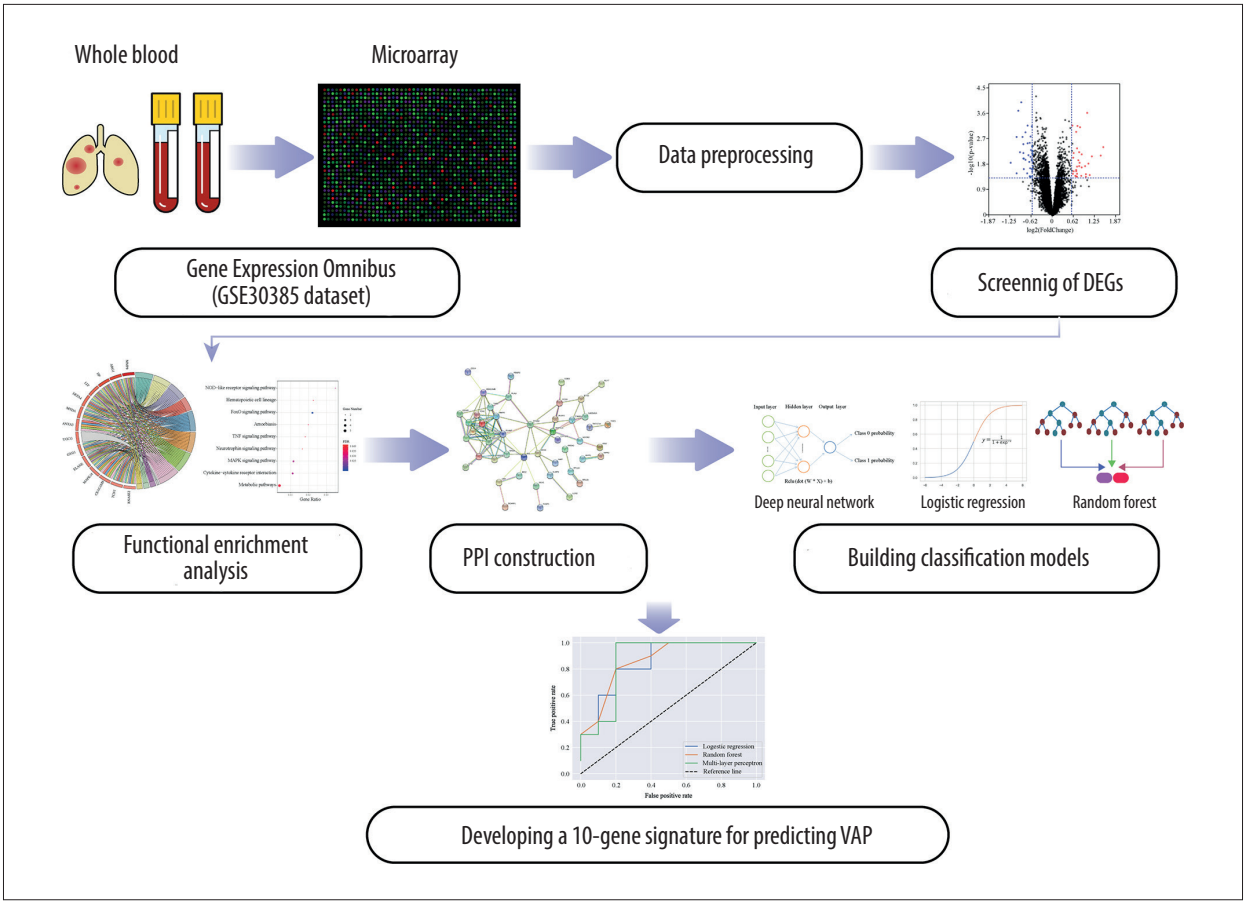

Figure 1. Flow diagram of the study design. The study design for developing a ten-gene signature for predicting ventilator-associated pneumonia (VAP), based on a deep learning neural network and bioinformatics analysis.

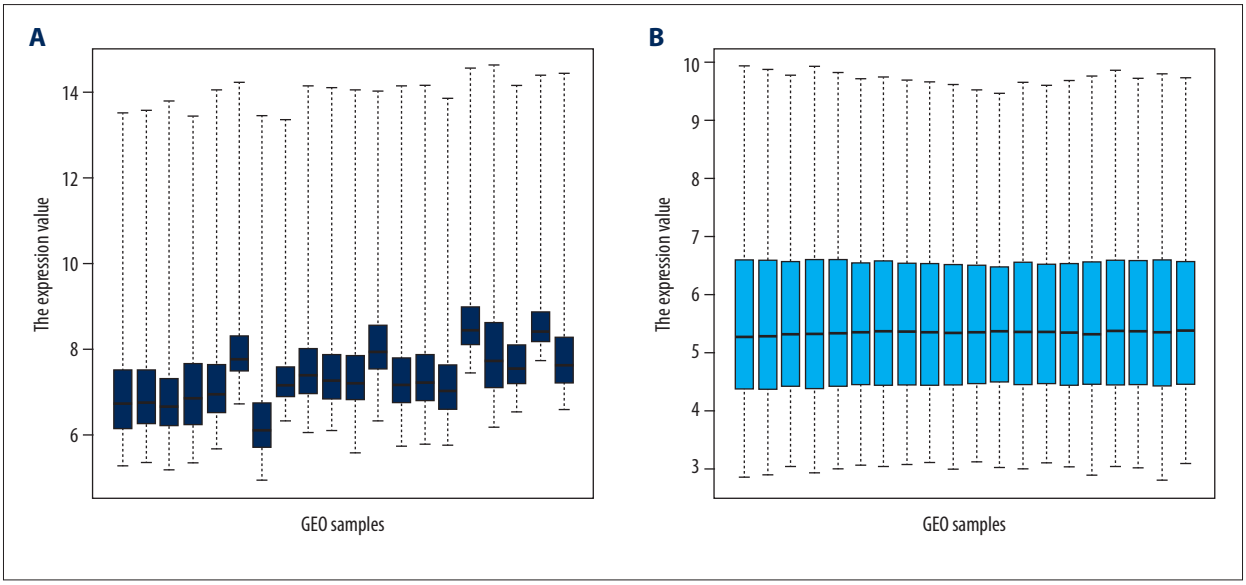

Figure 2. The distribution of gene expression values for ventilator-associated pneumonia (VAP). (**A**) Raw data box plot. (**B**) Normalized data box plot. The abscissa and ordinate represent the Gene Expression Omnibus (GEO) samples and the gene expression value, respectively.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 1. The top ten differentially expressed genes (DEGs) in ventilator-associated pneumonia (VAP).

Gene symbol	Gene name	log2-FC	p-Value	Regulation
MMP8	Matrix metallopeptidase 8	1.871450982	0.002938895	Up
ARG1	Arginase 1	1.488811973	0.003824453	Up
HP	Haptoglobin	1.411448445	0.007541712	Up
IL18R1	Interleukin 18 receptor 1	1.156128987	0.008158229	Up
NAIP	NLR family apoptosis inhibitory protein	1.095823284	0.036770186	Up
LTF	Lactotransferrin	1.051261502	0.014733312	Up
CYP1B1	Cytochrome P450 family 1 subfamily B member 1	0.982426681	0.041859932	Up
DEFA4	Defensin alpha 4	0.979954402	0.034299563	Up
MNDA	Myeloid cell nuclear differentiation antigen	0.959622573	0.017834746	Up
ANXA3	Annexin A3	0.848970244	0.019900178	Up
CTSZ	Cathepsin Z	–0.862858489	0.002942944	Down
THBD	Thrombomodulin	–0.871430229	0.009507344	Down
PPIF	Peptidylprolyl isomerase F	–0.872320694	0.006172789	Down
SLPI	Secretory leukocyte peptidase inhibitor	–0.890979857	0.001666654	Down
PI3	Peptidase inhibitor 3	–0.928381059	9.68E–05	Down
DUSP2	Dual-specificity phosphatase 2	–0.961242077	0.005718815	Down
LAMB3	Laminin subunit beta 3	–1.010787763	0.000192664	Down
PLAU	Plasminogen activator, urokinase	–1.063036567	0.032298508	Down
PHLDA2	Pleckstrin homology like domain family A member 2	–1.070158334	0.001626576	Down
CFD	Complement factor D	–1.242598099	0.013684517	Down

The gene expression files of VAP patients were downloaded from the Gene Expression Omnibus (GEO) database. The Affy R package was used to preprocess the raw data. As shown in Figure 2, box plots of the processed and raw data distribution were prepared. The data distribution was disordered before processing but was consistent after processing and could be used for subsequent analysis.

Identification of differentially expressed genes (DEGs) inVAP+ and VAP– patients

Based on the processed data of the VAP gene expression profiles, 66 significant DEGs were identified, including 35 down-regulated and 31 upregulated genes for the VAP+ and VAP– groups, respectively. The top ten DEGs are shown in Table 1. Upregulated genes included matrix metallopeptidase 8 (MMP8), arginase 1 (ARG1), haptoglobin (HP), interleukin 18 receptor 1 (IL18R1), and NLR family apoptosis inhibitory protein (NAIP). The top down-regulated genes included complement factor D (CFD), pleckstrin homology-like domain family A member 2 (PHLDA2), plasminogen activator, urokinase (PLAU), laminin subunit beta 3 (LAMB3), and dual-specificity phosphatase 2 (DUSP2).

Three-dimensional principal component analysis (PCA) was performed using the above DEGs (Figure 3A) and showed that VAP samples were divided into two groups. The volcano plot of the p-value and fold change are shown in Figure 3B. The whole gene expression of the 66 DEGs is shown in the heatmap in Figure 3C.

Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the DEGs

Functional and pathway enrichment analysis of the biological classification of the above DEGs was performed. The results of the analysis of the GO terms analysis are shown in Supplementary Table 1. In the biological process (BP) category, upregulated DEGs were associated with a significant increase in neutrophil activity, including neutrophil activation in the immune response and degranulation. In the molecular function (MF) category, upregulated genes were enriched in serine hydrolase activity, serine-type peptidase and endopeptidase activity, and glucosyltransferase activity. In the cellular component (CC) category, upregulated DEGs were associated

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

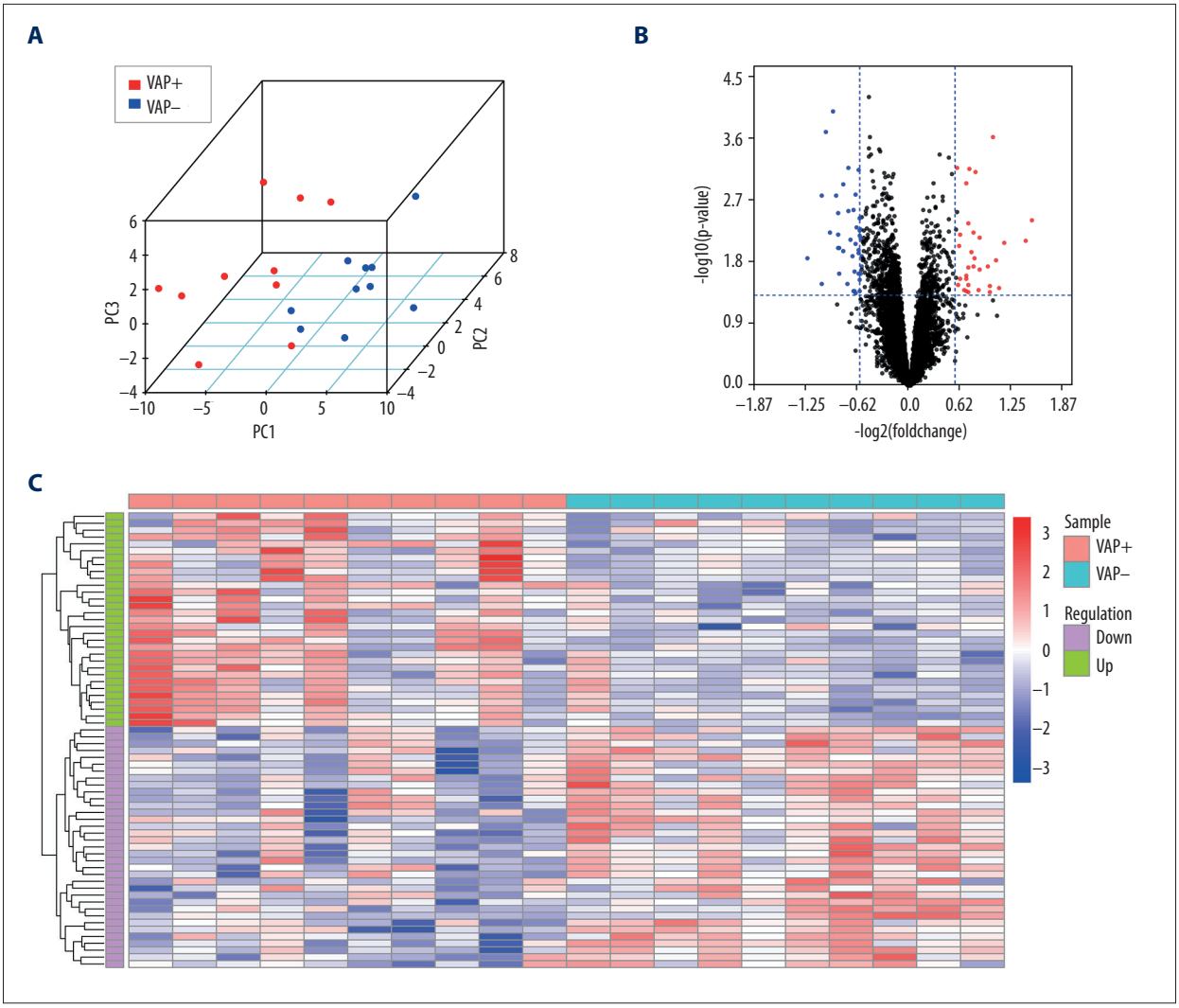

Figure 3. Identification of differentially expressed genes (DEGs) in ventilator-associated pneumonia (VAP) between the VAP+ and VAP– groups. (**A**) Three-dimensional principal component analysis (PCA). The red points represent VAP+ samples, and the blue points represent VAP– samples. (**B**) The volcano plot of DEGs. Blue dots denote down-regulated genes, and red dots represent upregulated genes. (**C**) The expression heatmap of DEGs.

with the cytoplasmic vesicle lumen, and secretory granule lumen (Figure 4A).

Down-regulated genes showed significant enrichment in the upregulation of responses to external stimuli, down-regulation of hydrolase activity, negative regulation of proteolysis and endopeptidase activity, and differentiation of brown fat cells in the BP category. In the MF category, down-regulated genes were enriched in endopeptidase inhibitor activity, peptidase inhibitor activity, endopeptidase regulator activity, cyclosporin A binding, and enzyme inhibitor activity. In the CC category, down-regulated genes were enriched in the endoplasmic reticulum lumen (Figure 4B).

The KEGG pathway enrichment was analyzed for the identified DEGs using the KOBAS database (Supplementary Table 2). The analysis showed a significant increase of upregulated DEGs in the hematopoietic cell lineage, interactions between cytokine receptors, metabolic pathways, amoebiasis, and MAPK, NOD-like receptor, FoxO, neurotrophin, and tumor necrosis factor (TNF) signaling pathways (Figure 4C). The down-regulated genes showed enrichment in the PI3K-Akt signaling pathway, proteoglycans in cancer, Epstein-Barr virus (EBV) infection, focal adhesion, pathways in cancer, herpes simplex virus (HSV) infection, transcriptional dysregulation in cancer, tuberculosis, the cGMP-PKG signaling pathway, and microRNAs in cancer (Figure 4D).

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

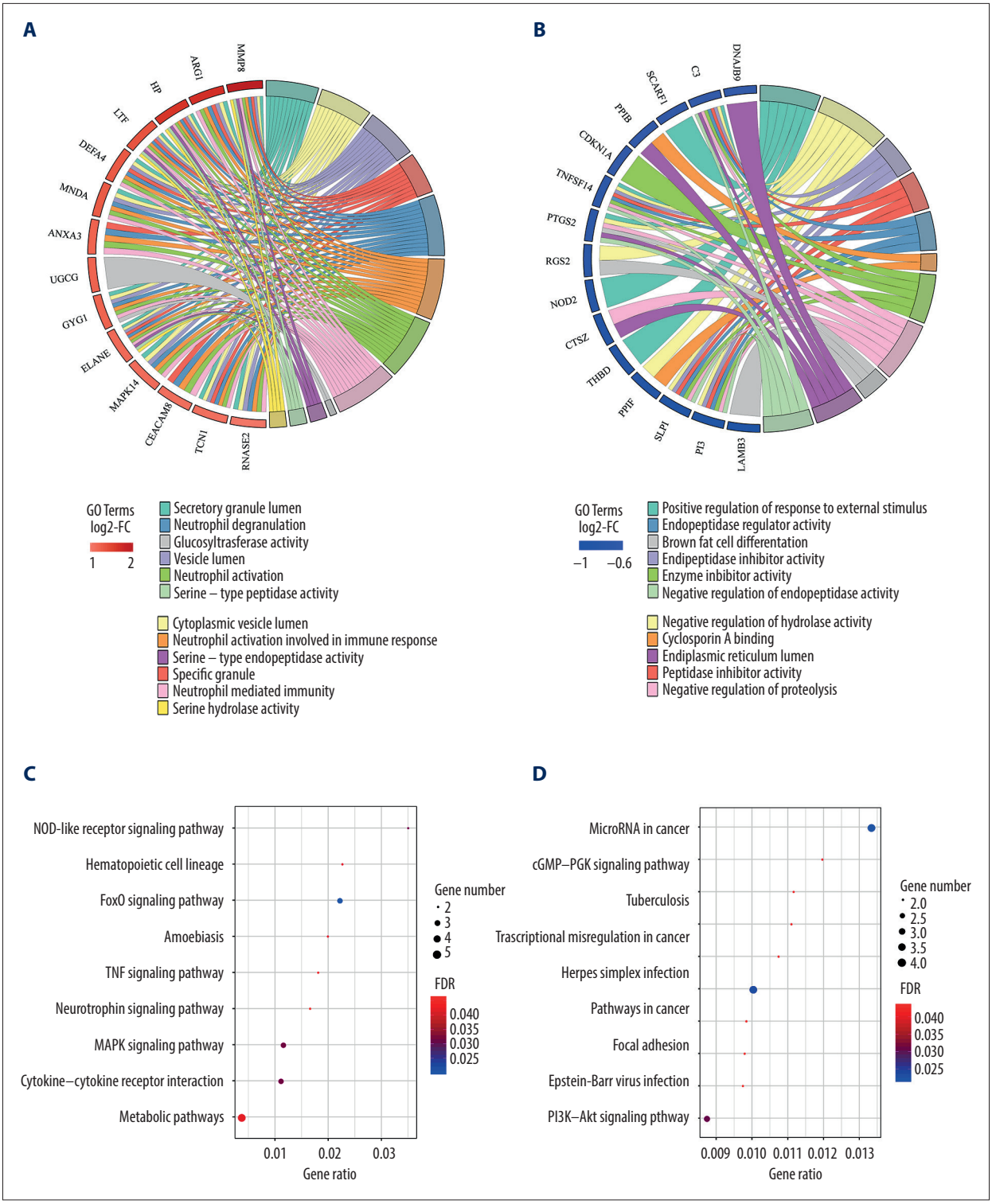

Figure 4. Functional annotation and pathway enrichment analysis of genes associated with ventilator-associated pneumonia (VAP). (**A**) The Gene Ontology (GO) enrichment term results of the upregulated differentially expressed genes (DEGs). (**B**) The GO terms enrichment results of the down-regulated DEGs. (**C**) The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment results of the upregulated DEGs. (**D**) The KEGG pathway enrichment results of down-regulated DEGs.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

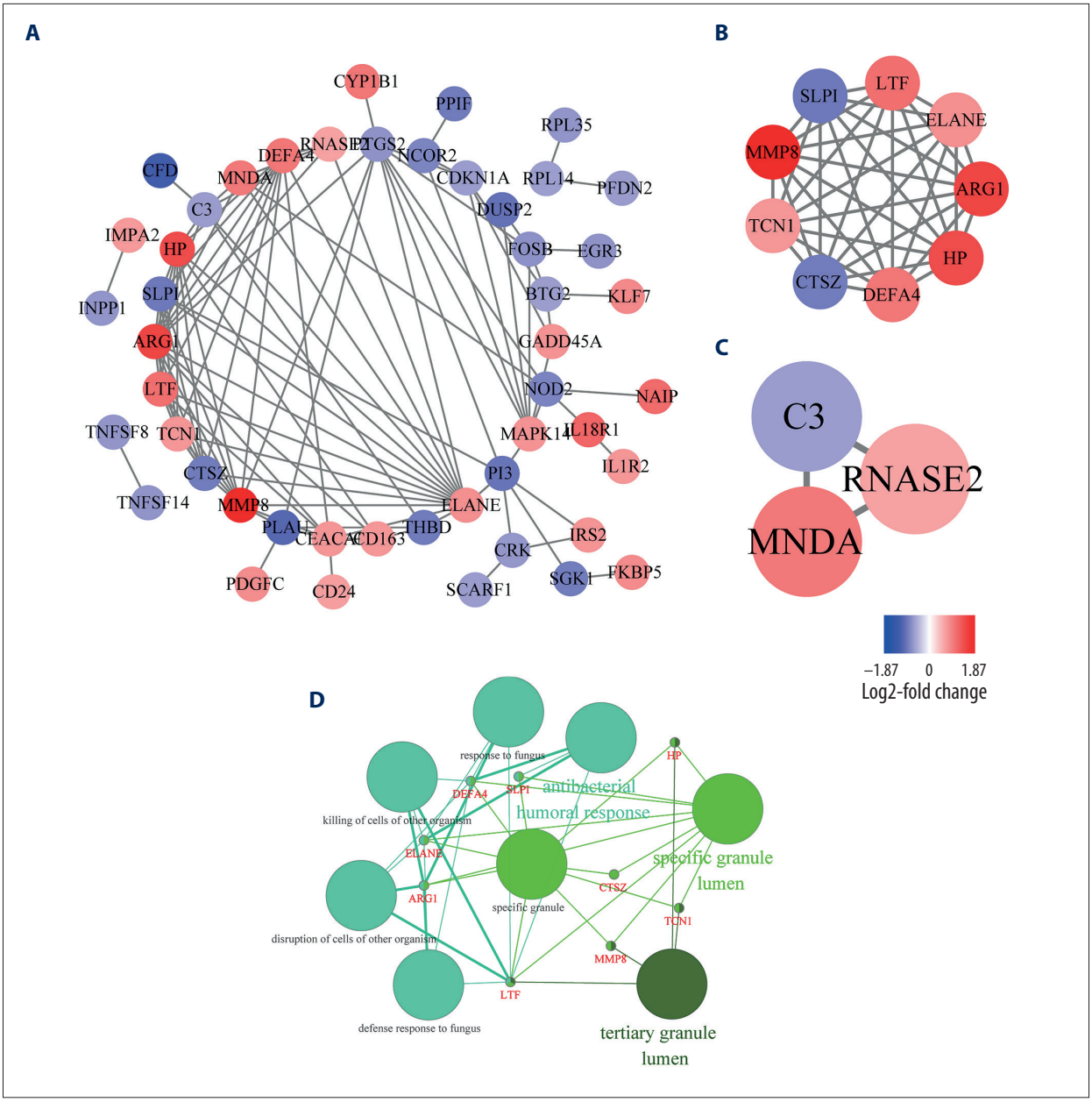

Figure 5. Protein–protein interaction (PPI) network construction and module analysis based on differentially expressed genes (DEGs). (**A**) The entire PPI network (**B**) Module 1 network. (**C**) Module 2 network. (**D**) The Gene Ontology (GO) enrichment term analysis of module 1.

Protein–protein interaction (PPI) network establishment and module analysis

The STRING database was used to create a PPI network to investigate the biological roles of the identified DEGs (Figure 5A). There were 48 nodes and 106 edges that were identified in the PPI network. Two key modules were identified from the whole network using MCODE (Figure 5B, 5C). ClueGO was used to perform GO term enrichment analysis of genes in module 1 (Figure 5D). Genes were enriched in tertiary granule lumen, specific granule, specific granule lumen, defense response to fungus, disruption of cells of other organisms, and antibacterial humoral response (Supplementary Table 3).

Screening for feature genes

To select feature genes and build a gene signature with diagnostic value for VAP among the DEGs, a series of analyses was performed (Figure 6A). The projected data identified by the Isomap algorithm transformed two dimensions, are shown

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

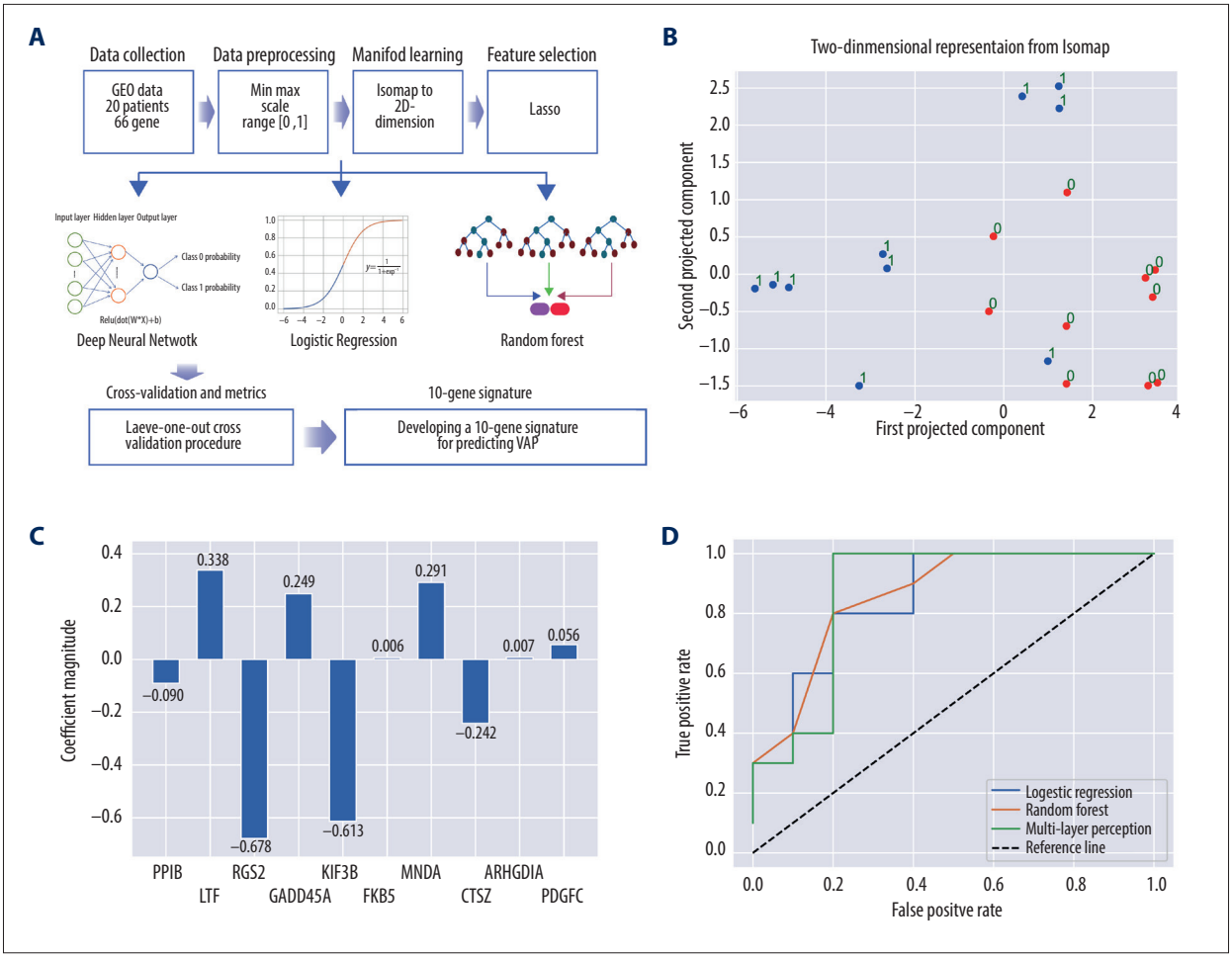

Figure 6. Development of a ten-gene signature for predicting ventilator-associated pneumonia (VAP) using the three modeling methods of logistic regression, random forest, and fully connected neural network analysis. (**A**) The workflow for developing the gene signature. (**B**) The manifold learning algorithm was used to project the data. The points with zero labels identify VAP– patients, with the remaining being VAP+ patients. (**C**) Feature selection of ten genes. (**D**) The receiver operating characteristic (ROC) curves of the diagnostic value of the three algorithms.

Table 2. Details of the ten featured genes in ventilator-associated pneumonia (VAP).

Gene symbol	Gene name	log2-FC	p-Value	Regulation
LTF	Lactotransferrin	1.051261502	0.014733312	Up
MNDA	Myeloid cell nuclear differentiation antigen	0.959622573	0.017834746	Up
FKBP5	FK506 binding protein 5	0.787232999	0.013536848	Up
PDGFC	Platelet derived growth factor C	0.782415364	0.00563081	Up
GADD45A	Growth arrest and DNA damage inducible alpha	0.724582865	0.000672126	Up
ARHGDIA	Rho GDP dissociation inhibitor alpha	0.720167209	0.033516623	Up
PPIB	Peptidylprolyl isomerase B	–0.608265883	0.0035643	Down
RGS2	Regulator of G protein signaling 2	–0.745549324	0.000658377	Down
KIF3B	Kinesin family member 3B	–0.746830664	0.002775964	Down
CTSZ	Cathepsin Z	–0.862858489	0.002942944	Down

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 3. The accuracy and area under the curve (AUC) of the three predicted models in ventilator-associated pneumonia (VAP).

Metrics	Accuracy	AUC
Logistic regression	0.75	0.85
Random forest	0.80	0.86
MLP	0.90	0.87

MLP – multi-layer perceptron.

in Figure 6B, where the binary classes of data are represented by different colors, and annotations on the right corner of each data point show the sample attribute. The majority of data points are mutually separated and can be distinguished by a simple decision boundary. Therefore, the machine learning algorithm was developed without a complex adjustment to fit the data and provide results.

Among the 66 genes, some were closely associated with the two types of patients and might be key biomarkers for identifying patients at increased risk of VAP. A reliable feature selection process was adopted in this study to identify essential genes. After LASSO on the 66 identified DEGs, ten essential genes with coefficients greater than zero were extracted as feature genes (Figure 6C, Table 2).

Building a gene signature exhibiting diagnostic value using three algorithms

The optimal identification of two patient populations was performed using robust machine learning algorithms to build a classification model on the selected feature genes. In this study, widely used and validated algorithms, including logistic regression and random forest algorithms, were applied to construct classification models. In particular, one type of deep neural network was applied, which was a fully connected network or dense-layer network, to construct a generalized model from the data.

Three prevalent and robust algorithms, including one type of deep neural network, the feed-forward multi-layer perceptron (MLP), were used to build the predictive models based on the ten selected genes. The area under the curve (AUC) values of the three models for logistic regression, random forest, and MLP were 0.85, 0.86, and 0.87, respectively (Table 3). In addition to the two metrics, the ROC curves were plotted for each model (Figure 6D). Considering the two metrics simultaneously, the predicted model based on the MLP algorithm was selected, and ten key genes were identified to distinguish between the two types of patients. The predictive ability of the MLP model also indicated that the ten selected essential genes were closely associated with the patients who were diagnosed with VAP. Among these ten genes, six were

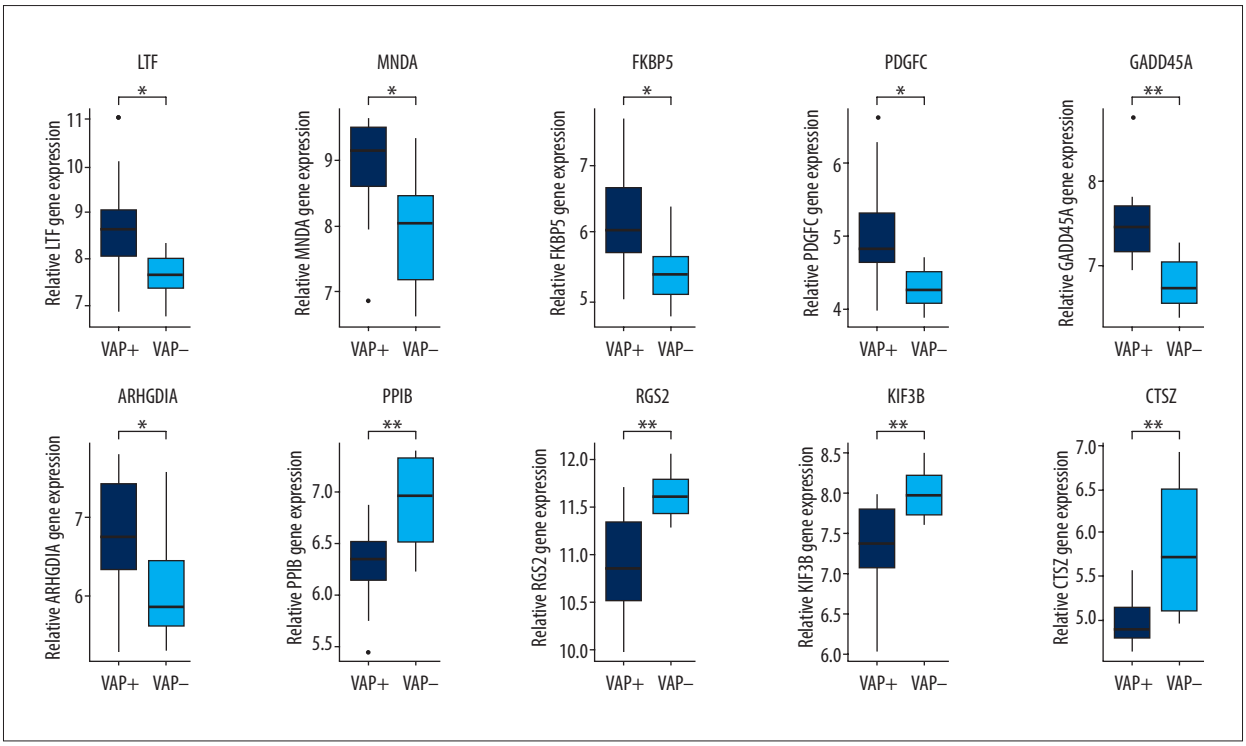

Figure 7. The expression box plot of the ten genes associated with ventilator-associated pneumonia (VAP). * Represents a p-value <0.05. ** Represents a p-value <0.01.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

upregulated, and four were down-regulated. The expression box plots are shown in Figure 7.

Discussion

The aim of this study was to use three modeling methods, logistic regression analysis, random forest analysis, and fully-connected neural network analysis, to develop a diagnostic gene signature for the diagnosis of ventilator-associated pneumonia (VAP). The multistep bioinformatics analysis was performed to identify a ten-gene signature for the diagnosis and prediction of VAP based on the three modeling methods. GSE30385 was identified from the Gene Expression Omnibus (GEO) database to identify differentially expressed genes (DEGs) associated with patients with VAP. A total of 66 significant DEGs were identified between VAP+ and VAP– patients. Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment identified the molecular functions of the DEGs. After applying the least absolute shrinkage and selection operator (LASSO) regression analysis algorithm to select key genes, ten essential genes were identified. Based on three modeling methods, including logistic regression, random forest, and fully-connected neural network methods, a ten-gene signature was identified with diagnostic value for VAP. This ten-gene signature may predict VAP in patients and could be used as potential diagnostic or predictive markers. However, these initial findings require validation with future clinical studies.

VAP acts as a potentially fatal hospital-acquired pneumonia that represents a global health problem [22]. Also, VAP is caused by multidrug-resistant bacteria that also represents an emerging global problem [23]. The diagnosis of VAP remains a challenge. Based on the use of the GEO database, bioinformatics analysis studies have been increasingly reported, but only three previous studies have been reported in gene expression associated with VAP. In 2015, Xu et al. [7] used the expression profile GSE30385 to identify 69 DEGs associated with VAP, including 36 down-regulated and 33 upregulated genes, which differed from the present study in which 66 DEGs were identified. Although the results of this previous study [7], and the present study were similar, the main reason for the differences in the number of DEGs was that the annotation platform GPL 201 was updated in July 2016.

In the present study, in the GO term enrichment analysis, the upregulated genes were found to be associated with the processes of the immune system, immune reaction, and kinase activity, while the down-regulated genes were associated with stronger stress response, peptidase inhibitor activity as well as programmed cell death. Also, upregulated genes exhibited a primary enrichment in the neurotrophic protein signaling pathway, MAPK signal pathway, and the nucleotide-binding oligomeric domain (NOD)-like receptor signal pathway. In contrast, down-regulated genes exhibited a primary enrichment in complement, the coagulation cascade, cancer, ribosomal, and other pathways.

In the present study, neutrophil activities were significantly enriched by upregulated genes, including neutrophil degranulation, neutrophil activation in the immune response, as well as immunity mediated by neutrophils. Neutrophil proteases are significantly increased in the alveolar space in VAP and may contribute to its pathogenesis [24]. Neutrophil extracellular traps are increased in the alveoli in patients with VAP [25]. Also, in the present study, upregulated genes were enriched in the defense responses to fungi and bacteria, immune responses, and antibacterial humoral responses. The findings from this study support the important role of immune responses in the etiology of VAP [26,27]. Genes associated with cell components were closely associated with the lumen, and down-regulated genes were enriched in the regulation of the response to external stimuli, as well as the negative regulation of hydrolase activity, proteolysis, and peptidase activity.

Also, in 2015, Xu et al. [28] reported the findings from a study that used the gene expression profile data of GSE30385 and compared the PPI pairs of all genes from the STRING database, followed by searching VAP-related genes in the National Center for Biotechnology Information (NCBI) to build a PPT network for these genes. Then, they searched the overlapping DEGs and those in the PPI network and showed that the MAPK cascade and processes related to the immune system were enriched in these overlapping genes [28]. Swanson et al. [8] used a logistic regression model with cross validation to develop a gene expression model (PIK3R3, ATP2A1, PI3, ADAM8, and HCN4) for predicting VAP in trauma patients, but this previous study used only one modeling method to build the gene signatures.

In the present study, three modeling methods were used to build a ten-gene signature with diagnostic value in VAP. Among these ten genes, lactotransferrin (LTF) is a multifunctional protein of the transferrin family. Specific receptors presenting on microbial cell surface also interpret lactoferrin antibacterial actions, and in humans, LTF is primarily expressed in mucosal epithelial cells and immune cells [29], and is known for its antimicrobial, antiviral, anti-inflammatory, and immunomodulatory functions [30].

A previously published study included proteomic profiling of bronchoalveolar lavage (BAL) fluid in critically ill VAP patients [31], and the protein lactotransferrin was also found to be a differentially expressed protein in VAP+ patients when compared with VAP– patients. Myeloid cell nuclear differentiation antigen (MNDA), is involved in the activation of the innate

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

immune response and cellular defense response and is an immunohistochemical marker used to distinguish marginal zone lymphomas from other small B-cell lymphomas [32]. Also, the role of MNDA on the proliferation, apoptosis, and migration of osteosarcoma cells has previously been studied [33]. FK506 binding protein 5 (FKBP5) is an important modulator of stress responses and affects the pathogenesis of stress-related disorders [34]. The critical roles of platelet-derived growth factor C (PDGFC) in the cardiovascular system as angiogenic and survival factors have been demonstrated [35].

Growth arrest and DNA damage-inducible alpha (GADD45A) acts as an indicator of DNA damage and responds to environmental stresses by mediating the p38/JNK pathway activation through MTK1/MEKK4 kinase, and has been studied in several human cancers [36–39]. Rho GDP dissociation inhibitor alpha (ARHGDIA) is expressed in glioma [40,41]. Peptidylprolyl isomerase B (PPIB) is expressed in both Gram-negative and Gram-positive bacteria and is an intracellular protein that controls bacterial cell division [42]. Regulator of G protein signaling 2 (RGS2) is expressed in prostate cancer [43], breast cancer [44], and ovarian cancer [45]. Inhibition of kinesin family member 3B (KIF3B) expression can inhibit hepatocellular carcinoma cell proliferation [46]. The gene polymorphisms of cathepsin Z (CTSZ) is expressed in pulmonary tuberculosis [47]. Also, by inducing epithelial-mesenchymal transition (EMT) in hepatocellular carcinoma, CTSZ overexpression is associated with tumor metastasis [48]. However, there have been no previous reports on the role of MNDA, FKBP5, PDGFC, GADD45A, ARHGDIA, PPIB, RGS2, KIF3B, and CTSZ in VAP.

This study had several limitations. The diagnostic signature identified in this study requires further validation in a larger sample size of patients with VAP. Although this model identified key genes associated with increased risk of VAP, a large number of genes were identified, which should be further narrowed to identify the most important genes that can be developed as predictive, diagnostic, or therapeutic biomarkers.

Supplementary Data

Supplementary Table 1. The GO terms enrichment analysis of up- and downregulated genes.

Supplementary Table 2. The KEGG pathway enrichment analysis of up- and downregulated genes.

Supplementary Table 3. The GO terms enrichment analysis of genes in module 1.

Conclusions

This study aimed to use machine learning models to develop a gene signature for the prediction of ventilator-associated pneumonia (VAP). The GSE30385 expression profile was downloaded from the Gene Expression Omnibus (GEO) database, and 66 significant differentially expressed genes (DEGs) were identified, including 35 down-regulated and 31 upregulated genes that distinguished between VAP+ and VAP– patients. According to Gene Ontology (GO) terms used for enrichment analysis, there was a significant increase in the number of upregulated DEGs in neutrophil activity. Down-regulated genes were increased in association with hydrolase activity. Based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, there was a significant increase in the number of upregulated DEGs in FoxO and MAPK signaling pathways. Down-regulated genes saw an enrichment in PI3K/Akt signaling pathway and focal adhesion. After applying the least absolute shrinkage and selection operator (LASSO) regression analysis algorithm on the 66 DEGs, ten essential genes were extracted as feature genes and a ten-gene signature was identified to predict VAP in patients, including LTF, MNDA, FKBP5, PDGFC, GADD45A, ARHGDIA, PPIB, RGS2, KIF3B, and CTSZ. The three modeling methods included logistic regression analysis, random forest analysis, and the feed-forward multi-layer perceptron (MLP), to build a ten-gene diagnostic signature for the diagnosis of VAP. The area under the curve (AUC) values using the three models were 0.85, 0.86, and 0.87, respectively. This ten-gene signature requires further clinical evaluation for the prediction of VAP in patients.

Acknowledgments

The authors thank the staff of the Department of Anesthesia in Zhejiang Cancer Hospital for their technical support.

Conflict of interest

None.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

References:

1. Chastre J, Fagon JY: Ventilator-associated pneumonia. Am J Respir Crit Care Med, 2002; 165(7): 867–903
2. Roberts KL, Micek ST, Juang P, Kollef MH: Controversies and advances in the management of ventilator associated pneumonia. Expert Rev Respir Med, 2017; 11(11): 875–84
3. Nair GB, Niederman MS: Ventilator-associated pneumonia: Present understanding and ongoing debates. Intensive Care Med, 2015; 41(1): 34–48
4. Damas P, Frippiat F, Ancion A et al: Prevention of ventilator-associated pneumonia and ventilator-associated conditions: A randomized controlled trial with subglottic secretion suctioning. Crit Care Med, 2015; 43(1): 22–30
5. Yang C, Ren J, Li B et al: Identification of gene biomarkers in patients with postmenopausal osteoporosis. Mol Med Rep, 2019; 19(2): 1065–73
6. Zeng M, Liu J, Yang W et al: Identification of key biomarkers in diabetic nephropathy via bioinformatic analysis. J Cell Biochem, 2018 [Epub ahead of print].
7. Xu X, Yuan B, Liang Q et al: Gene expression profile analysis of ventilator-associated pneumonia. Mol Med Rep, 2015; 12(5): 7455–62
8. Swanson JM1, Wood GC, Xu L et al: Developing a gene expression model for predicting ventilator-associated pneumonia in trauma patients: A pilot study. PLoS One, 2012; 7(8): e42065.
9. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 2002; 30(1): 207–10
10. Irizarry RA, Hobbs B, Collin F et al: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003; 4(2): 249–64
11. Gautier L, Cope L, Bolstad BM, Irizarry RA: affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 2004; 20(3): 307–15
12. Ritchie ME, Phipson B, Wu D et al: limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res, 2015; 43(7): e47
13. Yu G, Wang LG, Han Y, He QY: clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS, 2012; 16(5): 284–87
14. Xie C, Mao X, Huang J et al: KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res, 2011; 39(Web Server issue): W316–22
15. Szklarczyk D, Morris JH, Cook H et al: The STRING database in 2017: Quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res, 2017; 45(D1): D362–68
16. Shannon P, Markiel A, Ozier O et al: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res, 2003; 13(11): 2498–504
17. Bindea G, Mlecnik B, Hackl H et al: ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics, 2009; 25(8): 1091–93
18. Bindea G, Galon J, Mlecnik B: CluePedia Cytoscape plugin: Pathway insights using integrated experimental and in silico data. Bioinformatics, 2013; 29(5): 661–63
19. Balasubramanian M, Schwartz EL: The isomap algorithm and topological stability. Science, 2002; 295(5552): 7
20. Tenenbaum JB, de Silva V, Langford JC: A global geometric framework for nonlinear dimensionality reduction. Science, 2000; 290(5500): 2319–23
21. Gers FA, Schmidhuber E: LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Trans Neural Netw, 2001; 12(6): 1333–40
22. Poulakou G, Lagou S, Karageorgopoulos DE, Dimopoulos G: New treatments of multidrug-resistant Gram-negative ventilator-associated pneumonia. Ann Transl Med, 2018; 6(21): 423
23. Bassetti M, Vena A, Castaldo N et al: New antibiotics for ventilator-associated pneumonia. Curr Opin Infect Dis, 2018; 31(2): 177–86
24. Wilkinson TS, Conway Morris A, Kefala K et al: Ventilator-associated pneumonia is characterized by excessive release of neutrophil proteases in the lung. Chest, 2012; 142(6): 1425–32
25. Mikacenic C, Moore R, Dmyterko V et al: Neutrophil extracellular traps (NETs) are increased in the alveolar spaces of patients with ventilator-associated pneumonia. Crit Care, 2018; 22(1): 358
26. Smith MA, Hibino M, Falcione BA et al: Immunosuppressive aspects of analgesics and sedatives used in mechanically ventilated patients: An underappreciated risk factor for the development of ventilator-associated pneumonia in critically ill patients. Ann Pharmacother, 2014; 48(1): 77–85
27. Almansa R, Nogales L, Martin-Fernandez M et al: Transcriptomic depression of immunological synapse as a signature of ventilator-associated pneumonia. Ann Transl Med, 2018; 6(21): 415
28. Xu X, Yuan B, Shi Z et al: Identification of crucial genes in ventilator associated pneumonia through protein–protein interaction network. Exp Lung Res, 2015; 41(6): 316–23
29. Mayeur S, Spahis S, Pouliot Y, Levy E: Lactoferrin, a pleiotropic protein in health and disease. Antioxid Redox Signal, 2016; 24(14): 813–36
30. Kanwar RK, Kanwar JR: Immunomodulatory lactoferrin in the regulation of apoptosis modulatory proteins in cancer. Protein Pept Lett, 2013; 20(4): 450–58
31. Nguyen EV, Gharib SA, Palazzo SJ et al: Proteomic profiling of bronchoalveolar lavage fluid in critically ill patients with ventilator-associated pneumonia. PLoS One, 2013; 8(3): e58782
32. Manohar V, Peerani R, Tan B et al: Myeloid cell nuclear differentiation antigen (MNDA) positivity in primary follicles: potential pitfall in the differential diagnosis with marginal zone lymphoma. Appl Immunohistochem Mol Morphol. 2019 [Epub ahead of print]
33. Sun C, Liu C, Dong J et al: Effects of the myeloid cell nuclear differentiation antigen on the proliferation, apoptosis and migration of osteosarcoma cells. Oncol Lett, 2014; 7(3): 815–19
34. Zannas AS, Wiechmann T, Gassen NC, Binder EB: Gene-Stress-Epigenetic Regulation of FKBP5: Clinical and translational implications. Neuropsychopharmacology, 2016; 41(1): 261–74
35. Lee C, Li X: Platelet-derived growth factor-C and -D in the cardiovascular system and diseases. Mol Aspects Med, 2018; 62: 12–21
36. Ishiguro H, Kimura M, Takahashi H et al: GADD45A expression is correlated with patient prognosis in esophageal cancer. Oncol Lett, 2016; 11(1): 277–82
37. Wang HH, Chang TY, Lin WC et al: GADD45A plays a protective role against temozolomide treatment in glioblastoma cells. Sci Rep, 2017; 7(1): 8814
38. Liu J, Jiang G, Mao P et al: Down-regulation of GADD45A enhances chemosensitivity in melanoma. Sci Rep, 2018; 8(1): 4111
39. Wang J, Wang Y, Long F et al: The expression and clinical significance of GADD45A in breast cancer patients. Peer J, 2018; 6: e5344
40. Lu W, Wang X, Liu J et al: Down-regulation of ARHGDIA contributes to human glioma progression through activation of Rho GTPase signaling pathway. Tumour Biol, 2016; 37(12):15783–93
41. Lin X, Yang B, Liu W et al: Interplay between PCBP2 and miRNA modulates ARHGDIA expression and function in glioma migration and invasion. Oncotarget, 2016; 7(15): 19483–98
42. Skagia A, Zografou C, Venieraki A et al: Functional analysis of the cyclophilin PpiB role in bacterial cell division. Genes Cells, 2017; 22(9): 810–24
43. Linder A, Hagberg Thulin M, Damber JE, Welen K: Analysis of regulator of G-protein signalling 2 (RGS2) expression and function during prostate cancer progression. Sci Rep, 2018; 8(1): 17259
44. Lyu JH, Park DW, Huang B et al: RGS2 suppresses breast cancer cell growth via a MCPIP1-dependent pathway. J Cell Biochem, 2015; 116(2): 260–67
45. Cacan E: Epigenetic regulation of RGS2 (Regulator of G-protein signaling 2) in chemoresistant ovarian cancer cells. J Chemother, 2017; 29(3): 173–78
46. Huang X, Liu F, Zhu C et al: Suppression of KIF3B expression inhibits human hepatocellular carcinoma proliferation. Dig Dis Sci, 2014; 59(4): 795–806
47. Hashemi M, Eskandari-Nasab E, Moazeni-Roodi A et al: Association of CTSZ rs34069356 and MC3R rs6127698 gene polymorphisms with pulmonary tuberculosis. Int J Tuberc Lung Dis, 2013; 17(9): 1224–28
48. Wang J, Chen L, Li Y, Guan XY: Overexpression of cathepsin Z contributes to tumor metastasis by inducing epithelial-mesenchymal transition in hepatocellular carcinoma. PLoS One, 2011; 6(9): e24967

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]