International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 3, June 2026, pp. 1286∼1297
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i3.pp1286-1297

Sepsis detection using biomarkers and machine learning

Tuan Anh Vu[1], Dang Hoai Bac[2], Minh Tuan Nguyen[3]
[1]Center for Development of Information Technology and Communications, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[2]Posts and Telecommunications Institute of Technology, Hanoi, Vietnam
[3]Faculty of Telecommunications 1, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam

Article Info

Article history:

Received Jul 19, 2025
Revised Jan 23, 2026
Accepted Mar 16, 2026

Keywords:

Biomarker
Deep learning
Immune-related genes
Machine learning
Sepsis detection

ABSTRACT

Life-threatening dysfunction of organs, known as sepsis, is caused by an imbalanced response of host to infection. In this work, an efficient algorithm is proposed to address vital biomarkers for identification of sepsis using immune-related differential expression genes. A total of 16 gene datasets are processed for the extraction of a gene intersection between different gene datasets and the immune-related gene group, which improve the generalization of the final detection algorithm due to diversity of the input data. A novel gene selection method using sequential forward gene selection, machine learning, and ranked genes based on their importance calculated by a random forest model. A subset of 36 potential immune-related genes, which are identified as the biomarkers from 560 input genes, show an efficiency of the proposed gene selection algorithm. The biomarkers are validated the performance using various machine learning and deep learning related to sepsis diagnosis. The highest statistical performance is shown for the random forest model using the biomarkers as the input with an accuracy of 96.83%, sensitivity of 98.86%, specificity of 86.70%, and AUC of 98.67%. The proposed detection algorithm includes a random forest model and 36 biomarkers, which is simple, effective, and reliable for the applications in clinic environments.

Corresponding Author:

Minh Tuan Nguyen
Posts and Telecommunications Institute of Technology
No. 122, Hoang Quoc Viet, Hanoi 10000, Vietnam
Email: nmtuan@ptit.edu.vn

1. INTRODUCTION

Sepsis disease is caused by an imbalanced response of host to infection, which also known as life-threatening organ dysfunction [1]. For those who are in sepsis, the hyperactive inflammatory response in the early stages results in severe injuries, organ failures and even septic shock for the bodies [2]. Despite advances in the treatment of sepsis, the mortality proportion due to septic shock maintains a significant number, which is from 25% to 30% and even higher [3]. Furthermore, sepsis survivors quite frequently suffer long-term physical, psychological, and cognitive impairment [4] without effective treatments or approved drugs. Hence, intensive care, antibiotics medicine, and hemodynamic stabilization are the main medical treatment methods, emphasizing the urgent need to address biomarkers for the prompt and accurate sepsis identification, which leads to significantly improve the clinical decision-making of technicians and experts in practical environment [5].

Immune-related genes (IRGs) play an important role in response to infection, inflammation, and other immune-related processes of the immune system. In other words, IRGs are considered as biomarkers, which

are the diagnosis and prognostic signatures of various human diseases, such as cancer, exhibiting reliable sensitivity and specificity. Indeed, differential expression analysis of IRGs plays an essential role in biomarker identification related to rapid sepsis detection [6]–[8].

Although IRG expression datasets offer valuable insight into diverse biological processes, identification of essential biomarkers among high-dimensional databases is challenging due to redundant and irrelevant genes. Various differential expression analysis techniques have been developed to overcome such obstacles to improve the accuracy and efficiency for the existing differential expression analysis methods with respect to the selection of informative differential expression genes (DEGs) [9], which are primarily responsible for differences between biological states.

Recently, machine learning (ML) and deep learning (DL) approaches has been widely used to address vital biomarkers in terms of sepsis diagnosis [10]. Indeed, ensemble or multi-algorithm pipelines are proposed in different publications such as [11], which investigates the parameters of 18 ML models to select the optimal models based on the area under the curve (AUC)-receiver operating characteristic (ROC) curve values produced by 10-fold cross validation (CV). Here, ROC is a curve which shows the trade-off between sensitivity and specificity of a classifier for all possible classification thresholds, while AUC measures an area under this curve. The optimal models are then validated with 72 additional samples, resulting in a highest AUC of 85% of the extreme gradient boosting model. A large number of ML models are considered in [12] to address a significant gene subset among immune-related DEGs in which a weighted gene co-expression network analysis (WGCNA) is performed on the original data to identify sepsis-related genes. A number of 108 DEGs, also known as the overlapping gene subset between the immune-related DEGs and sepsis-related genes, are put into different ML models, which leads to a selection of 11 biomarkers. Among the ML models used for the estimation of optimal biomarkers, the penalized discriminant analysis model releases the highest AUC of 90.1% in the validation dataset. In [13], differentially expressed mRNAs are addressed by packages of "Limma" and "metaMA", which are then ranked by mean decrease accuracy values. Here, a forward-wrapper approach combined with different ML models is employed to identify a subset of 15 biomarkers in terms of sepsis detection. The largest validation performance is generated by the RF model with an AUC of 87.3% on the validation data. Lin *et al.* [14] propose the 5-methylcytosine (m5C)-related genes in terms of sepsis, focusing on different immune regulatory mechanisms. As a result, 3 biomarkers are identified from the above 3 subsets of ranked genes corresponding to the individual ML models, which generate AUC values over 70% on testing and validation data. A similar method development is proposed in [15], which addresses 44 DEGs by the use of "limma' package and then 4 biomarkers using different ML models. Performance analysis shows an AUC of 92% on the testing data evaluated by the CIBERSORT algorithm. In [16], GEO database including GSE65682 and GSE95233 is employed to investigate the role of PANoptosis-related genes (PRGs) and their association with characteristics of immune system related to sepsis. Here, the ConsensusClusterPlus algorithm classifies sepsis samples into molecular subtypes to address the DEGs using the package of "limma" with thresholds of $|logFC| > 1$ and $p - value < 0.05$. Furthermore, WGCNA in combination with the cluster analysis considers only sepsis samples of GSE65682 to select the red module genes, which results an intersection between the above genes and the PANoptosis-related DEGs, also known as a biomarker subset of 5 genes. In [17], a total of 308 potential genes are identified as an intersection between DEGs and MEturquoise module genes, which are subsequently subjected to 113 combinations using 12 ML algorithms for performance evaluation. The results indicate 22 biomarkers identified by the RF and Elastic Net models, which show the highest AUC of 88.1% among other model combinations.

Although many studies apply ML techniques to sepsis recognition, most of them rely on small gene-expression datasets, which limits the robustness and generalization of the resulting models [11]–[17]. Moreover, the identification process of DEGs is often insufficiently addressed in existing approaches using a fixed DEG-selection procedures. Therefore, it is potential to miss the informative genes, which are essential for accurate diagnosis. To address these limitations, we use 16 public datasets including various cell types, platforms, and age groups to ensure high generalization of the proposed prediction model. We further propose a novel algorithm to classify sepsis patients from normal people known as controls, which contains an effective ML model and a subset of biomarkers. Here, the sequential forward gene selection algorithm using a 5-fold cross-validation (CV) is employed to identify different potential genes as immune-related DEGs (IRDEGs) using gene importance computed by a ML model. The immune-related DEGs are then validated for their performance in a separated dataset by various ML and DL models to select the final biomarker subset of genes.

The most significant contributions of this work are as follows:

a. Investigation of different IRG frameworks for the extraction of a potential IRG subset which contributes significantly to the diagnosis of sepsis.
b. The utility of a novel gene selection algorithm using an intelligent method for the identification of the IRDEGs, which definitely maintain the relevant number of remarkable genes in terms of the distinction between sepsis and control people.
c. Proposal of an effective sepsis recognition algorithm based on ML techniques and immune-related biomarkers, which is powerful to deploy in medical facilities.

2. DATA

Table 1 shows 16 gene expression datasets, which are downloaded from the GEO and BioStudies databases including eight platforms namely Affymetrix Human Gene 2.0 ST Array, Custom Affymetrix Human Transcriptome Array, Affymetrix Human Gene 2.1 ST Array, Affymetrix Human Transcriptome Array 2.0, Agilent-026652 Whole Human Genome Microarray 4x44K v2, Affymetrix Human Genome U133 Plus 2.0, Affymetrix Human Genome U219 Array, and Agilent Humman Gene Expression 4x44K v2 Micorarry of Biostudies database. There are 2151 participants, which include 468 normal people known as controls and 1683 sepsis patients in the entire database. The total gene databases are randomly divided by datasets, which result in validation set of GSE26378, GSE26440, GSE57065, GSE95233, and GSE119217, while the remaining datasets are allocated to the training set.

Table 1. Data description

Order	Dataset	No. Genes	Control	Sepsis	Cell type	Age
1	GSE119217	28376	12	122	Peripheral blood	Children
2	GSE69686	20299	85	64	Peripheral blood	Post-natal age
3	GSE69063	25512	33	57	Peripheral blood	Adult
4	GSE134347	30905	83	215	Whole blood	Adult
5	GSE131761	21754	15	81	Peripheral blood	Adult
6	GSE57065	23520	25	82	Whole blood	Adult
7	GSE95233	23520	22	102	Whole blood	Adult
8	GSE28750	23520	20	10	Whole blood	Adult
9	GSE26378	23520	21	82	Whole blood	Children
10	GSE8121	23520	15	60	Whole blood	Children
11	GSE13904	23520	18	52	Whole blood	Children
12	GSE26440	23520	32	98	Whole blood	Children
13	GSE9692	23520	15	30	Whole blood	Children
14	GSE4067	23520	15	69	Whole blood	Children
15	GSE65682	19040	42	479	Whole blood	Adult
16	E-MTAB-1548	17028	15	80	Peripheral blood	Adult

3. METHOD

Figure 1 shows the proposed method including three steps, namely gene processing, gene selection, and gene estimation. In the first step, various gene databases are compared with different gene platforms such as the Affymetrix Human Genome U133 Plus 2.0, Affymetric Human Genome U129 Array, Agilent Human Gene Expression 4x44K v2 Microarray, Affymetrix Human Gene 2.0 ST Array, Custom Affymetrix Human Transcriptome Array, Affymetrix Human Gene 2.1 ST Array, Affymetrix Human Transcriptome Array 2.0, and Agilent-026652 Whole Human Genome Microarray 4x44K v2 for the identification of IRGs, which are then preprocessed by different techniques to improve data quality for further analysis. In the second step, the RF model and a 5-fold CV procedure are applied to calculate gene importance values for which the IRGs are ranked. A gene ranking based gene selection algorithm known as sequential forward gene selection (SFGS) is implemented with 3 ML models and 5-fold CV method to select 3 IRG combinations defined as 3 IRDEGs. Finally, different ML and DL models of RF, K-nearest neighbors (KNN), logistic regression (LR), and long short-term memory (LSTM) are used to validate the performance of the selected IRDEGs using 5-fold CV procedure to address the most informative biomarkers. These models are representative of widely used ML and DL techniques. Furthermore, LR, RF, and KNN models handle linear relationships, nonlinear interactions, and local similarity patterns, respectively, while LSTM is able to capture complex non-linear relationships in gene

expression data. The procedure of 5-fold CV includes the input dataset, which is divided into 5 folds. One of these fold is utilized for testing, and the others are applied for model training. The CV procedure is completed with 5 repetitions to ensure that all individual folds are used as the testing data.

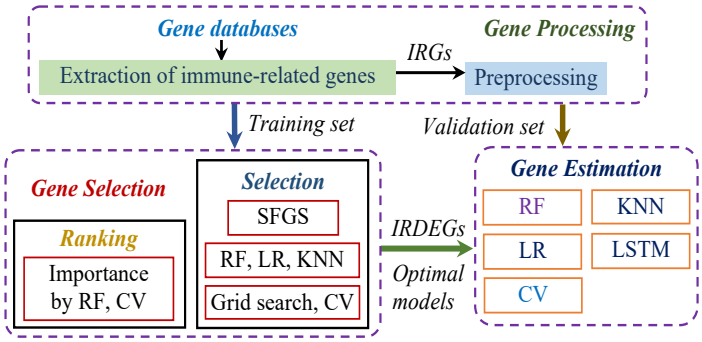

Figure 1. Method diagram

3.1. Gene processing

The gene processing workflow in this study consists of two stages: preprocessing of raw gene expression data and extraction of IRGs. A total of 16 publicly available gene expression datasets from various microarray platforms are aggregated to ensure broad coverage of heterogeneous patient cohorts and measurement conditions. Detailed descriptions of the preprocessing procedures and IRG extraction steps are provided in the following subsections.

3.1.1. Preprocessing

We consider 16 raw gene expression datasets, which are then preprocessed and normalized by the robust multi-array average (RMA) algorithm. Here, gene annotation is performed by mapping probe identifiers to gene symbols, based on the most recent SOFT files or chip description files (CDFs) which are available from the GEO database. SOFT files are used to process 14 gene datasets to set the gene expression level as the mean of the probes for common genes, while custom CDFs are adopted for GSE119217 and GSE69063 to ensure accurate gene mapping. Finally, gene data are preprocessed by Min-Max normalization with a scaling technique in the range of [0-1]. It is noteworthy that no method related to batch effects is considered to ensure the model generalization across independent datasets from various platforms.

3.1.2. Immune-related gene extraction

A total of 8 platforms of gene data are used to address the IRGs from 16 publicly available sepsis databases. Each database is map with IRGs reference set to identify the subset of IRGs related to sepsis. There are 770 IRGs collected from [6] using the publicly accessible NanoString database (*www.nanostring.com*), which are compared with 16 gene databases used in this work to identify potential IRGs related to sepsis. After filtering the IRGs of the platforms, overlapping genes across the platforms are identified by an intersection-based approach. Thereafter, these intersected genes are utilized as input for the gene selection step to identify IRDEGs.

3.2. Gene selection

The gene selection stage aims to identify IRDEGs for sepsis detection through a combination of gene ranking and SFGS. We employ gene ranking-based gene selection namely SFGS, which contains gene importance computed by the RF model in combination with different ML models as the fitness functions and 5-fold CV procedure. The gene selection framework is presented in detail in the following subsections.

3.2.1. Gene ranking

The preprocessed IRGs are put into the RF model to calculate their importance values, which represent the significance of the individual IRGs in terms of the final detection performance for sepsis detection. Specifically, the importance values are defined as scores for all input IRGs computed by a given ML model. Here, the total of IRGs are ranked by the above scores from highest to lowest values in which the higher score shows a greater impact of a specific IRG related to a ML model used to recognize sepsis disease.

3.2.2. Sequential forward gene selection

A gene selection namely SFGS in combination with 3 ML models such as KNN, LR, and RF and 5-fold CV procedure are deployed to select 3 optimal gene subsets, also known as 3 subsets of IRDEGs. The preprocessed IRGs are ranked according to their scores as presented in the previous step. The SFGS selects the first gene with the highest score as the input of 3 ML models to calculate classification performance related to sepsis detection. Then, two genes with the highest importance values are selected to put into the ML models to estimate their performance. The procedure is repeated until the entire preprocessed IRGs are considered for the performance calculation of the ML models. Algorithm 3.1 shows the SFGS combined with different ML models and 5-fold CV procedure.

Algorithm 1. Sequential forward gene selection with ML models

1) Sorting IRGs based on the importance values

G: input set of IRGs; $G(1)$: an IRG with highest importance value; $G(N)$: an IRG with lowest importance value; N: number of IRG; IRGs of G set are sorted descendingly by the importance values from the highest to lowest.

2) Calculating accuracy of ML models using different gene subsets

Training data: $P(i) = \{T(:, G(i)), y\}$; Where *i=1÷N; y = L×1*: label matrix; L: number of samples; *T = i×N*: sample matrix with i genes.

a) Starting with entire data and i genes:
$P(i) = \{T(:, G(i)), y\}$; $i = 1$;

b) **Repeat**
Separation of $P(i)$ into 5 folds by databases $P(i, k)$;
for k=1 to 5
- Model training with $V(i, t)$, $t \neq k$;
- Accuracy calculation on $S(i, k)$;

end
Calculation of the mean accuracy of CV;
Addition of a gene with highest score;
$i = i + 1$;

c) **Until i=N**

3) Immune-related differential gene expression selection

The subsets of IRGs namely IRDEGs are selected with the highest accuracies of the corresponding models. =0

In addition, a gird search-based optimization method is used for identification of the optimal learning and structure parameters of the models to address the overfitting problem. Indeed, the most important learning parameters are investigated for the RF model related to tree number of [25, 55, 75, 95], leaf number of [15, 25, 35, 55], while K of [5, 8, 11, 14, 17, 20, 23, 26, 29] is considered for the KNN model. The learning parameters of the LSTM model are the optimizer of [adam, SGD, RMSprop], batch sizes of [16, 32, 64], learning rate of [0.005, 0.01, 0.02], L2 regularization of [0.8, 0.9, 0.95], epochs of [40, 60, 80]. Here, 5 structures of the LSTM model are employed in which the first structure includes a LSTM and a Batch normalization layer. The second is a combination of 2 first structures, while the third contains the first and the second structure, etc. As a result, there are 1, 16, 9, and 1215 structures of LR, RF, KNN, and LSTM models, which are implemented to identify the optimal models corresponding to the individual subsets of IRGs.

3.3. Gene estimation

We use 3 ML and a DL models, namely RF [18], LR [19], KNN [20] and LSTM [21] to validate the entire input IRGs (AIRG) and 3 subsets of IRDEGs selected by the SFGS algorithm on the validation set using 5-fold CV procedure. Here, 5 folds are generated for the validation set in which each fold corresponds to a completed dataset. Then, 4 folds are for model training and one fold is for testing. The CV procedure is repeated 5 times to ensure that all individual datasets are used as the testing gene data. The mean validation performance of the models and their standard deviation are calculated for further analysis and comparison with previous studies.

4. SIMULATION RESULTS

4.1. Performance measurement

We use accuracy (*Ac*), sensitivity (*Se*), specificity (*Sp*), Mathews correlation coefficient (MCC), and area under the curve (AUC) to estimate the performance of different ML and DL models in this study. *Ac* shows

the rate of participants who are correctly predicted. *Se* and *Sp* present the number of correctly detected sepsis patients and control people, respectively. The discrepancy between patients and controls is measured by the MCC parameter. Furthermore, the AUC evaluates the ability of the ML and DL models to distinguish sepsis patients and control people.

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$Se = \frac{TP}{TP + FN} \tag{3}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

where TN, TP, FN, and FP are true negative, true positive, false negative, and false positive values.

4.2. Gene processing

4.2.1. Preprocessing

The preprocessing stage begins by applying the RMA method to all 16 gene expression datasets to perform background correction, normalization, and probe-level summarization. Following RMA, gene annotation is carried out using the corresponding SOFT and CDFs to accurately map probe identifiers to standardized gene symbols across different platforms. As a result of this procedure, the processed datasets contain between 17028 and 30905 genes, as detailed in Table 1, which are normalized by the Min-Max normalization-based scaling technique in the range of [0-1].

4.2.2. Immune-related gene extraction

The 16 gene expression datasets are filtered for IRGs based on a set of 770 IRGs. As a result, there are 760, 696, 742, 755, 751, 740, 737 extracted IRGs from GSE119217, GSE69686, GSE69063, GSE134347, GSE131761, GSE65682, E-MTAB-1548, respectively. Furthermore, the remaining gene datasets namely GSE57065, GSE95233, GSE28750, GSE26378, GSE1821, GSE13904, GSE26440, GSE9692, GSE4067 produce a similar number of 737 IRGs. We consider a subset of 560 IRGs, which is an intersection between 16 datasets for further analysis to ensure the inclusion of the most common characteristics of all input gene databases related to sepsis in the proposed algorithm.

Table 2. Gene ranked by the important values

Ord	Gene	Imp	Ord	Gene	Imp	Ord	Gene	Imp	Ord	Gene	Imp
1	IL1R2	2.99	13	GATA3	0.44	25	ITGA4	0.34	37	CEACAM8	0.28
2	S100A12	1.80	14	MAGEB2	0.42	26	IFIT1	0.33	38	KLRD1	0.27
3	CCR7	1.58	15	CD3E	0.42	27	CD274	0.32	39	AMMECR1L	0.26
4	IL6ST	0.74	16	ARG1	0.42	28	GZMA	0.32	40	PYCARD	0.26
5	ABCB1	0.66	17	CCR9	0.39	29	CR1	0.32	41	CD80	0.25
6	FCER1A	0.64	18	LRRN3	0.39	30	BATF	0.30	42	ST6GAL1	0.25
7	FCER1G	0.62	19	GNLY	0.38	31	LTB	0.30	43	TXK	0.25
8	C1QA	0.62	20	COLEC12	0.37	32	CR2	0.30	44	CD63	0.25
9	C3AR1	0.61	21	CD3D	0.37	33	HLA_DQA1	0.29	45	C5	0.25
10	CCL28	0.60	22	BCL2	0.36	34	KLRG1	0.29	46	SSX1	0.24
11	BST2	0.55	23	KLRF1	0.35	35	DUSP6	0.28	Others		<0.24
12	CFD	0.46	24	STAT3	0.35	36	IL18R1	0.28			

Imp: Importance value, Ord: Order

4.3. Gene selection

4.3.1. Gene ranking

A total of 560 IRGs are evaluated and ranked according to their importance values, which are computed using the RF model as shown in Table 2. These importance scores represent the contribution of each IRG to the overall classification performance, allowing us to identify genes that are most influential in distinguishing sepsis samples from non-sepsis samples. We only represent the first 46 IRGs with the highest important values due to large number of IRGs investigated in this work.

4.3.2. Sequential forward gene selection

We employ 3 ML models such as LR, KNN, RF as the fitness function of the SFGS algorithm in combination with 5-fold CV procedure to identify optimal IRG subsets. During the selection process, SFGS iteratively adds genes from the ranked list and evaluates each candidate subset using a 5-fold CV procedure to measure its classification performance. The optimal subsets, termed IRDEG1, IRDEG2, and IRDEG3, contain the first 31, 36, and 46 genes, respectively, corresponding to the highest average accuracy achieved by each ML model. These selected gene sets are summarized in Table 2, and their performance are illustrated in Figure 2.

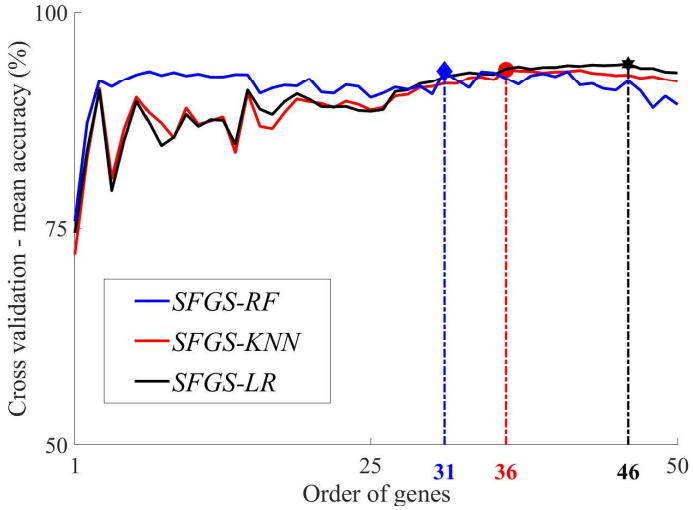

Figure 2. Average accuracy of 5-fold CV for the individual immune gene subsets

Table 3. The largest validation performance of various models using 3 IRDEGs and AIRG on the validation set

Model	DEG	Ac (%)	Se (%)	Sp (%)	MCC (%)	AUC (%)
RF	IRDEG1	94.71±5.68	95.71±6.08	90.65±10.57	83.08±20.29	97.60±4.41
	IRDEG2	**96.83±1.39**	**98.86±2.03**	**86.70±11.85**	**84.97±13.56**	**98.67±2.54**
	IRDEG3	95.60±3.73	97.46±3.18	85.56±15.33	83.02±19.28	97.48±5.16
	AIRG	96.03±5.03	98.07±3.11	81.32±27.81	80.69±31.36	97.17±6.02
KNN	IRDEG1	95.94±3.49	96.09±2.77	93.42±10.50	85.02±17.48	97.68±4.31
	IRDEG2	94.89±3.87	94.95±3.21	91.75±14.17	81.82±19.89	96.70±6.60
	IRDEG3	94.67±7.56	94.52±7.23	94.37±10.91	83.48±25.33	97.17±6.07
	AIRG	94.59±5.38	95.16±3.45	85.71±29.35	77.42±31.53	93.27±13.28
LR	IRDEG1	75.25±28.44	73.61±40.22	81.54±24.71	55.37±31.52	80.26±16.41
	IRDEG2	71.68±27.75	68.96±39.40	82.58±28.76	50.22±30.67	79.41±16.28
	IRDEG3	65.77±32.95	65.01±46.52	72.09±39.39	49.26±22.63	72.54±19.61
	AIRG	49.54±26.84	40.34±37.52	72.13±40.50	47.81±24.89	55.61±9.40
LSTM	IRDEG1	94.52±4.70	95.71±4.51	87.88±10.43	80.69±19.56	97.39±4.30
	IRDEG2	89.33±7.76	92.28±8.56	76.67±38.16	65.37±29.17	93.44±8.60
	IRDEG3	88.97±13.43	90.79±16.21	85.37±18.06	74.02±24.09	94.37±7.48
	AIRG	91.67±5.57	92.86±7.74	84.29±17.06	74.66±19.82	97.86±3.96

4.4. Gene estimation

The optimal parameters of the RF models include 55 trees and 25 leaves, while that of KNN is K=17. Moreover, the structure of optimal LSTM model consists of 3 sequential layers in which LSTM layer is followed by a batch normalization and dropout layer for training stabilization and overfitting reduction, respectively. The extracted temporal representations are then passed through dual layers, namely fully connected and softmax output layers for binary classification. The optimal LSTM model uses the Adam optimizer with a batch size of 32, a learning rate of 0.01, and a L2 regularization coefficient of 0.9. These optimal models are then used for the performance validation of 3 IRDEG subsets on the validation set using 5-fold CV procedure. The mean performance of different ML and DL models such as RF, LR, KNN, and LSTM is given in Table 3.

The RF model produces the highest average Ac of 96.83%, Se of 98.86%, Sp of 86.70%, MCC of 84.97% and AUC of 98.67%, which is selected as the proposed algorithm to classify sepsis disease.

5. DISCUSSION

Sepsis is a dangerous disease for human health, which has received intense attention from medical experts, technicians, and researchers. Existing studies certainly consider different gene databases to develop an effective method for the sepsis detection. However, the number of gene databases is frequently small resulting in unreliability, low performance of the proposed method, showing difficulties for practical application in the clinic environments [14]–[17]. A potential solution to enhance detection performance of the proposed algorithm in previous works is the use of IRG databases, which are involved in immune regulation, response, and proper functioning of the immune system to protect the human body from harmful substances, germs, and cell changes. Hence, we investigate a large number of 16 gene databases to produce better detection performance of the final algorithm in this work. Obviously, the utility of massive gene databases certainly results in avoidance of overfitting problems, improvement of the final classification performance, and increase in reliability of the proposed method. Additionally, the utility of common IRGs from 16 gene databases certainly significantly improves the generalization of the proposed method in terms of sepsis recognition.

Another significant characteristic is the gene selection. Most of existing studies adopt conventional methods such as log(Fold-change) and P-value to address the DEGs [16], [17]. Indeed, log(Fold-change) parameter represents the degree of gene expression in which the up- and down-regulation of genes are based on higher and lower values of log(Fold-change) than zero, respectively. Moreover, statistical method indicates a threshold of 0.05 for which p-value parameter being smaller than such threshold certainly represents biological expression changes. The combination of the above parameters results in an effective method for DEG identification. However, a large number of DEGs as the outcome of conventional method definitely poses an obstacle for the further step of biomarker identification among selected DEGs such as 6361, 1230, 405 [14], [16], [17]. Therefore, a gene ranking-based gene selection method namely SFGS is applied in this work to select the potential IRDEGs from the input gene set of 560 IRGs. Here, the gene important values, which are computed by a RF model, are used to rank 560 IRGs. A total of 560 IRG combinations with number of IRG ranging from 1 to 560 are evaluated by 3 ML models such as RF, LR, and KNN. Consequently, there are 3 subset including 31, 36, 46 IRGs selected by the SFGS in combination with 3 ML models and 5-fold CV procedure on the training set. It is clear that the gene number of the above subsets is smaller than those of [14], [16], [17], which makes it easily to identify a subset of biomarkers.

We implement ML and DL models for comparison with existing publications based on different performance metrics and the proposal of an effective sepsis diagnosis algorithm. Indeed, AUC and MCC performance parameters are widely used for estimation of the proposed methods in previous works [13], [15], [17]. Obviously, AUC metric emphasizes the ability of the models to distinguish between sepsis and control groups, while the overall prediction is measured by accuracy parameter. It is clear that the high classification performance of the final algorithm for sepsis is one of the most important elements for those who develop novel methods related to sepsis recognition. Therefore, the use of numerous metrics for the performance estimation of the sepsis detection algorithm plays an essential role. In this work, 5 parameters are employed for the performance evaluation of various models in terms of sepsis classification, which certainly provide reliable estimation of the proposed algorithm's ability with respect to sepsis diagnosis. Moreover, the grid search combined with 5-fold CV procedure is deployed for identification of the optimal learning and structure parameters, which leads to obtain the best model with relative high sepsis detection performance while avoiding fundamental problems such as overfitting.

The average performance of ML and DL models on the validation set is given in Table 3. The RF and KNN models generate high performance for the sepsis diagnosis with mean Ac and AUC over 94% and 93%, respectively, while LR model shows lowest performance with Ac and AUC less than 75% and 80%. The highest performance with mean Ac of 96.83%, Se of 98.86%, Sp of 86.70%, MCC of 84.97%, and AUC of 98.67% is released by the RF model selected as the final algorithm for the sepsis detection among the others. Here, high sensitivity of the proposed model implies an accurate diagnosis of sepsis cases, which are then definitely checked by clinical experts to make final decision of delivering effective treatment. In the clinical context, emphasizing sensitivity is essential for early detection, as timely intervention can significantly reduce the risk of severe complications and mortality in patients with sepsis. It is noteworthy that examination

of experts is applied for people being incorrectly identified by the proposed model, who are then given no medical treatment. A comparison of the proposed algorithm with existing publications is presented in Table 4 which shows outperformed performance of the proposed algorithm compared with existing studies. Hence, the proposed algorithm is effective for sepsis detection applications in practical facilities and hospitals.

Table 4. Comparisons of the proposed algorithm with previous studies

Ref.	Method	Data	Ac (%)	AUC (%)	MCC (%)	Pros	Cons
[17] 2025	Gene selection and classification using RF and Elastic Net	- 4 datasets (359 samples) - Separated training and testing	NA	88.1	NA	- 113 combinations models, enabling robust, thorough performance evaluation	- Small dataset - Non-optimized model - Only using ML - Only AUC
[13] 2022	- Gene selection using RF ranking and forward-wrapper - Classification using RF	- 5 datasets (958 samples) - 4 datasets for training and testing - A dataset for validation	NA	87.3	71.3	- Robust wrapper-based gene selection - Independent dataset validation	- Small dataset - Non-optimized model - Only using ML
[15] 2023	- Gene selection by intersecting LASSO, SVM-RF, and RF - Classification using CIBERSOFT	- 3 datasets (253 samples) - 2 datasets for training - A dataset for validation	NA	92	NA	Integration of multiple gene selection methods	- Small dataset - Non-optimized model - Only using ML - Model evaluation using only AUC
Our	- Gene importance-based gene ranking - Gene selection using SFGS and ML models - Classification using ML, DL	- 16 datasets (2151 samples) - 11 datasets for training - 5 datasets for validation - 5-fold CV	96.83	98.67	84.97	- Multiple datasets to improve generalizability - Grid search, CV to optimize model - SFGS and ML models for IRDEG selection - High classification performance	- High number of selected biomarkers - Limited exploration of DL models

Existing clinical tools for assessing sepsis, such as SOFA, qSOFA, and procalcitonin, are known as important diagnostic guidance, which still have several limitations. Indeed, evaluation of organ dysfunction across six physiological systems for the diagnosis of sepsis is considered as SOFA score, which is more reactive than predictive [22]. Consequently, sepsis disease is often identified only after the existence of significant organ damages [23], [24]. Similarly, qSOFA was developed and validated in populations with suspected sepsis, making it less suitable as an early screening tool [25]. Small sensitivity and specificity to differentiate sepsis disease based on other causes of systemic inflammatory responses is shown for Procalcitonin, which emphasizes the need for more reliable molecular markers [26]. In contrast, our proposed method utilizes a subset of 36 IRDEGs to detect sepsis promptly and with high performance such as sensitivity of 98.86% and AUC of 98.67%. These findings suggest that our model provides an alternative diagnostic approach with improved accuracy and timeliness in comparison with the existing clinical tools.

The first limitation of this work is the large number of 36 biomarkers, which definitely increases the time, complexity, and cost of gene expression measurement in real-world applications. Secondly, the datasets used in this study exhibit class imbalance, which may introduce bias to model learning and inflate sensitivity while reducing specificity, potentially affecting the generalization of the predictive model. Omission of external validation and analysis limited by the utility of 3 ML and a DL models are certainly other limitations. Indeed, the exploration of different DL models definitely generates a better chance to find a productive model with better sepsis recognition performance, which is absolutely considered in future research.

6. CONCLUSION

Sepsis is a main cause of serious medical conditions, which represent the body's uncontrolled response to infection, leading to organ failure and high mortality. Millions of cases are reported each year, which pose important burden on healthcare systems worldwide. Prompt and accurate recognition is essential to improve patient outcomes. In this work, we propose an algorithm for the sepsis prediction with high generalization

based on the utility of multiple cases, diverse age groups of input gene datasets collected from different medical platforms. The proposed algorithm is developed with the RF model and a subset of 36 biomarkers, which are choosen from the input IRGs. A gene ranking-based gene selection known as the SFGS algorithm utilizing different ML models as the fitness function and 5-fold CV procedure is deployed to select the optimal subset of IRGs, known as IRDEGs, which are then validated their sepsis detection performance by the ML and DL models. The relatively high performance confirms the effectiveness of the SFGS using ML techniques in comparison with the conventional method using log(Fold-Change) and p-value for the identification of DEGs. The RF model releases the highest average sepsis recognition performance with Ac of 96.83%, Se of 98.86%, Sp of 86.70%, MCC of 84.97% and AUC of 98.67% among the other ML and DL models, which shows successful utility of ML model and biomarkers for the sepsis diagnosis. Indeed, we propose an simple but efficient method to archive better massive gene data processing and high level of gene data separation for the sepsis detection. As a result, we suppose that the proposed algorithm is deployed as the application in clinic environments and hospitals. However, the number of biomarkers includes 36 genes, which may increase the practical complexity for clinical implementation, is the first limit of this work. Moreover, the imbalanced datasets, no external validation, and the use of small number of models such as 4 ML and DL models for method development are the additional limitations, which are certainly addressed in future researches.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Tuan Anh Vu		✓	✓		✓	✓		✓	✓		✓		✓	
Dang Hoai Bac						✓	✓	✓			✓			
Minh Tuan Nguyen	✓	✓		✓	✓				✓	✓		✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The supporting data of this study are openly available at https://www.ncbi.nlm.nih.gov/geo/ and https://www.ebi.ac.uk/biostudies/arrayexpress.

REFERENCES

[1] S. Lin *et al.*, "Multiple datasets to explore the molecular mechanism of sepsis," *BMC Genomic Data*, vol. 23, pp. 1–13, 2022, doi: 10.1186/s12863-022-01078-2.

[2] L.-W. Duan *et al.*, "Effects of viral infection and microbial diversity on patients with sepsis: A retrospective study based on metagenomic next-generation sequencing," *World Journal of Emergency Medicine*, vol. 12, pp. 29–35, 2021, doi: 10.5847/wjem.j.1920-8642.2021.01.005.

[3] L. La Via *et al.*, "The global burden of sepsis and septic shock," *Epidemiologia*, vol. 5, pp. 456–478, 2024, doi: 10.3390/epidemiologia5030032.

[4] Z. Li *et al.*, "Immune-associated molecular classification and prognosis signature of sepsis," *PLOS One*, vol. 20, pp. 1–24, 2025, doi: 10.1371/journal.pone.0326083.
[5] A. Yang *et al.*, "Time to treatment and mortality for clinical sepsis subtypes," *Critical Care*, vol. 27, pp. 1–10, 2023, doi: 10.1186/s13054-023-04507-5.
[6] Y. Yang *et al.*, "A robust and generalizable immune-related signature for sepsis diagnostics," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, pp. 246–3254, 2021, doi: 10.1109/TCBB.2021.3107874.
[7] Z.-H. Chen *et al.*, "A signature of immune-related genes correlating with clinical prognosis and immune microenvironment in sepsis," *BMC Bioinformatics*, vol. 24, pp. 1–19, 2023, doi: 10.1186/s12859-023-05134-1.
[8] Y. Peng *et al.*, "An immune-related gene signature predicts the 28-day mortality in patients with sepsis," *Frontiers in Immunology*, vol. 14, pp. 1–14, 2023, doi: 10.3389/fimmu.2023.1152117.
[9] A. Yaqoob *et al.*, "Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm," *Journal of Medical Systems*, vol. 48, pp. 1–10, 2024, doi: 10.1007/s10916-023-02031-1.
[10] H. Wang *et al.*, "Combining machine learning and single-cell sequencing to identify key immune genes in sepsis," *Scientific Reports*, vol. 15, pp. 1–17, 2025, doi: 10.1038/s41598-025-85799-1.
[11] L. Malic *et al.*, "A machine learning and centrifugal microfluidics platform for bedside prediction of sepsis," *Nature Communications*, vol. 16, pp. 1–13, 2025, doi: 10.1038/s41467-025-59227-x.
[12] W. Xiong *et al.*, "Advancing sepsis diagnosis and immunotherapy: Machine learning driven identification of stable molecular biomarkers and therapeutic targets," *Scientific Reports*, vol. 15, pp. 1–17, 2025, doi: 10.1038/s41598-025-93010-8.
[13] Y. Fan *et al.*, "Revealing potential diagnostic gene biomarkers of septic shock based on machine learning analysis," *BMC Infectious Diseases*, vol. 22, pp. 1–16, 2022, doi: 10.1186/s12879-022-07056-4.
[14] S. Lin *et al.*, "Identification of m5C-related gene diagnostic biomarkers for sepsis: A machine learning study," *Frontiers in Genetics*, vol. 15, pp. 1–14, 2024, doi: 10.3389/fgene.2024.1444003.
[15] Z. Jiang *et al.*, "Bioinformatic analysis and machine learning methods in neonatal sepsis: Identification of biomarkers and immune infiltration," *Biomedicines*, vol. 11, pp. 1–14, 2023, doi: 10.3390/biomedicines11071853.
[16] J. Xu *et al.*, "Machine learning screening and validation of panoptosis-related gene signatures in sepsis," *Journal of Inflammation Research*, vol. 17, pp. 4765–4780, 2024, doi: 10.2147/JIR.S461809.
[17] W. Zhang *et al.*, "A diagnostic model for sepsis using an integrated machine learning framework approach and its therapeutic drug discovery," *BMC Infectious Diseases*, vol. 25, pp. 1–13, 2025, doi: 10.1186/s12879-025-10616-z.
[18] B. Gérard, "Analysis of a random forests model," *Journal of Machine Learning Research*, vol. 13, pp. 1063–1095, 2012.
[19] M. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.
[20] L. Prechelt, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.
[21] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, 2012, pp. 37–45, doi: 10.1007/978-3-642-24797-2_4.
[22] C. Singer *et al.*, "The third international consensus definitions for sepsis and septic shock (Sepsis-3)," *Journal of the American Medical Association*, vol. 315, no. 8, pp. 801–810, 2016.
[23] A. Riahi *et al.*, "Exploring the potentials of artificial intelligence in sepsis management in the intensive care unit," *Critical Care Research and Practice*, vol. 2025, no. 1, p. 9031137, 2025.
[24] R. López-Izquierdo *et al.*, "Role of qSOFA and SOFA scoring systems for predicting in-hospital risk of deterioration in the emergency department," *International Journal of Environmental Research and Public Health*, vol. 17, no. 22, p. 8367, 2020.
[25] C. F. Duncan *et al.*, "Diagnostic challenges in sepsis," *Current Infectious Disease Reports*, vol. 23, no. 12, p. 22, 2021.
[26] A. Feng *et al.*, "Free radical–associated gene signature predicts survival in sepsis patients," *International Journal of Molecular Sciences*, vol. 25, no. 8, p. 4574, 2024.

BIOGRAPHIES OF AUTHORS

Tuan Anh Vu received the B.Sc. degree of engineer in information technology and the M.Sc. degree in information systems from the Post and Telecommunications Institute of Technology (PTIT), Hanoi, VietNam, in 2016 and 2018. He is currently a Ph.D. candidate in PTIT with research interests including machine learning, deep learning, optimization, and bigdata. He can be contacted at email: vtanh@ptit.edu.vn.

Dang Hoai Bac archived the B.Sc. degree in automation engineering from Hanoi University of Technology and Science, Hanoi, Vietnam in 1997, the M.Sc. degree in electronics & communications engineering an the Ph.D. degree in electronics & communications engineering from Posts and Telecommunications Institute of Technology (PTIT), in 2003 and 2008, respectively. He is now the director of PTIT with research interests of digital signal processing, telecommunication transmission, NGN technologies, satellite system, machine learning, deep learning, biomedical application designs, gene expression optimization. He can be contacted at email: bacdh@ptit.edu.vn.

Minh Tuan Nguyen obtained the B.Sc. and M.Sc. degrees of electronics and telecommunications engineering from the Post and Telecommunications Institute of Technology (PTIT) and Hanoi University of Science and Technology (HUST), Hanoi, Vietnam, in 2004 and 2008, respectively. In 2018, he was awarded a Ph.D. degree of Electrical engineering and computer science from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. He has worked as researcher at PTIT since 2021 with research interests of network security, internet of things, biomedical signal processing, gene analysis, sentiment analysis, brain computer interface, machine learning, deep learning, optimization, and biomedical application design. He can be contacted at email: nmtuan@ptit.edu.vn.