BIOMARKER SELECTION FOR PEDIATRIC SEPSIS DIAGNOSIS USING DEEP LEARNING

Ngoc Anh Thi Phung, Anh Thu Pham, Minh Tuan Nguyen
Posts and Telecommunications Institute of Technology

Abstract—This study proposes a new approach for diagnosing pediatric sepsis that utilizes a convolutional neural network and a combination of 7 immune-related genes (IRGs), including CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3, and CCRL2. A three-layer gene selection process involves a sequential procedure that combines differential gene expression analysis, selection of immune-related genes, and gene score calculation using the F-score algorithm. This process identifies the most informative differentially expressed genes, followed by the utilization of a deep learning model to determine the optimal gene combination. The performance of the proposed algorithm is evaluated using a 3-fold cross-validation procedure with deep learning models. The results show that the selected gene combinations achieve an accuracy of 91.92% and an area under the ROC curve of 87.86%, indicating that the proposed algorithm is reliable for predicting pediatric sepsis mortality. Additionally, the identification of a signature consisting of 7 IRGs associated with pediatric sepsis mortality has the potential to aid in the development of dependable diagnostic and prognostic biomarkers for sepsis.

Keywords— Pediatric sepsis, Differential expression gene, Immune-related genes, Gene selection, Deep learning.

I. INTRODUCTION

Sepsis is characterized by elevated rates of morbidity and mortality, stemming from an imbalanced inflammatory response of the host to infection [1]. Septic shock is a severe form of sepsis in which the blood pressure drops to dangerously low levels, causing organ failure and potentially leading to death [2]. Pediatric sepsis is a critical condition that occurs when the body's immune system responds excessively to an infection, leading to organ failure and a life-threatening condition [3]. It is a major global public health problem and one of the leading causes of death in critically ill children admitted to the intensive care unit (ICU) [4]. Despite the significant progress made in the diagnosis and treatment of sepsis, the number of cases is still increasing [5], [6]. Besides,there have been efforts to stratify the risk of sepsis, particularly in children, it remains a challenge due to the considerable variability among patients and the inadequate definitions of sepsis in pediatric populations that currently exist [7]. This highlights the urgent need to gain a deeper understanding. Ongoing research efforts are necessary to identify more sensitive and specific targets for the diagnosis and treatment of sepsis, particularly pediatric sepsis and septic shock, as the complexities of this condition require a comprehensive approach to ensure effective management and prevention. Early warning and accurate prediction on pediatric sepsis and septic shock provide opportunities for physicians to take preventative measures to alleviate its devastating consequences.

The molecular-level diagnosis and prognostic detection of diseases has become a prevailing trend, and researchers studying sepsis have also widely adopted this approach [8]. Recently advancement of multi-omics sequencing technologies has resulted in an increase in the number of genetic biomarkers available. To better understand the genes, RNA, and proteins involved, researchers are increasingly analyzing and testing single or combined biomarkers. Various strategies have been employed to uncover these biomarkers, such as massspectrometry, protein arrays, and gene-expression profiling. It has also been observed that the development of pediatric septic shock involves the participation of multiple genes and immune system-related pathways [9]. Differential expression (DE) analysis of transcriptomic data allows for the study of gene expression changes associated with specific biological conditions across the entire genome. Typically, this analysis generates a large list of genes that exhibit differential expression between two or more groups. These identified differential expression genes (DEGs) can be subject to further downstream analysis to gain additional biological insights, such as identifying enriched functional pathways or gene ontologies. Additionally, DEGs are considered as candidate biomarkers, and a smaller subset of DEGs may be identified as potential biomarkers using either data-driven or biological knowledge-based approaches [9], [10]. Immune-related genes (IRGs) are a group of genes that

Contact author: Minh Tuan Nguyen
Email: nmtuan@ptit.edu.vn
Manuscript received: 10/2023, revised: 11/2023, accepted: 12/2023.

play important roles in the immune system's response to infection, inflammation, and other immunerelated processes, have been used as biomarker for diagnosis and prognostic signatures for various types of cancer, exhibiting high sensitivity and specificity [11]. Recent studies have shown that using IRGs to diagnose sepsis can significantly improve the accuracy of the diagnostic method [5], [11].

Gene selection is crucial for reducing data dimensionality and improving prediction efficiency. The authors of studies in[10], [12] utilized a two-layer gene selection approach that combined DEGs analysis with feature selection using machine learning techniques to identify potential genes and enhance prediction accuracy with a limited gene dataset. In [10], 10 genes were selected, resulting in an accuracy of 87.06% and an AUC of 89%. In [12], 9 genes were selected, resulting in an accuracy of 91.79% and an AUC of 85.66%. In other studies, researchers utilized IRGs and machine learning (ML) methods for gene selection, as seen in studies such as [5], [11]. In our work, we propose a novel three-layer gene selection approach that integrates DEGs, IRGs, and F-score based on deep learning (DL) to identify the optimal gene combination, and therefore enhance the performance of mortality prediction in pediatric sepsis.

ML has become an increasingly popular method for detecting and predicting biomarkers in recent years [13]. A study conducted previously demonstrated that ML algorithms can accurately predict the onset of sepsis in an ICU patient between 4-12 hours prior to clinical recognition based on medical data [14]. Additionally, various ML techniques have been utilized in other studies to predict patient outcomes in cases of sepsis [10], [12], [15]. Although, DL algorithms have demonstrated significant success in healthcare, such as biomedical signal processing. However, the application of DL in the field of infection detection, particularly sepsis diagnosis, has not been extensively explored. This may be due to the complex and multifactorial nature of sepsis, which involves various physiological and molecular processes. Nevertheless, recent studies [16], [17] have shown the potential of DL models in predicting sepsis onset and identifying sepsis biomarkers using various types of data, such as clinical, genomics, and metabolomics data. These studies suggest that DL algorithms have promising applications in the field of sepsis diagnosis and management. Further research is needed to fully understand the potential and limitations of DL in this area and to develop more accurate and effective models for sepsis diagnosis and treatment. Therefore, the aim of this study is to identify diagnostic biomarkers genes for sepsis using DL.

In our work, a novel diagnostic algorithm for pediatric sepsis that combines 7 IRGs using convolution neural network (CNN) algorithm is introduced. To identify the most relevant gene combinations, we use three-layer gene selection. Firstly we employ a sequential gene selection procedure, also known as DEGs, which identifies a subset of genes that are most informative for sepsis diagnosis; then these genes are then filtered out for IRGs; finally the most potential IRG combinations are selected based on DL models to validate the gene combinations ranked by F-score algorithm. Besides, the performance of the DL model using the IRGs is estimated through the 3-fold CV method on the validation set to increase reliability, making the results more reliable for practical use in clinic environments. The main contribution of this work:

- A novel three-layer gene selection selection, including DEGs, IRGs, and DL models based on F-score algorithm to identify optimal the number of genes, and therefore enhance the performance of prediction pediatric sepsis.
- Proposing a simple algorithm model and high predictive efficiency for the mortality of pediatric sepsis.
- Proposing a subset of genes associated immune to diagnosis pediatric sepsis mortality.

II. METHOD

The proposed method includes four steps shown in Fig. 1. The first step is to preprocess the gene dataset to compute the gene expression levels and perform differential expression analysis. After that the dataset is split into two equal parts, 50% for training and 50% for testing. In the second stage, we identify potential biomarker genes through DEGs analysis and subsequently filter the obtained gene list with immune-related genes. The gene ranking process is implemented using the F-score algorithm to identify the most effective gene combinations for improving the diagnosis of pediatric sepsis. A variety of gene combinations are generated and utilized as input for the DL models in order to assess their performance. Next, the third step is model validation and gene validation to identify a the best gene combination and optimal DL models to achieve the high accuracy of the diagnosis of pediatric sepsis. In the last step, the selected gene combination along with the optimal CNN and LSTM models are then subjected to performance evaluation on the testing data. The aim is to compare the performance of the models and determine the best DL model that can be used to propose an algorithm for diagnosing pediatric sepsis.

A. Data

The dataset we use in our work is GSE66099, which is publicly available and sourced from six other GEO databases [18]. The dataset included 276 unique patients, including 47 healthy controls, 18 sepsis patients, 181 septic shock patients, and 30 patients with systemic inflammatory response syndrome (SIRS). The dataset consists of patients who were included in six additional GEO datasets that were previously published by Hector Wong [19] and the Genomics of Pediatric SIRS and Septic Shock Investigators. This comprehensive dataset comprises all distinct patients from GSE4607, GSE8121, GSE9692, GSE13904, GSE26378, and GSE26440.

In this paper, we only use the dataset involving pediatric patients classified as sepsis and septic shock. The dataset contains expression profiles of about 10,596 genes from 199 children. Among the 199 pediatric patients, 28 did not survive within 24 hours of admission to the ICU. All samples were saved in CEL format and renormalized

using the R package affy's gcRMA method [20]. We downloaded probe-to-gene files from GEO and calculated the gene expression level by taking the mean. It is noteworthy that the number of patients is relatively small, which results in equal amount of training and testing datasets to ensure sufficient number of survivals and non-survivals in both datasets.

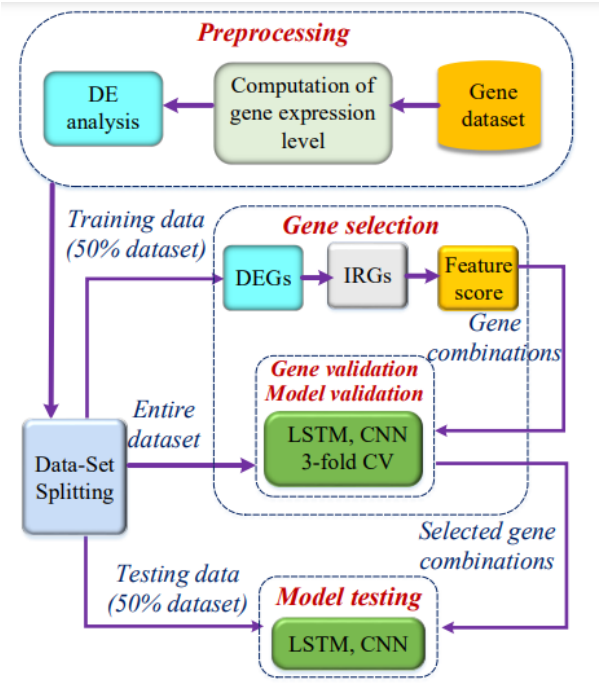

Figure 1. Method diagram

B. Identification differential expression genes

The gene dataset uses DE analysis, where the genes are tested individually for expression differences between conditions. The fold change is calculated and used as a crucial factor to distinguish between survivors and nonsurvivors. Specifically, non-surviving samples exhibited higher expression levels of up-regulated genes than surviving samples. To implement DEG analysis in this study, we utilized the R package as the simulation tool and employed the 'limma' package of R with the BenjaminiHochberg (BH) correction method [21] to identify DEGs. Additionally, the screening criteria for DEGs were adjusted P-value ≤ 0.05 and log fold change (LogFC) ≥ 1.5 to select representative DEGs for pediatric sepsis patients who either survived or did not survive [10]. The selection of a P-value threshold of 0.05 and a LogFC threshold of 1.5 in the identification of DEGs is grounded in statistical and biological relevance. It represents the probability of obtaining observed results, or more extreme results, under the null hypothesis. The logFC threshold of 1.5, on the other hand, reflects a practical determination of biological significance. This threshold ensures that only genes with a substantial magnitude of expression change are considered, aiding in the focus on alterations that are likely to be biologically meaningful.

C. Immune-related genes

IRGs are considered as potential biomarkers due to the association of sepsis with the immune system. These genes have been used in numerous studies related to pathogen infection and host response. A total of 770 IRGs were gathered from the nanoString database (www.nanoString.com), which has been utilized in numerous studies involving pathogen infection and the host response. Following the completion of the DEG analysis, we compared the genes identified as DEGs with this gene database of 770 IRGs to select immune genes.

D. Gene selection method

The selection of genes crucial for realizing biomarker potential in diagnosing sepsis poses a challenge. Therefore, we propose a novel three-layer gene selection approach, utilizing a sequential gene selection method to identify informative genes for sepsis diagnosis. This step aims to identify a set of genes related to the outcome of interest and identify small sets of genes suitable for diagnostic purposes in clinical practice. The three-layer gene selection includes DEGs to identify a subset of genes that are most informative for sepsis diagnosis, IRGs and gene validation based on DL model to validate the the gene combinations ranked by the Fscore algorithm. Also known as the Fisher score, this algorithm evaluates individual features in a dataset, measuring of the discriminatory power of each feature in distinguishing between two classes. The calculation involves both between-class and within-class variance.

E. Deep learning models

In order to optimize the LSTM and CNN models, their parameter structures are fine-tuned on the training set. Essential for avoiding overfitting and identifying the best models, hyperparameter tuning is carried out through an iterative process that optimizes external configuration settings, or hyperparameters. These hyperparameters, distinct from parameters learned during training, significantly impact a model's effectiveness. For CNNs, hyperparameters include the learning rate, batch size, and architecture-specific parameters such as the number of convolutional layers, filter sizes, and pooling strategies. In the case of LSTMs, tuning focuses on parameters like the learning rate, batch size, and LSTM-specific parameters such as the number of hidden units, the number of layers, and the dropout rate for regularization. The structure and parameters of CNN and LSTM models are shown in Table 1 and 2. Furthermore, the performance of the selected model is assessed on the validation set. To determine the optimal values for the model parameters, a combination of grid search and the 3-folds CV method is utilized in this study.

Convolutional Neural Networks: The neurons are a crucial component of the CNN, as they form the layers that make up the network. These neurons are arranged in three dimensions: the height and width of the input (known as the spatial dimensionality), as well as the depth. The CNN is composed of several layers, including the input layer, convolutional layer, rectified linear unit layer, pooling layer, fully connected layers, and output layer.

Table 1: The structure of CNN and LSTM models

Model	Layer	Number
CNN	Convolutional layer	6
	Relu	6
	Max pooling	4
	Fully connected	2
	Softmax	1
LSTM	LSTM layer	3
	Batch Normalization	3
	Drop out	1
	Fully connected	2
	Softmax	1

Table 2: Parameters of CNN and LSTM models

Model	Parameter	Value
CNN, LSTM	Learning rate	0.001
	Epoch	200
	Batchsize	100
	Optimizer	sgdm
	Momentum	0.95

1) Input layer: The preprocessed gene data is arranged as the input, which is then fed into the CNN.

2) Convolutional layer: It applies a set of filters, also known as kernels, to the input data. The filters convolve over the input data by performing a dot product operation between the filter weights and the corresponding input. The result of this computation determines the output of the neurons.

3) Rectified linear unit (ReLU): This is an activation function that is commonly used to transform the output of the previous layer. It is designed to be a more efficient activation function than the sigmoid function.

4) Pooling layer: In a CNN, the pooling layer is responsible for down-sampling the input data along its spatial dimensionality, thereby reducing the number of parameters in the activation. It accomplishes this by applying a fixed function over a sliding window of the input, such as taking the maximum or average value of the window. This results in a smaller output size compared to the input size, making it easier to process by subsequent layers.

5) Fully-connected layer: A fully-connected layer aims to produce class scores based on the activation obtained from the previous layer. This layer is used for classification purposes. To improve performance, ReLu can be added between the fully-connected layers.

6) Output layer: The output layer of the CNN includes both the Softmax and classification layers in which the former represents the probability distribution of a particular class assigned by the corresponding unit in classification and the later identifies the output as survival or non-survival, respectively.

Long Short-Term Memory: The problem of vanishing gradients, which occurs during the learning of long-term dependencies, even when the time lags are quite long, is a major issue. However, the LSTM model is an effective solution to address this problem. To prevent this issue, a constant error carousel is used that keeps the error signal within each cell of the unit. The LSTM architecture consists of a series of recurrently connected sub-networks, called memory blocks, which maintain state over time and regulate the flow of information through non-linear gating units. The output of the block is connected back to the input of the block, as well as to all of the gates.

F. Performance evaluation

The DL models' diagnosis performance using various gene combinations is evaluated based on different measured parameters, including accuracy (Acc), sensitivity (Se), specificity (Sp), Matthews correlation coefficient (Mcc), and area under the curve (AUC). The Acc parameter indicates the number of correctly identified pediatric patients. Se and Sp measure the number of correctly detected deaths and survivals due to sepsis, respectively. The Mcc measures the discrepancy between predicted and actual patients. Additionally, the AUC parameter evaluates the DL classifier's ability to differentiate between pediatric deaths and survivals caused by sepsis.

III. SIMULATION RESULTS

A. Differentially expressed genes, Immune-related gene and Gene ranking

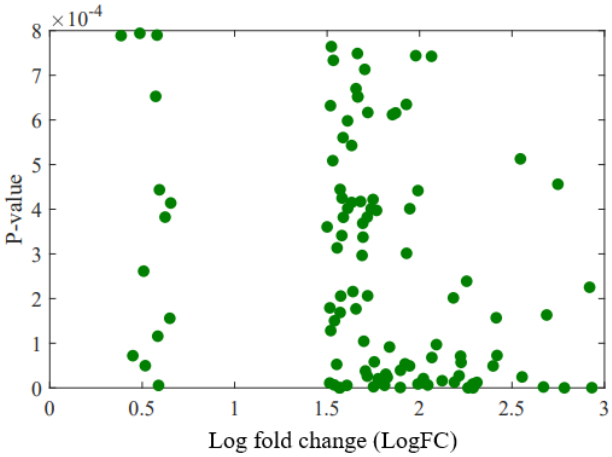

Figure 2. The scatter plot of p-value and log fold change for 108 DEGs

By applying a threshold of absolute log fold change LogFC 1.5 and p-value 0.05, 108 genes were identified as DEGs from 10596 genes of septic pediatric patients who survived and those who did not shown in Fig. 2. From 108 genes we matched with 770 IRGs, we got 12 genes, namely: CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3,

CCRL2, TFRC, STAT4, CCL20, CCR2, EBI3 that are related to the immune system.

The dataset consists of 12 genes obtained from IRGs that are normalized and preprocessed. The F-score algorithm is used to score the genes in the training dataset, and the corresponding score values are shown in Fig. 3. The genes are ranked in descending order according to their score values. A total of 12 gene combinations are generated, each combination consisting of a variable number of genes ranging from 1 to 12. The first gene combination is created from the gene with the highest score. In the second gene combination includes genes that are arranged with the first ranked score and the second ranked gene. Similarly to the 12th gene combination, there are 12 genes.

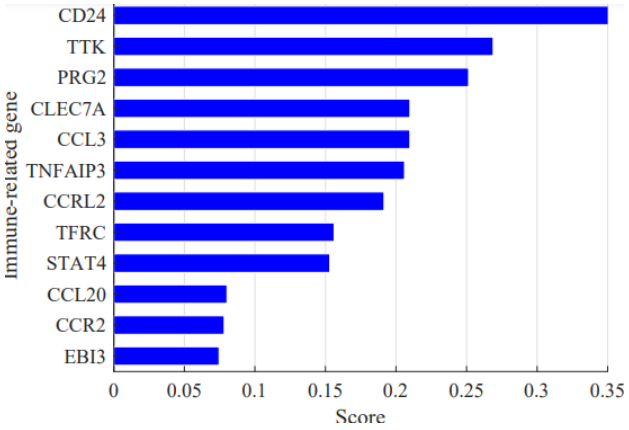

Figure 3. 12 immune-related gene ranked by F-score algorithm.

B. Gene validation and Model validation

In this stage, we trained and validated the models using the entire dataset via 3-fold CV. Because the gene dataset is relatively small, we validate the gene combinations using the entire dataset. The dataset is randomly divided into 3 folds, with 2 folds used for training the DL models and the remaining fold use for testing. This 3-fold CV process is repeated 3 times to complete a comprehensive procedure. The mean diagnostic performance of the DL models is then calculated for analysis and comparison. The selection of the optimal gene combinations is based on the highest performance of the corresponding DL models, which use these gene combinations as input for diagnosing pediatric sepsis. CNN and LSTM are employed to assess the diagnostic performance of various gene combinations in pediatric sepsis. we use different gene combinations in 3-fold CV on both CNN and LSTM models. The corresponding gene combination is used to evaluate the performance of each DL algorithm, and the metric score is calculated and analyzed. Specifically, each DL model is evaluated using 12 gene combinations. Based on the results of the DL models, we select the algorithm with the highest metric mean score, along with its corresponding gene combination, as the most effective algorithm and feature combination for the diagnosis model. Table 3 shows the highest validation performance results of the individual DL models. In the two DL models, the CNN model achieved the highest mean validation performance score using a combination of 7 genes, including CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3, CCRL2 which is a relatively small number of genes.

Table 3: The highest validation performance of the DL models on the entire dataset

DL model	Number of genes	Acc (%)	Se (%)	Sp (%)	Mcc (%)	AUC (%)
CNN	**7**	**91.92**	**33.33**	**96.72**	**46.67**	**87.86**
LSTM	7	83.84	20	89.07	12.38	61.31

Table 4: The performance of diagnosis pediatric sepsis on the testing set using DL models

DL model	Acc (%)	Se (%)	Sp (%)	Mcc (%)	AUC (%)
CNN	**91.92**	**33.33**	**96.72**	**46.67**	**87.86**
LSTM	83.84	20	89.07	12.38	61.31

Table 5: Confusion matrix of CNN/LSTM models on testing set based on gene combinations

Actual \ Predicted	Survival	Mortality
Survival	81/82	4/3
Mortality	5/12	9/2

C. Model testing

The proposed algorithm for diagnosing pediatric sepsis using DEGs and IRGs involves training and testing different DL models with the optimal gene combinations. The performance of each model is evaluated on the training and testing set using the selected gene combinations. The algorithm identifies the final gene combination and corresponding DL model that exhibits the highest diagnosis performance on the testing set. This approach allows for the development of an effective and accurate diagnosis method for pediatric sepsis using differential expression genes and DL models.The performance of DL models on the testing set is shown on Table 4, 5, where Table 4 shows the testing results of those models using different gene combinations and Table 5 shows the confusion matrix of different DL models on the testing set using the selected gene combinations.

IV. DISCUSSION

In this study, we propose an efficient three-layer gene selection method for selecting potential biomarkers to increase the accuracy of predicting mortality in pediatric septic patients. Indeed, in papers [10], [12], [5], [11], [22], by applying two-layer gene selection, they selected a small set of genes with potential to increase the efficiency of predicting the mortality rate of sepsis. In our work, a three-layer gene selection is proposed including DEGs, IRGs and gene validation based on DL models to valid gene combinations ranked by F-score algorithm.

DE analysis is a common method used to examine gene expression profiles and understand the underlying biological mechanisms of complex diseases. The analysis of gene expression data can be beneficial for predicting sepsis in pediatric patients. This type of data offers a wealth

of information that can be utilized to identify significant biomarkers and genetic pathways linked to sepsis. Gene expression profiles are typically highdimensional, with tens of thousands of genes and high correlations between them. Therefore, DE analysis tools often identify hundreds of highly correlated genes. In this work, from an original gene dataset of 10,596 genes, a subset of 108 genes is the outcome of DEGs analysis. The DE analysis is used to eliminate irrelevant genes, which contribute insignificantly to the diagnosis of pediatric sepsis. Molecular biomarkers have been recognized as noninvasive clinical tools that can provide objective predictions or evaluations of disease status and progression. The immune system's regulation of response and function has been shown to be crucial in the development and advancement of sepsis [11]. Study [23] demonstrated that certain genes linked with innate immune response could be utilized to predict the prognosis and diagnosis of children with clinical sepsis and were found to have promising clinical efficacy. Moreover, sepsis is a disease closely associated with the immune system of patients, so IRGs are being considered as potential biomarkers. Therefore, in our work we filter out 12 IRGs (namely CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3, CCRL2, TFRC, STAT4, CCL20, CCR2, EBI3) from 108 DEGs. To select the most potential IRGs, we used a set of DL models to evaluate the genomes that were ranked by F-score algorithm and combined into 12 combined gene sets.

Using LSTM and CNN to examine this data and recognize crucial features can be used to diagnose sepsis. LSTM is capable of capturing the temporal relationships in gene expression data, which is usually recorded as a time series with regular measurements over time. This makes LSTM well-suited to modeling the changes in gene expression over time that may indicate sepsis. On the other hand, CNN is suitable for identifying significant features in high-dimensional data, such as gene expression data. With the use of convolutional layers, CNN can extract patterns and motifs that are indicative of sepsis. Our study used both LSTM and CNN models to diagnose pediatric sepsis, and the results in Table 3,4, and 5 demonstrate this approach. Obviously, the CNN model outperforms the LSTM model in terms of diagnosis performance on the entire gene dataset and on the testing set. Because the convolutional layers of CNNs autonomously learn relevant patterns and relationships within the DEG, providing a powerful mechanism for discerning key genetic signatures associated with sepsis. Additionally, the parameter sharing property of CNNs allows them to efficiently identify crucial features, contributing to improved generalization on genomic datasets. While LSTMs are proficient in modeling sequential dependencies, CNNs' ability to exploit spatial structures in gene data makes them particularly well-suited for enhancing the accuracy and interpretability of sepsis classification based on DEG.

Therefore, we propose an effective simple algorithm that is the CNN model in combination with 7 IRGs selected from three-layer gene selection, including CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3 and CCRL2, for pediatric sepsis diagnosis. By using the 3-fold CV procedure in both gene selection and model validation, this makes our results more reliable. A comparison between the proposed algorithm and an existing method using two-layer gene selection and ML models with a similar GSE66099 dataset is presented in Table 6.

Table 6: Validation and testing performance comparison of the proposed algorithm to existing works using the same GSE6609 data set

Ref	No. genes	Acc (%)	Se (%)	Sp (%)	Mcc (%)	AUC (%)
Proposed algorithm on validation set	7	91.92	33.33	96.72	46.67	87.86
[10] on validation set	10	87.06	55.00	93.00	50.00	89.00
[12] on validation set	9	91.97	57.33	100	69.73	85.66
[12] on testing set	9	88.90	28.57	98.82	43.59	81.10
Proposed algorithm on testing set	**7**	**90/91**	**64.28**	**95.29**	**61.47**	**91.60**

V. CONCLUSIONS

Our study proposes a novel approach for predicting mortality in pediatric sepsis. This approach involves utilizing a combination of a CNN model and a set of 7 IRGs signature, specifically CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3, and CCRL2. The selection of these 7 marker genes was performed using a threelayer gene selection, which is a sequential gene selection procedure that involves identifying differential expression genes, immune-related genes and gene validation utilizing deep learning based F-score algorithm models to identify the most optimal gene combinations. By applying this approach, we were able to narrow down the list of potential biomarkers from 10,569 genes to the most relevant set of 7 IRGs, which significantly improved the accuracy and reliability of our mortality predictions.

REFERENCES

[1] C.-Y. Hsu, Y.-H. Tsai, C.-Y. Lin, Y.-C. Chang, H.-C. Chen, Y.-P. Chang, Y.-M. Chen, K.-T. Huang, Y.-H. Wang, C.-C. Wang et al., "Application of a 72 h national early warning score and incorporation with sequential organ failure assessment for predicting sepsis outcomes and risk stratification in an intensive care unit: A derivation and validation cohort study," *Journal of Personalized Medicine*, vol. 11, no. 9, p. 910, 2021.

[2] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith et al., "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[3] M. R. Atreya and H. R. Wong, "Precision medicine in pediatric sepsis," *Current opinion in pediatrics*, vol. 31, no. 3, p. 322, 2019.

[4] L. Evans, A. Rhodes, W. Alhazzani, M. Antonelli, C. M. Coopersmith, C. French, F. R. Machado, L. Mcintyre, M. Ostermann, H. C. Prescott et al., "Executive summary: surviving sepsis campaign: international guidelines for the management of sepsis and septic shock 2021," *Critical care medicine*, vol. 49, no. 11, pp. 1974–1982, 2021.

[5] H. She, L. Tan, R. Yang, J. Zheng, Y. Wang, Y. Du, X. Peng, Q. Li, H. Lu, X. Xiang et al., "Identification of featured necroptosis-related genes and imbalanced immune

infiltration in sepsis via machine learning," *Frontiers in Genetics*, vol. 14, p. 1158029, 2023.
[6] L. Li, L. Huang, C. Huang, J. Xu, Y. Huang, H. Luo, X. Lu, S. He, G. Yuan, L. Chen et al., "The multiomics landscape of serum exosomes during the development of sepsis," *Journal of Advanced Research*, vol. 39, pp. 203–223, 2022.
[7] L. J. Schlapbach and N. Kissoon, "Defining pediatric sepsis," *JAMA pediatrics*, vol. 172, no. 4, pp. 313–314, 2018.
[8] Y. Fan, Q. Han, J. Li, G. Ye, X. Zhang, T. Xu, and H. Li, "Revealing potential diagnostic gene biomarkers of septic shock based on machine learning analysis," *BMC Infectious Diseases*, vol. 22, no. 1, p. 65, 2022.
[9] A. Mohammed, Y. Cui, V. R. Mas, and R. Kamaleswaran, "Differential gene expression analysis reveals novel genes and pathways in pediatric septic shock patients," *Scientific reports*, vol. 9, no. 1, p. 11270, 2019.
[10] M. Abbas and Y. El-Manzalawy, "Machine learning based refined differential gene expression analysis of pediatric sepsis," *BMC medical genomics*, vol. 13, no. 1, pp. 1–10, 2020.
[11] Z.-H. Chen, W.-Y. Zhang, H. Ye, Y.-Q. Guo, K. Zhang, and X.-M. Fang, "A signature of immune-related genes correlating with clinical prognosis and immune microenvironment in sepsis," *BMC bioinformatics*, vol. 24, no. 1, p. 20, 2023.
[12] L. D. Vu, M. T. Nguyen, H.-C. Le et al., "Pediatric sepsis diagnosis based on differential gene expression and machine learning method," in 2022 14th International Conference on Knowledge and Systems Engineering (KSE). IEEE, 2022, pp. 1–6.
[13] N. Radakovich, M. Nagy, and A. Nazha, "Machine learning in haematological malignancies," *The Lancet Haematology*, vol. 7, no. 7, pp. e541–e550, 2020.
[14] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the icu," *Critical care medicine*, vol. 46, no. 4, p. 547, 2018.
[15] L. Ke, Y. Lu, H. Gao, C. Hu, J. Zhang, Q. Zhao, Z. Sun, and Z. Peng, "Identification of potential diagnostic and prognostic biomarkers for sepsis based on machine learning," *Computational and Structural Biotechnology Journal*, vol. 21, pp. 2316–2331, 2023.
[16] E. A. Strickler, J. Thomas, J. P. Thomas, B. Benjamin, and R. Shamsuddin, "Exploring a global interpretation mechanism for deep learning networks when predicting sepsis," *Scientific Reports*, vol. 13, no. 1, p. 3067, 2023.
[17] T. As¸uroglu and H. O ˘ gul, "A deep learning approach for sepsis monitoring via severity score estimation," *Computer methods and programs in biomedicine*, vol. 198, p. 105816, 2021.
[18] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse66099.
[19] T. E. Sweeney, A. Shidham, H. R. Wong, and P. Khatri, "A comprehensive time-course–based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set," *Science translational medicine*, vol. 7, no. 287, pp. 287ra71– 287ra71, 2015.
[20] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "affy—analysis of affymetrix genechip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.
[21] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.
[22] T. Gong, Y. Liu, Z. Tian, M. Zhang, H. Gao, Z. Peng, S. Yin, C. W. Cheung, and Y. Liu, "Identification of immune-related endoplasmic reticulum stress genes in sepsis using bioinformatics and machine learning," *Frontiers in Immunology*, vol. 13, p. 995974, 2022.
[23] H. R. Wong, "Pediatric sepsis biomarkers for prognostic and predictive enrichment," *Pediatric Research*, vol. 91, no. 2, pp. 283–288, 2022.

LỰA CHỌN GENE NỔI BẬT NHẰM NÂNG CAO HIỆU QUẢ CHẨN ĐOÁN NHIỄM TRÙNG MÁU DỰA TRÊN HỌC SÂU

Tóm tắt: Nghiên cứu đề xuất một phương pháp mới để chẩn đoán nhiễm trùng máu ở trẻ em sử dụng mô hình mạng lưới thần kinh tích chập CNN và sự kết hợp của 7 gene liên quan đến miễn dịch (IRG), bao gồm CD24, TTK, PRG2, CLEC7A, CCL3, TNFAIP3 và CCRL2. Bên cạnh đó, nghiên cứu cũng đề xuất quy trình chọn lọc gen ba lớp bao gồm quy trình tuần tự kết hợp phân tích biểu hiện gen khác biệt, sau đó là chọn lọc các gene có liên quan đến miễn dịch, cuối cùng là tính toán điểm gene bằng thuật toán F-score để xác định các gen biểu hiện khác biệt có nhiều thông tin nhất, sau đó sử dụng mô hình học sâu để xác định sự kết hợp gene tối ưu. Hiệu suất của thuật toán đề xuất được đánh giá bằng quy trình xác thực chéo 3 lần với các mô hình học sâu. Kết quả cho thấy các tổ hợp gene được chọn đạt độ chính xác 91.92% và diện tích dưới đường cong ROC là 87.86%, cho thấy thuật toán đề xuất là đáng tin cậy để dự đoán tỷ lệ tử vong do nhiễm trùng máu ở trẻ em. Ngoài ra, việc xác định dấu hiệu bao gồm 7 IRG gene liên quan đến tỷ lệ tử vong do nhiễm trùng máu ở trẻ em có khả năng hỗ trợ phát triển các dấu ấn sinh học để chẩn đoán và tiên lượng đáng tin cậy cho bệnh nhiễm trùng máu.

Từ khóa: Nhiễm trùng máu trẻ em, gene biểu hiện khác biệt, gene miễn dịch, lựa chọn gene nổi bật, học sâu.

Ngoc Anh Phung Thi is a B.E. student in Information Technology Department of Posts and Telecommunications Institute of Technology (PTIT) of Vietnam. Her research interests include machine learning, bioinformatics.

Anh Thu Pham received B.E degree of Telecommunication engineering from Posts and Telecommunications Institute of Technology (PTIT), Viet Nam, in 2003, and M.E degree of Telecommunication engineering from Royal Melbourne Institute of Technology, Australia, in 2008. She received the Ph.D. degree in Telecommunication engineering from PTIT, in 2010. Now, she is a lecturer in Telecommunication faculty of PTIT. Her research interests include networking, radio over fiber, and broadband networks.

Minh Tuan Nguyen received the B.S. degree from the Post & Telecommunications Institute of Technology, Hanoi, Vietnam, in 2004, the M.S. degree from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2008, both in electronics and telecommunications engineering, and the Ph.D. degree at the Gwangju Institute of Science and Technology, Gwangju, South Korea, in 2018. He is with Posts and Telecommunications Institute of Technology. His research interests include network security, internet of things, biomedical signal processing, gene analysis, sentiment analysis, brain computer interface, machine learning, deep learning, optimization, and biomedical application design.
Email: nmtuan@ptit.edu.vn