Efficient Machine Learning-based Gene Selection Exploiting Immune-related Biomarkers and Recursive Feature Elimination for Sepsis Diagnosis

Duc-Long Vu
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
longvd@ptit.edu.vn

Hai-Chau Le
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
chaulh@ptit.edu.vn

ABSTRACT

Differential expression gene (DEG) analysis of transcriptomic data allows for a comprehensive examination of the regulation in gene expression profiles related to specific biological states. The result of this analysis typically consists of an extensive record of genes that display varying levels of expression among two or more groups. A portion of these genes with altered expression could potentially function as candidate biomarkers, chosen through either existing biological insights or data-driven techniques. In diagnosing sepsis, a life-threatening health problem, our work proposes a novel approach using immune-related gene data to identify the optimal gene combination as signature biomarkers to improve the diagnosis performance. Our proposed method involves sequential gene selection procedures, including the DEG analysis and the machine learning-based importance assessment, and a Recursive Feature Elimination (RFE) process supported by Principal Component Analysis (PCA). The selected gene combination, which consists of twelve immune-related genes, shows remarkable cross-validation results with an accuracy of 99.35%, AUC score of 99.56%, Sensitivity and a Specificity of 99.44% and 90.00%, respectively. Besides, the proposed 12 gene markers combined with the XGBoost algorithm were also tested in three individual cohorts with appropriately significant results, demonstrating the effectiveness of our developed method in different cohorts and the reliability of the proposed gene selection procedure.

CCS CONCEPTS

• **Computing methodologies** → **Boosting**; **Feature selection**; **Cross-validation**.

KEYWORDS

Machine Learning, Sepsis, Immune-Related Gene, Feature Selection, Recursive Feature Elimination

SOICT 2023, December 7–8, 2023, Ho Chi Minh, Vietnam

ACM ISBN 979-8-4007-0891-6/23/12…$15.00
https://doi.org/10.1145/3628797.3628989

ACM Reference Format:
Duc-Long Vu and Hai-Chau Le. 2023. Efficient Machine Learning-based Gene Selection Exploiting Immune-related Biomarkers and Recursive Feature Elimination for Sepsis Diagnosis. In *The 12th International Symposium on Information and Communication Technology (SOICT 2023), December 7–8, 2023, Ho Chi Minh, Vietnam.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3628797.3628989

1 INTRODUCTION

Sepsis is the condition where the body responds to infection in human tissues and organs with severe inflammation. This disease has evolved into a major global problem in human health due to a significant increase in both incidence and mortality rates in recent years [3, 11]. Sepsis is categorized into three types based on the levels of severity including sepsis, severe sepsis, and septic shock, in which sepsis shock is renowned for being the most intricate and life-threatening condition due to the dysfunction syndromes in multiple patient organs [17]. Similar to some popularly dangerous diseases like trauma, stroke, or sudden cardiac arrest, sepsis is a condition where timing is critical and hence, it is crucial to promptly recognize and accurately predict the sepsis in order to enhance the efficiency of sepsis diagnosis and ensure timely and effective treatment.

Conventional methods for sepsis diagnosis typically entail utilizing microbiological culture techniques to identify and classify the pathogen. These methods still have various limitations; they are time-consuming and may not yield positive results quickly, and lead to false negative outcomes [9, 10]. Moreover, another approach involves the use of physiological scoring tools within intensive care units (ICUs) primarily relying on clinical and laboratory data to assess the severity of critical illness [6, 15]. Unfortunately, those tools provide limited insight into the likelihood of a negative outcome, i.e. mortality, at the early stages of the disease. Additionally, differential expression gene analysis is a computational and analytical technique used to identify the genes that show significant differences in expression between two different phenotypes. By applying statistical methods, such as associating a p-value threshold with each gene in the two groups, DEG analysis determines which genes exhibit the most significant differences in expression. Because the number of identified genes is extremely high and many correlated biomarkers may still be available, it makes the processes of the diagnosis and prognosis procedure become exceedingly exhaustive and need a lot of professional knowledge to construct an effective prediction model.

On the other hand, in order to develop effective sepsis prediction models, several methods have been introduced to refine the outcomes of the statistical analysis for differential expression by taking

advantage of data analysis and machine learning (ML) models for reducing the number of biomarkers required [1, 7, 16, 22]. In [16], the authors apply the two stages of gene number reduction using DEGs and sequential forward features selection-based Boosting algorithm to determine a group of nine genes that enable a remarkably precise mortality prediction in sepsis. Likewise, a 10-gene subset has been pinpointed among a pool of differentially expressed genes observed in pediatric patients suffering from sepsis and sepsis shock. [1]. These genes were ranked again using the MRMR score obtained from the maximum relevance minimum redundant algorithm, then utilizing a repetitive cross-validation process for feature selection. Besides, an ML algorithm was employed to validate the selected gene set in the diagnosis of pediatric sepsis patient mortality. Furthermore, an additional set of 18 genes was spotted as diagnostic markers using a sequential filter method based on LASSO, RFE, and Random Forest Variable Hunting feature selection algorithms [22]. In the work of [7], an ensemble of 8 genes was selected using the LASSO feature selection algorithm, then the Random Forest algorithm was used to build a sepsis diagnosis model. Additionally, in [21], a novel approach called Recurrent Logistic Regression (RLR) identified a limited subset of just 5 genes that are notably associated with the immune system. Clearly, machine learning algorithms are instrumental in the process of analyzing genetic data, and methods based on machine learning uncover valuable insights that are instrumental in developing models for sepsis diagnosis and prognosis, as well as in dissecting the biology of sepsis. Nevertheless, to the best of the author's knowledge and comprehensive literature review result, current research on the sepsis diagnosis utilizing gene expression data still lacks the ability of various data adaption; most research only effectively deals with a very limited number of datasets and consequently, it is not generalizable and may be not efficiently applicable for other datasets. Moreover, the selected features (genes) always show a lot of correlation together, which would make the prediction model tend to be overfitting when training and receive not very good results. Besides, most of the introduced biomarkers do not have appropriate verification.

To resolve the above-mentioned issues, in this research work, we propose an efficient gene selection method that applies data mining that leverages machine learning techniques to specify a small number of genes with much information to predict sepsis effectively. Due to the severe dysregulation of the immune system in sepsis patients, [8, 13], our work's primary focus has been on the signature immune-related gene markers. Moreover, our approach to gene selection involves a sequential procedure comprising three phases of data dimension reduction. These stages encompass the filter of genes associated with the immune system, the analysis of differential expression, and the utilization of principal component analysis (PCA) to aid in recursive feature elimination (RFE) through data transformation. Our procedure generates a subset of genes that contains the highest quantity of information relevant to sepsis. Additionally, we utilize cross-validation in conjunction with several ML algorithms for evaluating the performance of the chosen gene combination for sepsis diagnosis. Furthermore, we validate the identified biomarkers by testing them on three distinct sets of gene expressions, each obtained from a different *microarray* platform. Our main contributions in this work are listed as follows:

- Proposing a machine learning-based gene selection procedure that exploits immune-related biomarkers and recursive feature elimination based on PCA for sepsis diagnosis.
- Developing an extremely accurate prediction model and evaluating the effectiveness of the proposed biomarkers in notable platforms.

The rest of this paper is presented as follows: In Section II, we introduce the typical datasets and outline the preprocessing methods employed in our study. Next, in Section III, we delve into the sequential gene analysis and advanced gene selection methods. We present the experimental results and engage in discussions in Sections IV and V, respectively. Finally, in Section VI, we provide a concise summary and draw our conclusions.

2 DATA AND PREPROCESSING

2.1 Genomics Data

Our study utilizes seven public genome datasets downloaded from the Gene Expression Omnibus (GEO) public database. These datasets comprise both sepsis and healthy samples, with a breakdown of four datasets for adults [4, 12, 14, 20] and three datasets for children [2, 18, 19]. There is a total of 1164 samples have been collected and available across three microarray platforms. All the samples have undergone preprocessing and normalization by RMA algorithms. The gene expression levels were determined by calculating the mean of the genes associated with multiple probe positions based on the probe mapping table obtained as a SOFT file from the corresponding GEO dataset. The discovery cohort consists of four datasets: GSE57065 [4], GSE26378 [19], GSE95233 [14], and GSE26440 [18], to identify the most informative genes and train the diagnostic model for sepsis. Three datasets, including GSE4607 [20], E-MTAB1548 [2], and GSE65682 [12], are designated as validation cohorts and used to assess the capability of the proposed gene combination and evaluate the capability for working in multiple platforms of selected biomarkers.

2.2 Immune-related Gene Analysis

We acquired approximately 770 genes that related significantly to the immune system from the nanoString database, a widely utilized resource in numerous studies involving pathogen infection and host response. There are three main *microarray* platforms are used in this work. For the Affymetrix Human Genome U133 Plus 2.0 (AffyU133P2) and Affymetrix Human Genome U219 (AffyU219) platforms, we compiled IRG counts of 737 and 740, respectively. In the case of the Agilent Human Gene Expression 4x44K v2 Microarray (AgilentV2) platform, we collected a count of 627 IRGs. To ensure the identification of cross-platform applicable biomarkers, our focus centered on the 608 IRGs that were common to all three systems. These shared IRGs formed the basis for our computational modeling, as illustrated in the accompanying Figure.1.

2.3 Differential Expression Analysis

In this research project, we utilized the differential gene expression analysis procedure to pinpoint immune-related genes exhibiting the most notable differential expression. To execute the differential expression analysis, we harnessed the R package as a simulation

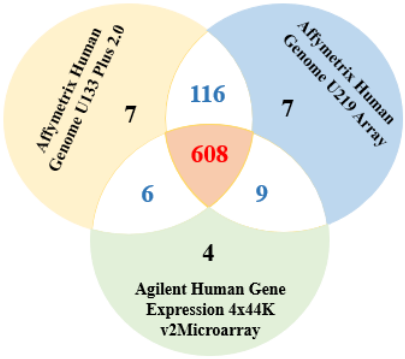

Figure 1: Overlapping immune-related genes of in three platforms

tool. Specifically, the limma R package is obtained along with the Benjamini-Hochberg (BH) correction method to identify the genes that are significantly different from others. We estimated the fold change, which signifies the ratio of gene expression in sepsis samples when compared to normal samples. In the process of selecting differentially expressed immune-related genes (DEIRGs) that accurately represent sepsis patients and normal individuals, we took into account the *p-values* and *log-fold* change components. These criteria enabled us to pinpoint the DEIRGs that display substantial alterations in expression between sepsis patients and normal samples.

3 METHODOLOGY

Our proposed method includes three procedures for reducing the high data dimension of the gene expression data shown in Fig. 2. The first process is gene preprocessing, we investigate the immune-related gene and perform an analysis of differential expression genes. Then, the immune gene selection procedure strives to reduce the number of genes used for the diagnosis of sepsis, in this section, only the immune-related genes are considered. The final stage is to estimate the proposed optimal gene set in the validation cohorts corresponding with various ML algorithms.

3.1 Machine Learning Algorithm

In this study, to estimate the performance of gene combinations, we use three machine learning models, including Extreme Gradient Boosting (XGB), Random Forest (RF), and K-Nearest Neighbors (KNN) [5], to perform the gene score estimation task and evaluate individual gene combinations generated by the Recursive Feature Elimination algorithm. The intelligent diagnostic model is also constructed using the optimal subsets and tested in individual gene expression datasets.

3.2 Gene Importance Score Estimation

In this stage, the differential expressed immune-related genes results are investigated to rank and select based on their absolute log fold-change values. Particularly, the gene with a high expression between two types of the sample shows a significant gap; hence it is considered the most informative gene related to sepsis. Next,

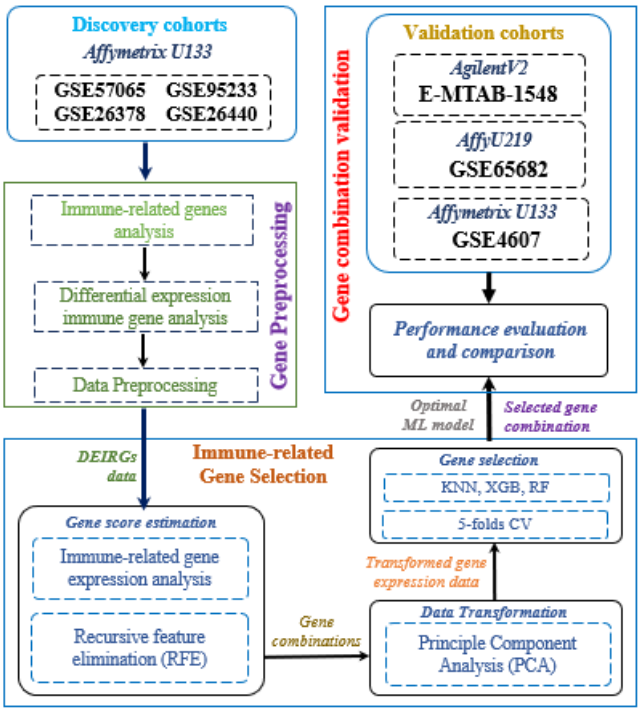

Figure 2: Flow method

Algorithm 1 Gene Ranking Based on RFE

1) Initialization: n genes; S is set of n genes
2) Feature Importance Analysis
Repeat
a) Divide Discovery Cohorts into k folds of data set F(i)
b) Initial XGBoost Model
c) Estimate Gene Importance
for i = 1 to k

- Training the XGBoost model with $F(j)$, $j \neq i$.
- Calculating the importance value of each gene

end
- Estimate the average importance value for each gene
d) Ranking genes based on their importance score
e) Eliminate the gene with the lowest importance score
f) Saving a subset of n - 1 genes S(n-1)
n = n - 1
Until n = 1
3) Ranking genes based on the number of selection times
4) Collection of n gene subsets

We employed the Recursive Feature Elimination (RFE) method to construct multiple subsets of genes, which were subsequently transformed into the component space through Principal Component Analysis (PCA). Initially, the input genes are ranked based on their important values, which are determined using the XGB model along with the cross-validation procedure. The gene subsets were created by iteratively eliminating the gene with the lowest average importance value. Importantly, we applied the XGBoost (XGB) model

and cross-validation (CV) repeatedly to calculate the gene importance for each gene subset until no genes were left for removal. The detailed RFE algorithm for generating the gene combinations is illustrated in Algorithm 1.

3.3 Data Transformation

Various of gene subsets generated by the RFE algorithms and the cross-validation procedure, are then transformed into principal component space by PCA, where each subset corresponds to a TGC. The primary motivation for using PCA transformation is the presence of strong correlations among features within the original differential immune-related gene subset. By employing PCA, the resulting component subset consists of individuals with a significantly reduced correlation between genes, therefore improving the performance of a selection of the most informative genes and consequently improving the total performance of the diagnosis model. It's important to highlight that throughout the transformation process, an equivalent number of genes (i.e., data dimensions) in a subset are retained as components, preserving the entirety of the information.

3.4 Machine Learning-based Gene Selection

All the gene subsets are then fed into different ML algorithms to explore the optimal model corresponding with the gene combinations on the discovery set. Clearly, it is crucial to perform hyperparameter tuning in order to identify the best models, as this plays a vital role in mitigating the issue of overfitting. Furthermore, we assessed the performance of the chosen models linked to the Top Gene Combinations (TGCs) by evaluating their classification accuracy on the validation set. In our study, we utilized a combination of random search with grid search and the fivefold Cross-Validation (CV) method to identify the optimal parameter values for the models across the complete set of TGCs.

3.5 Gene Combination Validation

The optimal gene subsets, which are considered the output of the gene selection phase, are estimated for the diagnosis performance on the validation set. Different ML models are then retrained using the discovery data and tested on the validation cohorts using optimal gene combinations. The gene combination corresponding with the ML model performs the best result and is considered the most effective model as well as the optimal gene set to predict sepsis. Moreover, the cross-platform capability of the proposed model was demonstrated by evaluating two different platform cohorts GSE65682 and E-MTAB-1548.

4 EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Performance Metrics of Classification

Several evaluation metrics including Accuracy (Acc), Sensitivity (Sn), Specificity (Sp), Area Under the Curve (AUC), and Balance Error Rate (BER), are considered in our work to estimate the efficiency of both the ML algorithm and the combinations of immune-related genes. Accuracy and AUC assess the precision of the classification model, while Sensitivity and Specificity gauge the model's capability to diagnose sepsis and normal cases, respectively. The BER score, calculated using the formula 1 - 0.5*(Sn + Sp), represents the average error rate across the two classification categories.

4.2 Identification of Differential Expressed Immune-related Genes

Since our primary research emphasis revolved around immune-related genes (IRGs), recognizing their critical involvement in the molecular mechanisms underlying sepsis. By evaluating the intersection of multiple platforms, 608 IRGs hold potential as biomarkers for future investigations. A wise-known sequential analysis is applied to identify the differential expressed genes, including 127 genes that have absolute log-Fold Change (aFC) $\geq$ 1.75 and the p-value $\leq$ 0.05. Table 1 shows the ranking result using the absolute fold-change values of 10 DEIRGs. Specifically, up-regulated genes in sepsis samples exhibit higher expression levels compared to the same genes in normal samples. The fold change indicated the intensity of expression value variance between the two classes.

Table 1: Top 10 DEIRGs ranked top-down based on the absolute fold change values

Idx	GID	logFC	p-value	Regulation
1	IL1R2	3.72	2.95E-96	Up
2	LCN2	3.58	6.84E-56	Up
3	LTF	3.41	5.54E-54	Up
4	CEACAM8	3.3	4.06E-43	Up
5	S100A12	3.1	1.50E-110	Up
6	CLEC5A	2.98	4.89E-61	Up
7	IL18R1	2.91	5.96E-71	Up
8	KLRF1	2.55	1.22E-56	Down
9	GNLY	2.46	3.96E-53	Down
10	LRRN3	2.39	3.29E-28	Down

4.3 Gene Ranking

There are 127 genes that are the result of the gene expression analysis procedure and are considered as the input of this process. Particularly, the feature selection method RFE is applied to ranking individual genes using the importance score. Additionally, the XGB algorithm with the cross-validation method, which is implemented in the RFE procedure, is applied 10 times to calculate the mean importance value of each gene. After that, depending on the average importance values, the gene with the lowest importance value is eliminated. It is noteworthy that the importance score considered in our work can be known as the number of selection times that the gene is kept to construct the classifier by the RFE method. The final genes ranking table is obtained by ranking in decreasing order the selected time value for each gene.

4.4 Gene Validation

The gene ranking outcomes are utilized to establish the quantity of m genes to pick. This is achieved by employing any classifier and supplying it with a subset of genes that have been ranked, commencing with the highest-ranked gene and progressing downward. The objective is to retain the m genes that yield the greatest

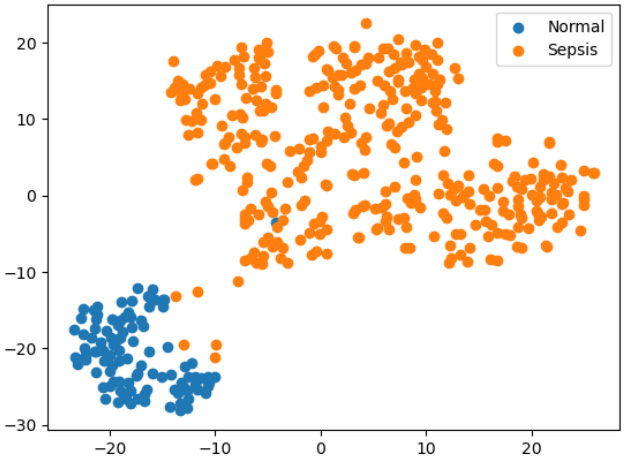

Figure 3: Visualization of TGC sets of 12 genes using tSNE

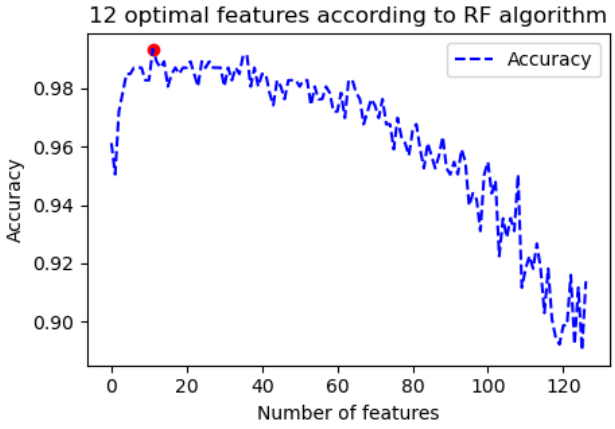

Figure 4: 12 features according to RF algorithms

classification score. Hence, there are 127 gene combinations are estimated to figure out the most effective gene group. It is worth noting that there are 3 ML algorithms considered to perform this estimation procedure and 127 gene subsets corresponding to 127 transformed gene combinations (TGCs), which applied the PCA method to convert into different data spaces. Fig.3 illustrates the gene expression data that was transformed into different spaces using PCA. Obviously, by applying the PCA transform, the gene expression data in the new data space perform remarkable clustering performance that could help develop the diagnosis model more effectively, which is demonstrated in the following experiments.

There are a total of 381 models trained using TGCs, which are then evaluated for their diagnosis efficiency on the discovery cohort by the 5-fold CV procedure. Table 2 illustrates the performance of three ML models on the discovery cohort. The RF algorithm produces the highest classification accuracy corresponding with the optimal gene combinations. Notably, each model in Table 2 was selected by a comparison of another classification accuracy among 127 models using TGCs. Clearly, the average results for the ideal algorithms align with the efficient combination of immune genes generated by the RF classifier, which specifically incorporates 12 genes associated with the immune system. This is presented in detail in Figure 4. As a result, we designate the sets of these 12 genes as the most informative gene subsets related to sepsis.

4.5 Diagnostics Performance Estimation

To demonstrate the effectiveness of the minimized gene groups with 12 candidates, we investigate the performance of this combination on validation cohorts. Three machine learning models, namely RF, XGB, and KNN, were trained on the discovery cohort. Subsequently, their performance was assessed using three distinct gene sets. Additionally, we also demonstrate the effectiveness of our proposed gene combination worked well across three different *micro-array* platforms. Table 3 depicts the detail of the validation results of three ML models on validation cohorts. Obviously, the performance of the XGB model for classification sepsis on the validation procedure shows the most elevated score compared with other ML algorithms in terms of AUC score, Accuracy, Sn, and Sp. As a result, we propose the XGB model with the genes set related to the immune system, including S100A12, TLR5, IL1R2, MAPK14, FCER1A, FCER1G, LCN2, RUNX3, C3AR1, IL18R1, EOMES, KLRF1, as the most useful gene subset and machine learning algorithm for diagnosis of sepsis.

4.6 Discussion

Differential gene expression analysis is a commonly utilized technique to investigate gene expression patterns and unveil the intricate biological mechanisms associated with complex diseases. Typically, gene expression profiles exhibit a large number of genes and significant correlations between them. Consequently, when conducting differential expression analysis, the results often yield hundreds of genes that are highly correlated. However, it is not feasible to utilize all of these differentially expressed genes when developing diagnostic and prognostic prediction tools. Instead, researchers commonly employ either domain knowledge or data-driven methods to identify a smaller set of marker genes, known as a gene signature, for such purposes. In our work, we introduced a new data-driven approach for prioritizing marker genes in the context of predicting sepsis. Specifically, we employed an instance of the RFE algorithm with the XGB model to rank the genes based on their important values followed by a sequential gene combination generation procedure. The originality of our research stems from our integration of feature selection methods into the statistical framework used to conduct DEG analysis. Additionally, we introduce a novel importance score, which is based on the frequency with which a gene is chosen as a significant biomarker by the Recursive Feature Elimination (RFE) algorithm.

Furthermore, in this work, we used the immune-related genes were used as the foundation for the model. Due to the dysregulated of the host immune system to infection in sepsis, the immune-related genes represent prior knowledge for the diagnosis model and prevent overfitting, resulting in a robust model for patient heterogeneity. From the original gene feature set, a different gene combination will be selected with unexpected noise. Hence, if we start constructing a diagnosis model with all genes, the model may

Table 2: Feature Validation Results

Model	# of genes	Acc (%)	Sn (%)	Sp (%)	AUC (%)	BER (%)
RF	***12***	***99.35 ± 0.52***	***99.44 ± 0.67***	***99.00 ± 0.20***	***99.56 ± 0.84***	***0.77 ± 0.91***
KNN	12	99.13 ± 0.43	99.17 ± 0.67	99.00 ± 0.20	99.30 ± 0.97	0.91 ± 0.83
XGB	61	99.13 ± 0.43	98.90 ± 0.54	100.00 ± 0.00	99.94 ± 0.06	0.54 ± 0.27

Table 3: Performance Estimation of Proposed Model on Three Independent Datasets

Data	ML	Acc%	Sn%	Sp%	AUC%	BER%
GSE4607	RF	95.23	95.65	93.33	94.29	5.50
	KNN	94.04	94.20	93.33	96.08	6.23
	XGB	**96.42**	**97.10**	**93.33**	**96.71**	**4.78**
GSE65682	RF	96.92	96.65	100.00	98.68	1.67
	KNN	95.39	94.98	100.00	98.85	2.50
	XGB	**97.50**	**97.28**	**100.00**	**99.12**	**1.35**
EMTAB-1458	RF	97.89	97.50	100.00	99.46	1.25
	KNN	97.89	97.50	100.00	99.37	1.25
	XGB	**97.89**	**97.50**	**100.00**	**99.67**	**1.25**

get extremely high performance for the training cohort but worse when it comes to an evaluation in the testing set and validation cohorts. Moreover, with a significant number of genes, the training and process procedure will consume computational resources and time. Additionally, it is important to highlight that supervised machine learning methods, in conjunction with feature selection algorithms, can be directly applied to pinpoint distinct gene markers from extensive transcriptomic datasets. However, this particular method has significant drawbacks with the computational time required can be exceptionally long due to certain feature selection methods.

Our DE approach and machine learning analysis aided by immune-related genes suggested a combination including 12-gene markers for diagnostic sepsis with an average AUC score of 99.56 % and Accuracy of 99.35 %. It is interesting that our approach includes three reliable procedures for data dimension reduction and also for analyzing the principle component of gene expression using PCA, including expression analysis and feature selection. The immune-related genes obtained from the filtered method based on the immune gene data, play a vital role in reducing the number of genes that contributed significantly to the diagnosis of sepsis. The DEG procedure is used to eliminate the normal genes that do not regulate significantly between two classes and reveal potential biomarkers for the next analysis process using machine learning. The machine learning technique based on RFE is employed to gauge the efficacy of gene combinations. This approach aims to optimize classification performance while minimizing the complexity of the diagnostic model. Consequently, several machine learning models are assessed using 127 gene combinations generated by specialized algorithms for gene combination generation. Ultimately, a final subset of 12 genes is identified, enhancing the diagnostic performance of the proposed algorithm. The selection of a very limited number of genes through this process underscores the practicality and effectiveness of the suggested algorithm in a clinical setting.

5 CONCLUSION

In this research, we have explored the diagnostic challenge of sepsis, a critical global health concern, using machine learning. Our approach introduces an effective gene selection method that combines machine learning techniques with a focus on immune-related genes. A signature of 12 marker genes related to the immune system has been identified for enhancing the performance of the sepsis diagnosis. These 12 genes were identified using a data-driven machine-learning approach for feature decomposition. They were selected as the optimal subset from a pool of 127 DEIGRs combinations, which were identified through a statistical analysis using seven publicly genome datasets related to sepsis. The selected gene combination combined with the XGB algorithms shows significant accuracy in diagnostics and performs robustness on validation cohorts. Furthermore, we conducted validation across three distinct gene expression platforms to showcase the effectiveness of both our sequential gene selection method and the intelligent prognosis model associated with the proposed biomarkers. The numerical validation outcomes, boasting an accuracy of 99.35%, a sensitivity of 99.44%, and a specificity of 90.00%, underscore the potential applicability of our developed approach in a real clinical setting.

REFERENCES

[1] Mostafa Abbas and Yasser EL-Manzalawy. 2020. Machine learning based refined differential gene expression analysis of pediatric sepsis. *BMC Medical Genomics* 13 (12 2020), 122. Issue 1. https://doi.org/10.1186/s12920-020-00771-4

[2] Raquel Almansa, Eduardo Tamayo, María Heredia, Sandra Gutierrez, Patricia Ruiz, Elisa Alvarez, Esther Gomez-Sanchez, David Andaluz-Ojeda, Rafael Ceña, Lucia Rico, et al. 2014. Transcriptomic evidence of impaired immunoglobulin G production in fatal septic shock. *Journal of critical care* 29, 2 (2014), 307–309.

[3] Derek C. Angus and Tom van der Poll. 2013. Severe Sepsis and Septic Shock. *New England Journal of Medicine* 369 (8 2013), 840–851. Issue 9. https://doi.org/10.1056/NEJMra1208623

[4] Marie-Angélique Cazalis, Alain Lepape, Fabienne Venet, Florence Frager, Bruno Mougin, Hélène Vallin, Malick Paye, Alexandre Pachot, and Guillaume Monneret. 2014. Early and dynamic changes in gene expression in septic shock patients: a genome-wide approach. *Intensive care medicine experimental* 2, 1 (2014), 1–17.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

[6] William A. Knaus, Douglas P. Wagner, Elizabeth A. Draper, Jack E. Zimmerman, Marilyn Bergner, Paulo G. Bastos, Carl A. Sirio, Donald J. Murphy, Ted Lotring, Anne Damiano, and Frank E. Harrell. 1991. The APACHE III Prognostic System. *Chest* 100 (12 1991), 1619–1636. Issue 6. https://doi.org/10.1378/chest.100.6.1619

[7] Jianhai Lu, Rui Chen, Yangpeng Ou, Qianhua Jiang, Liping Wang, Genglong Liu, Yayun Liu, Ben Yang, Zhujiang Zhou, Liuer Zuo, and Zhen Chen. 2022. Characterization of immune-related genes and immune infiltration features for early diagnosis, prognosis and recognition of immunosuppression in sepsis. *International Immunopharmacology* 107 (6 2022), 108650. https://doi.org/10.1016/j.intimp.2022.108650

[8] Leo McHugh, Therese A. Seldon, Roslyn A. Brandon, James T. Kirk, Antony Rapisarda, Allison J. Sutherland, Jeffrey J. Presneill, Deon J. Venter, Jeffrey Lipman, Mervyn R. Thomas, Peter M. C. Klein Klouwenberg, Lonneke van Vught, Brendon Scicluna, Marc Bonten, Olaf L. Cremer, Marcus J. Schultz, Tom van der Poll, Thomas D. Yager, and Richard B. Brandon. 2015. A Molecular Host Response Assay to Discriminate Between Sepsis and Infection-Negative Systemic Inflammation in Critically Ill Patients: Discovery and Validation in Independent Cohorts. *PLOS Medicine* 12 (12 2015), e1001916. Issue 12. https://doi.org/10.1371/journal.pmed.1001916

[9] Ithan D. Peltan, Samuel M. Brown, Joseph R. Bledsoe, Jeffrey Sorensen, Matthew H. Samore, Todd L. Allen, and Catherine L. Hough. 2019. ED Door-to-Antibiotic Time and Long-term Mortality in Sepsis. *Chest* 155 (5 2019), 938–946. Issue 5. https://doi.org/10.1016/j.chest.2019.02.008

[10] Ithan D. Peltan, Kristina H. Mitchell, Kristina E. Rudd, Blake A. Mann, David J. Carlbom, Catherine L. Hough, Thomas D. Rea, and Samuel M. Brown. 2017. Physician Variation in Time to Antimicrobial Treatment for Septic Patients Presenting to the Emergency Department. *Critical Care Medicine* 45 (6 2017), 1011–1018. Issue 6. https://doi.org/10.1097/CCM.0000000000002436

[11] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, Carolin Fleischmann-Struzek, Flavia R Machado, Konrad K Reinhart, Kathryn Rowan, Christopher W Seymour, R Scott Watson, T Eoin West, Fatima Marinho, Simon I Hay, Rafael Lozano, Alan D Lopez, Derek C Angus, Christopher J L Murray, and Mohsen Naghavi. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* 395 (1 2020), 200–211. Issue 10219. https://doi.org/10.1016/S0140-6736(19)32989-7

[12] Brendon P Scicluna, Peter MC Klein Klouwenberg, Lonneke A van Vught, Maryse A Wiewel, David SY Ong, Aeilko H Zwinderman, Marek Franitza, Mohammad R Toliat, Peter Nürnberg, Arie J Hoogendijk, et al. 2015. A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission. *American journal of respiratory and critical care medicine* 192, 7 (2015), 826–835.

[13] Brendon P. Scicluna, Maryse A. Wiewel, Lonneke A. van Vught, Arie J. Hoogendijk, Augustijn M. Klarenbeek, Marek Franitza, Mohammad R. Toliat, Peter Nürnberg, Janneke Horn, Marc J. Bonten, Marcus J. Schultz, Olaf L. Cremer, and Tom van der Poll. 2018. Molecular Biomarker to Assist in Diagnosing Abdominal Sepsis upon ICU Admission. *American Journal of Respiratory and Critical Care Medicine* 197 (4 2018), 1070–1073. Issue 8. https://doi.org/10.1164/rccm.201707-1339LE

[14] Olivier Tabone, Marine Mommert, Camille Jourdan, Elisabeth Cerrato, Matthieu Legrand, Alain Lepape, Bernard Allaouchiche, Thomas Rimmelé, Alexandre Pachot, Guillaume Monneret, et al. 2019. Endogenous retroviruses transcriptional modulation after severe infection, trauma and burn. *Frontiers in immunology* 9 (2019), 3091.

[15] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. 1996. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine* 22 (7 1996), 707–710. Issue 7. https://doi.org/10.1007/BF01709751

[16] Long Duc Vu, Van Su Pham, Minh Tuan Nguyen, and Hai-Chau Le. 2022. Pediatric Sepsis Diagnosis Based on Differential Gene Expression and Machine Learning Method. *2022 14th International Conference on Knowledge and Systems Engineering (KSE)* (10 2022), 1–6. https://doi.org/10.1109/KSE56063.2022.9953619

[17] Scott L. Weiss, Julie C. Fitzgerald, John Pappachan, Derek Wheeler, Juan C. Jaramillo-Bustamante, Asma Salloo, Sunit C. Singhi, Simon Erickson, Jason A. Roy, Jenny L. Bush, Vinay M. Nadkarni, and Neal J. Thomas. 2015. Global Epidemiology of Pediatric Severe Sepsis: The Sepsis Prevalence, Outcomes, and Therapies Study. *American Journal of Respiratory and Critical Care Medicine* 191 (5 2015), 1147–1157. Issue 10. https://doi.org/10.1164/rccm.201412-2323OC

[18] Hector R Wong, Natalie Cvijanovich, Richard Lin, Geoffrey L Allen, Neal J Thomas, Douglas F Willson, Robert J Freishtat, Nick Anas, Keith Meyer, Paul A Checchia, et al. 2009. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC medicine* 7, 1 (2009), 1–12.

[19] Hector R Wong, Natalie Z Cvijanovich, Geoffrey L Allen, Neal J Thomas, Robert J Freishtat, Nick Anas, Keith Meyer, Paul A Checchia, Scott L Weiss, Thomas P Shanley, et al. 2014. Corticosteroids are associated with repression of adaptive immunity gene programs in pediatric septic shock. *American journal of respiratory and critical care medicine* 189, 8 (2014), 940–946.

[20] Hector R Wong, Thomas P Shanley, Bhuvaneswari Sakthivel, Natalie Cvijanovich, Richard Lin, Geoffrey L Allen, Neal J Thomas, Allan Doctor, Meena Kalyanaraman, Nancy M Tofil, et al. 2007. Genome-level expression profiles in pediatric septic shock indicate a role for altered zinc homeostasis in poor outcome. *Physiological genomics* 30, 2 (2007), 146–155.

[21] Yueran Yang, Yu Zhang, Shuai Li, Xubin Zheng, Man Hon Wong, Kwong Sak Leung, and Lixin Cheng. 2021. A robust and generalizable immune-related signature for sepsis diagnostics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021), 1–1. https://doi.org/10.1109/TCBB.2021.3107874

[22] Jianchao Ying, Qian Wang, Teng Xu, and Zhongqiu Lu. 2021. Diagnostic potential of a gradient boosting-based model for detecting pediatric sepsis. *Genomics* 113 (1 2021), 874–883. Issue 1. https://doi.org/10.1016/j.ygeno.2020.10.018