ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

A Hybrid of Modified Capsule and Transformer Model for Sepsis Diagnosis

Tuan Anh Vu[1], Hoai Bac Dang[2] and Minh Tuan Nguyen[3]

ABSTRACT

Sepsis is a critical and urgent medical condition that imposes a global health burden due to high mortality and the risk of long-term disability without prompt treatment. In this study, we propose a novel hybrid modified capsule and a transformer encoder (CaT) using a selected subset of biomarkers for the diagnosis of sepsis. The biomarkers are identified by a dual selection strategy that combines the differential expression analysis of immune-related genes with the Boruta algorithm using a random forest model. The modified capsule network consists of 4 parallel capsule layers, each implemented as a feedforward unit comprising a linear transformation followed by ReLU activation. On the validation set using Leave-One-Dataset-Out Cross-Validation, the CaT model shows better performance compared to other machine learning and deep learning models, with an accuracy of 96.8%, sensitivity of 98.0%, specificity of 87.9%, Mathews correlation coefficient of 85.6%, and area under curve of 98.0%. These findings highlight the robustness, generalization, and effectiveness of the proposed CaT model, demonstrating its potential as a reliable tool for the prediction of sepsis in clinical practice.

Article information:

Keywords: Sepsis, Immune-related Genes, Differentially Expressed Genes, Modified Capsule, Transformer Model

Article history:
Received: August 6, 2025
Revised: December 4, 2025
Accepted: December 25, 2025
Published: January 24, 2026 (Online)

DOI: 10.37936/ecti-cit.2026201.263411

1. INTRODUCTION

Sepsis is a major global health problem, which causes a high mortality rate and certainly long-term disabilities [1]. The complexity of this disease stems from the abnormal response of the host to infection, often associated with acute organ dysfunction [2], which requires essentially an early diagnosis of sepsis for effective and timely treatment. Sepsis induces a systemic, maladaptive response that often leads to remote multiorgan failure [3]. In addition, factors such as the source of infection and pathogen or patient-specific variables influence its expression [4]. Furthermore, the typical pathophysiological pattern in sepsis, with more than 80% of the transcriptional response in leukocytes, which are independent of the sources in terms of infection or pathogen, suggests a common transcriptional pattern among sepsis patients, providing a potential target for personalized strategies [5, 6].

Recent advances in genetics have provided significant insight into the molecular mechanisms of sepsis, which emphasize the important role of immune-related genes (IRGs) in terms of sepsis diagnosis [7]. Additionally, the identification of differentially expressed genes (DEGs) has been proven their importance due to the reflection of dynamic changes in gene expression profiles, which are associated with sepsis progression and patient outcomes [8]. Therefore, differential analysis of IRG expression allows the discovery of potential biomarkers, which contribute dramatically to early diagnosis, prognosis, and therapeutic targeting, thus facilitating timely and effective clinical management [9, 10].

The advancement of machine learning (ML) and deep learning (DL) has facilitated the emergence of numerous computational approaches for disease detection and diagnosis. Indeed, ML models have demonstrated a better ability to detect sepsis rapidly and accurately in comparison with traditional methods, which allow prompt intervention and potentially

[1]The author is with Center for Development of Information Technology and Communications (CDIT), Posts and Telecommunications Institute of Technology (PTIT), Vietnam, E-mail: vtanh@ptit.edu.vn
[2,3]The authors are with Faculty of Telecommunications, Posts and Telecommunications Institute of Technology (PTIT), Vietnam, E-mail: bacdh@ptit.edu.vn and nmtuan@ptit.edu.vn
[3]Corresponding author: nmtuan@ptit.edu.vn

improve patient outcomes [11]. Indeed, the authors of [12] identify vital IRGs, which are associated with sepsis, using integrated Weighted gene correlation network analysis (WGCNA), Estimation of STromal and Immune cells in MAlignant Tumor tissues, and different ML models such as Elastic net, Least absolute shrinkage and selection operator (LASSO) regression, random forest (RF), Boruta, and eXtreme Gradient Boosting (XGBoost). As a result, 11 potential IRGs are selected as the input of 6 ML models to predict sepsis, which obtains an area under the curves (AUC) value greater than 75%. In [13], gene selection and model estimation stages are designed with different ML models. The former uses a combination of 3 topological analysis methods and 4 ML models namely RF, LASSO, Support vector machine (SVM), and XGBoost for the identification of potential genes, the later considers Logistic regression (LR), AdaBoost, K-nearest neighbor (KNN), and XGboost as prediction models using the outcome of the former. Consequently, the KNN model achieves the highest performance related to sepsis recognition with an AUC of 99%.

Another significant method to address DEGs is the utility of intersection between the WGCNA and MEturquoise module genes as shown in [14]. Here, a total of 308 potential genes are identified for the input of 113 combinations, which are generated from 12 ML models to estimate their detection performance. The validation results indicate 22 biomarkers identified by the RF and Elastic Net models, which show the highest AUC of 88.1% among other combinations of models. In addition, Fan *et at.* [15] use Limma and metaMA packages to find differentially expressed mRNAs, which are then ranked by mean decrease accuracy values calculated by the RF model. Subsequently, a subset of 15 biomarkers is identified by a forward-wrapper approach in combination with various ML models. The highest validation performance with an AUC of 87.3% on the validation data is produced by RF model, which is proposed as a final algorithm for sepsis diagnosis. In [16], a modified LASSO penalized regression and SVM are employed to identify hub autophagy-related genes, which are then combined with an artificial neural network for the prediction of sepsis. Consequently, the proposed algorithm achieves a relatively high AUC performance over 85%. The authors of [17] consider Protein-Protein Interaction network analysis using STRING and Cytoscape to address 5 hub DEGs, used as the input of RF model, which shows strong predictive AUC performance of 84.94% on different datasets.

Recently, DL-based diagnostic models have been widely used for the designs of medical applications. Obviously, a large number of studies [18-20] emphasize the potential applications of the transformer model with respect to cancer disease using gene expression data. In the context of sepsis prediction, the authors of [21-23] employ Convolutional neural networks (CNN), Long short-term memory (LSTM) networks, and Transformer models for the design of sepsis detection algorithm. However, these models are largely relied on physiological signals and clinical measurements, such as vital signs, electrocardiograms, and biochemical laboratory values, which are definitely time-consuming. Furthermore, the application of DL models to gene expression data for the prediction of sepsis remains relatively unexplored. Only a few existing works follow this research direction to explore potentials of the DL modes. Typically, an AI-driven integrative framework, which combines a Transformer-based DL model and established ML models such as LASSO, SVM-Recursive feature elimination, RF and neural networks is proposed to uncover complex, non-linear interactions among gene-expression biomarkers [24]. As a result, AUC values ranging from 93.95% to 99.96% of the Transformer-based classifier confirm a subset of 5 genes as optimal diagnostic biomarkers.

The utility of small datasets collected from one or few platforms definitely reduces the reliability and restricts the generalization of the proposed diagnosis models [12-17]. Motivated by this gap, our work focuses on the development of an efficient DL framework using gene expression data collected from various platforms. Moreover, we propose a novel algorithm to predict sepsis disease, which contains an effective hybrid modified capsule and transformer encoder (CaT) model and a subset of biomarkers in this paper. Here, the Boruta and RF models are used to identify biomarkers from differentially expressed immune-related genes (DEIRGs). The proposed method validates biomarkers using Leave-One-Dataset-Out Cross-Validation (LODO-CV) procedure on the validation set, which contains 5 datasets randomly selected under constraints: datasets should originate from different platforms when possible; if platforms overlap, selected datasets must differ in age groups. The main contributions of this study are as follows:

- Investigation of multiple gene expression databases including various cell types, platforms, and age groups to ensure high generalization of the proposed prediction model.
- Identification of robust gene biomarkers associated with sepsis disease.
- Proposal of a simple CaT architecture as an efficient DL framework for sepsis recognition, which is optimized and estimated by grid search-based algorithm and statistical method using LODO-CV, respectively.

The rest of the paper is organized as follows. Section 2 describes the gene datasets with different platforms. Section 3 presents the methodology, followed by Section 4, which reports the simulation results.

Table 1: *Data description*

Dataset	No.Genes	Control	Sepsis	Cell type	Age	Platform
GSE119217	28376	12	122	Peripheral blood	Children	GPL16686
GSE69686	20299	85	64	Peripheral blood	Post-natal age	GPL20292
GSE69063	25512	33	57	Peripheral blood	Adult	GPL19983
GSE134347	30905	83	215	Whole blood	Aldult	GPL17586
GSE131761	21754	15	81	Peripheral Blood	Aldult	GPL13497
GSE57065	23520	25	82	Whole blood	Aldult	GPL570
GSE95233	23520	22	102	Whole blood	Adult	
GSE28750	23520	20	10	Whole blood	Aldult	
GSE26378	23520	21	82	Whole blood	Children	
GSE8121	23520	15	60	Whole blood	Children	
GSE13904	23520	18	52	Whole blood	Children	
GSE26440	23520	32	98	Whole blood	Children	
GSE9692	23520	15	30	Whole blood	Children	
GSE4067	23520	15	69	Whole blood	Children	
GSE65682	19040	42	479	Whole blood	Aldult	GPL13667
E-MTAB-1548	17028	15	80	Peripheral blood	Aldult	BioStudies

Section 5 provides a discussion, while Section 6 summarizes our study.

2. DATA

In this study, a total of 16 gene expression datasets are downloaded from the Gene expression omnibus (GEO) [25] and BioStudies [26] databases. A summary of these datasets is provided in Table 1. The datasets include 2 cell types, namely peripheral blood and whole blood, and span a wide range of age groups, including children, adults, and post-natal subjects. These datasets are derived from 8 different platforms, including GPL16686 (Affymetrix Human Gene 2.0 ST Array), GPL20292 (Custom Affymetrix Human Transcriptome Array), GPL19983 (Affymetrix Human Gene 2.1 ST Array), GPL17586 (Affymetrix Human Transcriptome Array 2.0), GPL13497 (Agilent-026652 Whole Human Genome Microarray 4x44K v2), GPL570 (Affymetrix Human Genome U133 Plus 2.0), GPL13667 (Affymetrix Human Genome U219 Array), and Agilent Human Gene Expression 4x44K v2 Microarray from the BioStudies database.

A total of 5 datasets are randomly selected under constraints: datasets should originate from different platforms when possible; if platforms overlap, selected datasets must differ in age groups. Therefore, GSE26378, GSE69063, GSE134347, GSE57065, and GSE119217 are chosen for the validation set, while the remaining datasets are used for training.

3. METHOD

The proposed method is shown in Figure 1, which includes 4 stages namely gene processing, dual selection of biomarkers, model optimization, and model validation.

- In the first stage, raw gene datasets are preprocessed using the Robust Multi-array Average (RMA) procedure [8], followed by the identification of the potential IRGs [10] using the intersection between expression profiles and curated immune gene databases.
- In the second stage, a dual selection strategy of biomarkers includes the differential expression analysis [8] and Biomarker selection is implemented to address the most relevant gene subset known as biomarkers. Indeed, the former plays a role to identify the DEIRGs among input IRGs in terms of sepsis detection using absolute fold-change and adjusted p-value, which are higher and lower than thresholds of 1.5 and 0.05, respectively. The latter employs a Boruta with a Random Forest (Boruta-

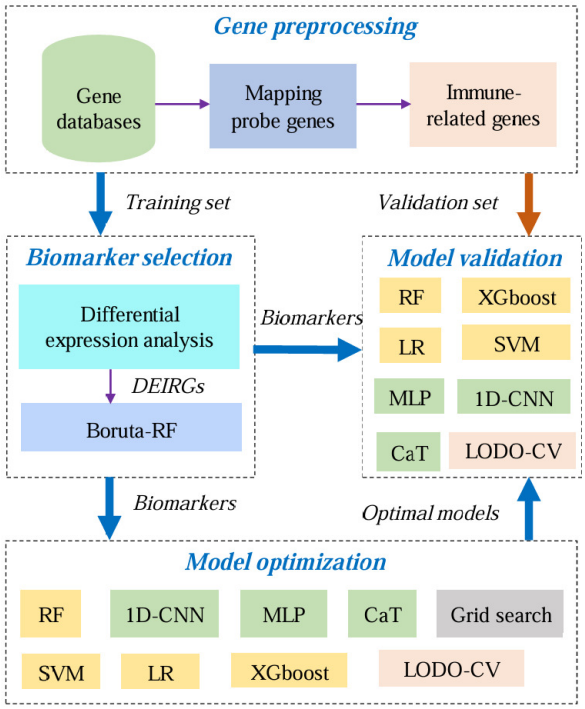

Fig.1: *Method diagram.*

RF) wrapper [28] to refine the most informative, contributive DEIRGs using a configuration of 100 trees and a maximum of 50 Boruta iterations.
- In the third stage, different classification models namely RF [31], LR, XGboost, SVM [13], one-dimensional CNN (1D-CNN) [32], MLP [33], and a hybrid CaT [29] are trained with the above selected biomarkers. Here, a grid search-based optimization algorithm is deployed to select the optimal hyper-parameters and structures of various models.
- In the last stage, the optimal models are validated for their classification performance on the validation set using the selected biomarkers and LODO-CV procedure for further comparisons and proposal of the final algorithm with respect to sepsis recognition.

3.1 Gene Processing

3.1.1 Mapping probe genes

We use Robust multi-array average (RMA) algorithm to preprocess the 16 raw gene expression datasets. Probe-to-gene annotation is conducted by aligning probe identifiers with corresponding gene symbols, utilizing the latest SOFT files or Chip Description Files (CDFs) obtained from the GEO database. Moreover, custom CDFs and SOFT files are applied for the input gene databases in which the former maps gene data of GSE119217 and GSE69063, and the latter processes the remaining gene databases using the average expression of probes corresponding to the similar genes, which represent the gene expression level.

3.1.2 Immune-related gene extraction

A referred database of IRGs is used to extract IRGs, which have a significant impact on sepsis disease, derived from the NanoString database (www.nanostring.com). Consequently, a total of 770 IRGs are obtained according to [10], which are considered as a standard IRG dataset for the selection of genes related to both sepsis disease and the immune system. Indeed, the individual gene datasets are compared with the above standard IRG dataset to address various subsets of IRG corresponding to different individual platforms. A gene intersection between all subsets of IRGs coming from various platforms is then identified as the potential IRGs, which represents the IRGs shared across all datasets. This step ensures a consistent gene space for the subsequent analysis. The resulting IRGs are then normalized using a Min–Max scaler to the range [0–1].

3.2 Dual selection of biomarkers

3.2.1 DEIRG selection

Differential analysis [8] is performed for the IRGs using the Limma package using R software, which is a part of the Bioconductor project for the analysis of gene expression data using microarrays and RNA-seq. The package provides linear models and empirical Bayes moderation in which the formers are fitted to expression data and the latter to stabilize the variances estimated across genes. Empirical Bayes moderation is particularly important in genomic studies using small databases with massive genes.

Besides, p-values and fold-change are the reliably statistical measurements, which are applied for the correction of multiple hypothesis testing. Indeed, the false discovery rate is controlled by p-values, which are certainly adjusted by Benjamini-Hochercg method [27]. This adjustment definitely reduces the possibility of false positive results (type I errors), which result in a statistical reliability with respects to the identification of the DEIRGs. In addition, fold-change is computed to quantify the magnitude of expression differences between septic patients and healthy controls. Positive and negative fold-change values show the up-regulated and down-regulated characteristics of genes related to sepsis. As a result, adjusted p-value less than 0.05 and fold-change over 1.5 are the main conditions to mark significantly differential expression as DEIRGs.

3.2.2 Biomarker selection

The Boruta method [28], known as a wrapper feature selection algorithm that works in conjunction with the RF classifier, is used to identify the biomarkers. Here, the RF model has the ability to measure the importance of each gene by evaluating how much predictive accuracy decreases when the values of those features are permuted. Unlike conventional feature selection methods, which frequently aim to find a minimal subset of features with a minimal error on a selected classifier, the Boruta method is proposed to retain all features which are either strongly or weakly relevant to the target variable. Therefore, this method is well suited for biomedical applications, which essentially make a decision of biomarkers certainly involved in a particular disease condition.

In this study, the DEIRG dataset is used as input of Boruta algorithm for the generation of shadow genes using shuffled copies of the total genes. Moreover, a RF model is trained on the orginal gene dataset and shadow gene dataset to calculate the important values of genes, which are further evaluated by decreases in mean accuracy. Then, the importance of each DEIRG is compared with the most important shadow gene in each iteration until either entire DEIRGs are labeled or repetition of RF model is completed. The final result is a subset of identified DEIRGs, known as biomarkers, which definitely are robust and effective for the sepsis diagnosis.

3.3 Model Optimization

There are 3 DL models namely 1D-CNN [32], MLP [33], the proposed CaT model, and 4 ML models in-

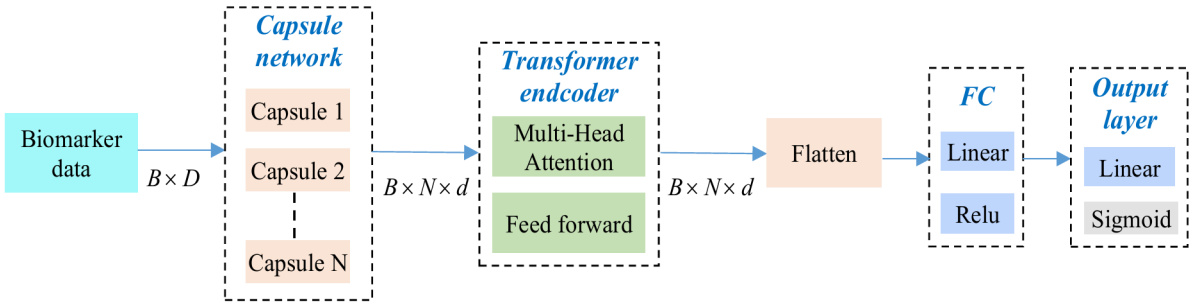

Fig. 2: *CaT model.*

cluding RF [31], LR, SVM, XGboost [13], which are implemented for performance analysis and comparisons. In addition, a grid search-based method combined with LODO-CV procedure is also applied for optimization of model hyperparameters to select the optimal parameters and structures of different ML and DL models.

The parameters are considered for the RF model namely number of trees in a range of [75, 85, 95, 105] and the maximum tree depth searched in [3, 5, 10, 15, 20], which results in a total of 20 RF structures. The LR model is implemented with a fixed configuration, which employs a L2 regularization solver of "lbfgs" and a maximum of 1000 iterations. Moreover, RBF kernel is selected for SVM model using parameter C identified in a range of [0.1, 10] using a step size of 0.1, which results in 100 structures of SVM model. The optimal structure of XGboost model is addressed by optimization of parameter values from different ranges such as tree number of [100, 200, 300], learning rate of [0.01, 0.05, 0.1], and maximum tree depth of [3, 5, 7], which result in 27 structures. Moreover, the learning rate, batch size, optimizer, and number of epochs are necessary to fine-tuned for the 1D-CNN and MLP models to maximize the performance and avoid overfitting. Here, we apply various range of values for the grid search to identify the optimal parameters of the 1D-CNN and MLP models such as optimizers of [RMSprop, Adam, SGD], batch-size of [16, 32, 64], epochs of [40, 50, 60], and learning-rate of [0.0001, 0.0005, 0.001]. Besides, a group of 4 different layers including 1D convolutional, ReLU activation, batch normalization, and max-pooling is defined to make different 1D-CNN structures. Indeed, we consider 10 1D-CNN structures containing 1 to 10 groups, which leads to 810 1D-CNN structures. In addition, the structures of the MLP models are optimized with the number of hidden layers ranging from 1 to 5 and the number of neurons of [16, 32, 64], which results in 1215 model structures.

In the CaT model as shown in Figure 2, we propose a modification of the capsule network, which is implemented via parallel feedforward projections without dynamic routing. The capsule-like embeddings are subsequently contextualized by a Transformer encoder to capture global dependencies between genes. The input in the CaT architecture is a $B \times D$ matrix of biomarkers, where B denotes the batch-size and D denotes the number of biomarkers. The input is first passed through a capsule network module with N parallel capsule units. Unlike capsule network in study [29], which uses squashing functions and dynamic routing mechanisms, the capsules in the proposed CaT model are implemented as simplified feedforward subnetworks, each consisting of a linear layer followed by a ReLU activation. This maps the input biomarkers into a tensor of shape $B \times N \times d$, where, N and d denote the number of capsules and the dimension of the capsule, respectively. The output of stacked capsules is passed to a transformer encoder to model contextual interactions among the capsules. The transformer encoder [30] consists of 2 main layers, namely multi-head self-attention and feedforward. The multi-head self-attention mechanism enables each capsule embedding to attend to all other capsule embeddings, allowing the model to capture both local and global dependencies. The feedforward network refines the attended representations by increasing the ability of the model to learn complex transformations. After the transformer encoder, a flatten layer is used, which is then passed through a fully connected (FC) layer. Finally, an output layer is applied to generate prediction probabilities. To optimize the performance of the CaT model, a grid search is applied for the structure parameters such as the number of capsules ranging from 2 to 6 , the dimension of the capsule of [16, 32, 64], the number of transformer encoder layers ranging from 1 to 3, and the number of attention heads ranging from 2 to 4. In addition, optimal learning parameters such as optimizers of [RMSprop, Adam, SGD], batch-size of [16, 32, 64], epochs of [40, 50, 60], and learning-rate of [0.0001, 0.0005, 0.001] are also addressed by the grid search algorithm for the CaT model, which results in a total of 10.935 CaT model structures.

3.4 Model Validation

The validation set includes 5 datasets in which children and adult samples are available in GSE126378 and GSE57065 datasets of the GPL570 platform

while the remaining datasets namely GSE69063, GSE134347, and GSE119217 are of the GPL19983, GPL17586, and GPL16686 platforms, respectively. The LODO-CV procedure is implemented to estimate the model performance on the validation set. Indeed, 4 datasets are combined to use for training, and the last is for testing. The repetition of the procedure is run 5 times to ensure that every single dataset is considered as a testing dataset.

4. SIMULATION RESULTS

4.1 Performance measurement

To comprehensively and reliably evaluate the diagnostic ability of each model in classifying between sepsis and control samples, 5 metrics [8] including Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), Mathews correlation coefficient (Mcc), and the Area under curve (AUC) are calculated to estimate the performance of the models in this study. Acc represents the proportion of participants who are correctly predicted. Sen and Spe represent the number of correctly detected sepsis patients and control people, respectively. The quality of a binary classification between sepsis patients and control people is measured by the Mcc parameter. Furthermore, AUC evaluates the ability of the ML and DL models to discriminate between sepsis patients and controls.

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

$$Sen = \frac{TP}{TP + FN} \tag{2}$$

$$Spe = \frac{TN}{TN + FP} \tag{3}$$

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

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

4.2 Gene processing

4.2.1 Mapping probe genes

The RMA technique is adopted to process 16 gene expression datasets followed by gene annotation by mapping using SOFT and CDF files as detailed in Table 1, which result in preprocessed gene datasets containing the number of genes between 17,028 and 30,905.

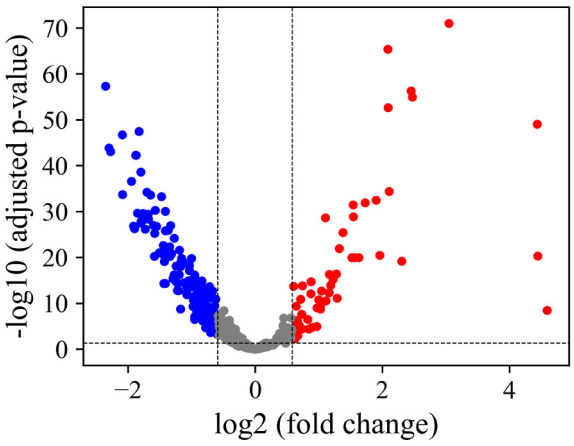

Fig.3: *DEIRGs of sepsis and controls; Red and Blue represent up-regulated and down-regulated DEIRGs, while Gray indicates insignificant IRGs.*

4.2.2 Immune-related gene extraction

The 16 gene expression datasets are compared with a standard dataset, which includes 770 IRGs. The number of IRGs, which are related to the immune system are 760, 696, 742, 755, 751 IRGs corresponding to GSE119217, GSE69686, GSE69063, GSE134347, GSE131761 datasets, respectively. Moreover, the GPL570 platform provides 11 gene datasets in which the numbers of IRGs are 740 and 627 of GSE65682 and E-MTAB-1548, and the remaining gene datasets consist of 737 IRGs. A total of 16 subsets of IRGs are then compared to extract an intersected part of IRGs, which results in a common subset of 560 IRGs for further selection of DEIRGs.

4.3 Dual selection of biomarkers

4.3.1 DEIRG selection

A subset of 219 DEIRGs is identified by differential expression analysis, which includes 56 up-regulated and 163 down-regulated DEIRGs using thresholds for the adjusted p-value and the fold-change as shown in Figure 3. The up-regulated DEIRGs, which have adjusted the p-value < 0.05 and absolute fold-change value > 1.5, show higher expression levels in sepsis patients compared to healthy controls, while down-regulated DEIRGs exhibit reduced expression, which have adjusted p-value over 0.05 and absolute fold-change value lower than 1.5.

4.3.2 Biomarker selection

We implement the Boruta-RF method using a subset of 219 DEIRGs as input for selection of biomarkers. Consequently, a subset of 135 biomarkers is addressed in which 55 and 80 biomarkers are up-regulated and down-regulated, respectively, as shown in Table 2. These biomarkers are then validated for their classification performance using various ML and

Table 2: *Selected biomarkers from DEIRGs.*

No.	Biomarkers	FC	P.Value	adj.P.Val	Regulation	No.	Biomarkers	FC	P.Value	adj.P.Val	Regulation
1	CEACAM8	24.09	9.42E-10	3.38E-09	Up	68	STAT4	-2.37	1.01E-19	8.04E-19	Down
2	LCN2	21.79	4.78E-22	4.78E-21	Up	69	CSF1R	-2.34	3.00E-14	1.67E-13	Down
3	IL1R2	21.65	1.20E-51	9.62E-50	Up	70	ITGAM	2.34	7.97E-17	5.58E-16	Up
4	S100A12	8.26	1.69E-74	9.46E-72	Up	71	HLA.DMA	-2.33	7.69E-16	4.79E-15	Down
5	ARG1	5.55	1.02E-57	1.14E-55	Up	72	GZMB	-2.31	3.27E-14	1.78E-13	Down
6	IL18R1	5.47	3.68E-59	5.15E-57	Up	73	ILF3	-2.29	7.98E-18	5.96E-17	Down
7	LRRN3	-5.11	2.67E-60	4.99E-58	Down	74	IFNGR1	2.28	1.97E-15	1.20E-14	Up
8	CLEC5A	4.95	8.20E-21	7.06E-20	Up	75	HLA.DRA	-2.26	4.53E-10	1.69E-09	Down
9	CD3G	-4.92	2.44E-46	1.37E-44	Down	76	CARD11	-2.25	1.31E-20	1.08E-19	Down
10	CD3E	-4.84	1.68E-45	8.57E-44	Down	77	CAMP	2.25	6.66E-18	5.04E-17	Up
11	CD163	4.31	1.18E-36	4.12E-35	Up	78	CD55	2.24	9.88E-14	5.12E-13	Up
12	C3AR1	4.27	2.35E-55	2.19E-53	Up	79	SH2D1A	-2.23	1.90E-21	1.68E-20	Down
13	CD96	-4.26	3.11E-49	1.94E-47	Down	80	ST6GAL1	-2.20	9.38E-21	7.84E-20	Down
14	TLR5	4.25	1.54E-68	4.31E-66	Up	81	TNFSF13B	2.16	6.22E-12	2.77E-11	Up
15	ITK	-4.25	6.59E-36	2.05E-34	Down	82	IL1R1	2.15	1.19E-30	2.22E-29	Up
16	S100A8	3.89	3.28E-22	3.34E-21	Up	83	SIGIRR	-2.12	3.37E-15	1.97E-14	Down
17	KLRB1	-3.85	6.78E-39	2.53E-37	Down	84	TBP	-2.11	1.18E-13	6.05E-13	Down
18	CCR7	-3.76	9.79E-29	1.48E-27	Down	85	CD81	-2.11	2.56E-13	1.25E-12	Down
19	IL18RAP	3.74	1.20E-34	3.19E-33	Up	86	NCF4	2.07	1.99E-11	8.42E-11	Up
20	HLA.DQA1	-3.72	3.91E-28	5.61E-27	Down	87	TLR8	2.06	3.91E-14	2.09E-13	Up
21	TXK	-3.69	1.08E-44	5.04E-43	Down	88	CXCR3	-2.05	9.58E-19	7.45E-18	Down
22	KLRG1	-3.66	1.24E-44	5.35E-43	Down	89	TLR2	2.04	4.91E-10	1.82E-09	Up
23	MS4A1	-3.61	1.04E-31	2.25E-30	Down	90	DOCK9	-2.01	1.92E-21	1.68E-20	Down
24	FCER1A	-3.55	4.87E-50	3.41E-48	Down	91	NFATC3	-2.00	4.21E-15	2.41E-14	Down
25	GNLY	-3.48	1.96E-29	3.22E-28	Down	92	TLR1	2.00	3.74E-12	1.69E-11	Up
26	KLRF1	-3.48	6.50E-41	2.60E-39	Down	93	ANXA1	1.96	2.27E-10	8.83E-10	Up
27	CTSW	-3.47	7.44E-30	1.26E-28	Down	94	CTSG	1.96	4.42E-06	1.12E-05	Up
28	CD2	-3.39	1.30E-31	2.70E-30	Down	95	TBX21	-1.96	6.53E-13	3.13E-12	Down
29	CD247	-3.38	3.32E-30	5.81E-29	Down	96	AKT3	-1.94	8.90E-18	6.56E-17	Down
30	LCK	-3.32	5.18E-28	7.25E-27	Down	97	LTB	-1.93	2.35E-08	7.43E-08	Down
31	LY96	3.31	4.63E-34	1.18E-32	Up	98	HLA.DPA1	-1.93	1.01E-07	3.03E-07	Down
32	LY9	-3.26	1.75E-36	5.78E-35	Down	99	CD8B	-1.93	6.33E-16	4.03E-15	Down
33	ZAP70	-3.21	1.99E-31	3.98E-30	Down	100	SLAMF1	-1.93	3.88E-17	2.79E-16	Down
34	CD5	-3.21	2.45E-30	4.42E-29	Down	101	FCER2	-1.92	9.70E-17	6.71E-16	Down
35	CD40LG	-3.13	8.25E-36	2.43E-34	Down	102	REPS1	-1.90	2.48E-13	1.23E-12	Down
36	CD63	3.10	1.21E-21	1.11E-20	Up	103	IFITM1	1.86	7.26E-06	1.79E-05	Up
37	ETS1	-3.03	6.50E-29	1.04E-27	Down	104	FUT7	1.84	3.02E-16	1.99E-15	Up
38	CD3D	-3.00	6.03E-22	5.72E-21	Down	105	CD58	1.83	1.63E-13	8.24E-13	Up
39	ITGA4	-2.99	4.40E-27	5.60E-26	Down	106	LCP1	1.79	1.78E-05	4.04E-05	Up
40	EOMES	-2.98	2.31E-32	5.39E-31	Down	107	RELA	-1.79	1.79E-09	6.30E-09	Down
41	FCER1G	2.95	1.28E-21	1.16E-20	Up	108	NFKBIA	1.77	1.07E-07	3.19E-07	Up
42	CD6	-2.93	1.11E-28	1.64E-27	Down	109	CYLD	-1.77	3.43E-09	1.19E-08	Down
43	BST1	2.91	7.06E-31	1.36E-29	Up	110	SPN	-1.76	6.95E-11	2.84E-10	Down
44	MAPK14	2.91	1.41E-33	3.43E-32	Up	111	ADA	-1.75	8.49E-10	3.11E-09	Down
45	BCL6	2.87	1.14E-21	1.06E-20	Up	112	NFATC1	-1.74	1.80E-11	7.75E-11	Down
46	GZMK	-2.84	8.24E-23	9.04E-22	Down	113	CLEC7A	-1.71	4.42E-08	1.37E-07	Down
47	DPP4	-2.78	2.00E-35	5.60E-34	Down	114	PDGFC	1.68	2.45E-15	1.47E-14	Up
48	RUNX3	-2.71	1.48E-22	1.56E-21	Down	115	IL6ST	-1.67	1.89E-07	5.46E-07	Down
49	PRF1	-2.69	8.74E-21	7.41E-20	Down	116	LGALS3	1.67	1.76E-05	4.00E-05	Up
50	CD7	-2.68	5.73E-22	5.54E-21	Down	117	C1QA	1.66	7.06E-09	2.38E-08	Up
51	CD160	-2.66	3.93E-32	8.79E-31	Down	118	SF3A3	-1.65	1.04E-08	3.40E-08	Down
52	CD28	-2.66	1.08E-27	1.44E-26	Down	119	NT5E	-1.64	3.95E-15	2.28E-14	Down
53	IL10RA	-2.66	7.78E-16	4.79E-15	Down	120	CHIT1	1.64	2.96E-12	1.35E-11	Up
54	CR1	2.60	2.80E-27	3.65E-26	Up	121	CCR1	1.62	2.60E-06	6.85E-06	Up
55	GZMM	-2.60	4.50E-23	5.04E-22	Down	122	SELL	1.59	0.000653	0.001184	Up
56	NFATC2	-2.58	8.01E-28	1.09E-26	Down	123	BATF	1.59	3.28E-07	9.28E-07	Up
57	BLNK	-2.54	3.99E-24	4.76E-23	Down	124	JAK2	1.58	1.83E-07	5.33E-07	Up
58	GATA3	-2.51	7.11E-29	1.11E-27	Down	125	LYN	1.57	0.000183	0.000371	Up
59	FPR2	2.50	9.82E-24	1.15E-22	Up	126	PRKCD	1.56	3.71E-06	9.49E-06	Up
60	BCL2	-2.50	1.35E-22	1.45E-21	Down	127	MERTK	1.56	1.11E-10	4.36E-10	Up
61	KLRD1	-2.49	1.62E-22	1.68E-21	Down	128	FEZ1	-1.56	9.06E-11	3.63E-10	Down
62	CD74	-2.45	5.60E-17	3.97E-16	Down	129	IFITM2	1.55	0.002973	0.004771	Up
63	CEBPB	2.45	1.70E-12	7.85E-12	Up	130	PYCARD	1.55	8.81E-05	0.000183	Up
64	C1QB	2.42	5.29E-18	4.06E-17	Up	131	CSF3R	1.54	0.000458	0.00086	Up
65	HLA.DPB1	-2.42	9.30E-18	6.76E-17	Down	132	C1QBP	-1.53	8.02E-06	1.94E-05	Down
66	SH2D1B	-2.41	5.39E-26	6.71E-25	Down	133	FCGR2A	1.53	0.00026	0.000517	Up
67	GZMA	-2.38	2.59E-16	1.72E-15	Down	134	IL10	1.52	3.33E-15	1.96E-14	Up
						135	CFD	-1.51	0.000265	0.000523	Down

Table 3: *Performance of models on validation set.*

Model	Acc (%)	Sen (%)	Spe (%)	Mcc (%)	AUC (%)
1D-CNN	90.0 ± 9.8	96.0 ± 5.0	75.00 ± 27.5	71.4 ± 25.6	96.7 ± 6.8
MLP	89.7 ± 9.3	93.5 ± 10.1	77.5 ± 23.1	71.9 ± 26.2	97.4 ± 5.3
RF	90.0 ± 8.4	96.2 ± 8.4	70.1 ± 24.0	70.3 ± 27.0	96.5 ± 7.2
LR	83.3 ± 12.4	86.5 ± 16.2	73.2 ± 28.6	60.5± 25.7	96.7 ± 6.7
SVM	88.8 ± 13.2	95.3 ± 10.6	71.1 ± 39.8	66.2 ± 40.4	97.3±5.7
XGboost	94.4 ± 4.3	98.2 ± 2.7	78.4± 18.9	79.3 ± 20.8	97.4± 4.3
CaT	**96.8 ± 3.7**	**98.0 ± 2.2**	**87.9 ± 18.5**	**85.6 ± 21.3**	**98.0 ± 4.0**

DL models in combination with LODO-CV procedure.

4.4 Model optimization

A grid search-based method in combination with LODO-CV procedure is applied for selection of the optimal parameters and structures of the ML and DL models. As a result, the optimal RF model is selected with 95 trees and a maximum depth of 5. The SVM model achieves its best performance with a RBF kernel and a regularization parameter C of 1.0. The XGBoost model is optimized with 300 trees, a learning rate of 0.05, and a maximum depth of 5. Moreover, the optimal 1D-CNN model is selected with 3 groups using Adam optimizer, learning-rate of 0.001, batch-size of 32, epoch of 50. The optimal parameters of the MLP model is identified by the grid search method with 2 hidden layers using 64 neurons, optimizer of Adam, learning-rate of 0.0005, batch-size of 32, and epoch of 50. The CaT model is addressed with an optimal structure, which includes 4 capsules using a dimension of 32 followed by a transformer encoder using 2 attention heads, optimizer of Adam, learning-rate of 0.001, batch-size of 32, and epoch of 40.

4.5 Model validation

Table 3. shows the validation performance of different models using LODO-CV procedure. The proposed CaT model produces the highest performance of all evaluation metrics in comparison with other models using the selected biomarkers, which demonstrates superior generalization and robustness. Hence, the CaT model and the selected biomarkers are proposed as the final algorithm for the diagnosis of sepsis.

5. DISCUSSION

Sepsis is a life-threatening disease caused by the dysregulated host response to the infection and is the leading cause of death in patients in the intensive care unit. Early and accurate prediction of sepsis is crucial to improving patient outcomes, as timely interventions significantly reduce the risk of complications and mortality. Therefore, studies related sepsis detection are paid intensive attention from clinic experts and technicians to improve the performance and reliability of the proposed algorithms for the applications in real-world hospital environments.

Unfortunately, limitations are still available in many existing publications in which different approaches rely on small gene expression datasets downloaded from one or several platforms, which restrict the generalization and real-world applicability of the final proposed algorithms. In addition, while ML models such as KNN [13] or RF [15] have been employed for the designs of proposed algorithms with respect to the diagnosis of sepsis, very few studies have explored DL methods. Indeed, the DL architectures, which are productive to capture complex non-linear patterns and contextual dependencies in gene expression data, have been considered insufficiently. This highlights the need for robust and scalable DL models, which are specifically designed for sepsis prediction applications. Therefore, we address the gaps of previous studies by the utility of 16 gene expression datasets from 8 different platforms, which are collected from the GEO database and BiOstudies, representing diverse experimental conditions and population groups in this work. Furthermore, a novel proposal of a framework that includes a hybrid CaT model with 135 biomarkers to diagnose sepsis is also presented.

The dual selection strategy of biomarkers is proposed to improve the robustness of diagnostic modeling, which includes differential analysis and the Boruta-RF method. The differential expression analysis initially identifies 219 DEIRGs, reflecting the dysregulation of immune pathways during the progression of sepsis. However, relying solely on differential analysis retains redundant features, potentially compromising the generalization of models. Therefore, the Boruta-RF algorithm is subsequently applied to identify the most relevant DEIRGs by comparing them with random shadow DEIRGs, which result in a subset of 135 biomarkers as given in Table 2. This approach ensures that the final biomarkers capture genes, which are altered during infection-induced systemic inflammation and retain DEIRGs with potential predictive value.

Another significant characteristic is that the CaT model, which combines simple capsule layers with a transformer encoder, inspired by Capsule Networks,

is proposed in this study. Instead of using complex dynamic routing mechanisms, the capsule component in CaT is modified for the construction of independent and parallel feedforward subnetworks in which the individual subnetwork consists of a linear transformation followed by a ReLU activation. The linear layer projects the biomarker features into a capsule-specific latent space, while the ReLU activation introduces nonlinearity to capture complex biomarker interactions. Therefore, this network ensures efficient gene extraction and enhances the ability of the CaT model to learn discriminative local representations. This design enables the CaT model to project the input gene expression vector into multiple capsule embeddings, which captures distinct aspects of the underlying biological patterns. The resulting embeddings are then contextualized by self-attention mechanisms in the Transformer encoder. Moreover, the optimization of hyperparameters and structures for the models plays a vital role in improving the detection performance of the proposed algorithm. Hence, a grid search with LODO-CV procedure is applied for identification of optimal models, namely RF, LR, SVM, XGboost, CNN, MLP and CaT in this work, which systematically explores optimal parameter combinations and selects the most effective structures based on the highest validation performance, leading to overfitting avoidance and generalization improvement. The optimal CaT structure includes 4 capsules in which each of them adopts a dimension of 32, followed by a transformer encoder with 2 attention heads.

A comparison with existing works, as shown in Table 4, highlights the superior performance of the proposed algorithm. Indeed, the authors of [29] introduce a capsule network combined with a transformer to develop a sepsis diagnostic model using single-cell RNA sequencing data and subsequently transfer it to bulk RNA data. Their capsule network consists of fully connected and capsule layers in which the former projects gene expression into eight primary capsules and the latter contains 20 capsules, each of them represented as a 16-dimensional vector computed through dynamic routing. However, we propose a simple capsule network, including 4 capsules, each consisting of a linear projection followed by a Relu activation. This simple modification significantly reduces computational complexity while still enabling the generation of diverse capsule embeddings. As shown in Table 4, the proposed CaT model achieves similar value of AUC but remains simplicity and generalization in comparison with [29].

The first limitation of our work is a large subset of 135 final biomarkers selected by the proposed method, which requires a time-consuming, complexity process, which possibly increases the overall cost to profile them in real-world clinical settings and results in a significant challenge for application deployment on a large-scale. The second limit is the imbalanced class between sepsis and controls, which is exhibited in the datasets used in this study.

Table 4: *Performance comparison between the proposed algorithm and existing works.*

Ref	Acc	Mcc	AUC
[15]	NA	71.3	83.7
[17]	NA	NA	84.9
[129]	NA	NA	98.0
Proposed algorithm	96.8	85.6	98.0

It is obvious that the limitations of this study will be addressed in future work. Indeed, differently advanced methods to identify a small and efficient biomarker subset are implemented to enhance practicality and cost-effectiveness in clinical applications. Besides, variously effective strategies to address imbalanced class problem are proposed to improve the robustness of the models . Additionally, advanced deep learning architectures, which have emerged as promising algorithms to enhance predictive performance, will be investigated and employed for further studies.

6. CONCLUSION

Sepsis is a serious disease associated with high mortality and long-term sequelae, which places a substantial burden on healthcare systems. Therefore, accurate and timely prediction is essential to alleviate this burden.

In this paper, we propose a novel method that includes dual selection of biomarkers and an efficient and simple CaT model consisting of 4 capsule layers and a transformer encoder. The dual selection of biomarkers including differential expression analysis of IRGs to identify DEIRGs, followed by Boruta-RF to select biomarkers, allows an effective search for the most informative biomarkers related to sepsis detection. The utility of a large number of gene datasets from different platforms covering diversity of the participants, which reflects the heterogeneity commonly encountered in real-world clinical settings, definitely improves the generalization and reliability of the proposed algorithm. Additionally, the quality of selected biomarkers, which represent the most important characteristic of the sepsis patients, is also significantly improved due to the use of the massive gene datasets. A significant modification is implemented in CaT model, which considers the individual capsule layer as a feedforward subnetwork comprising a linear transformation followed by a ReLU activation for projection of input biomarkers into vectorized capsule embeddings processed by the transformer encoder to capture global interactions across the biomarkers. The above modification of CaT model combined with transformer encoder produces a relatively high performance with an Acc of 96.8%,

Sen of 98.0%, Spe of 87.9%, Mcc of 85.6%, and AUC of 98.0% using LODO-CV procedure and the selected biomarkers. The simulated results confirm the simplicity, robustness, and reliability of the proposed algorithm for the recognition of the sepsis, which is potential application in practical environments.

ACKNOWLEDGEMENT

This work was funded by Vietnam Ministry of Science and Technology (MST) under Grant number ĐT.68/25.

AUTHOR CONTRIBUTIONS

Conceptualization, Minh Tuan Nguyen; methodology, Minh Tuan Nguyen and Tuan Anh Vu; software, Tuan Anh Vu; validation, Hoai Bac Dang; formal analysis, Tuan Anh Vu; investigation, Tuan Anh Vu; data curation, Tuan Anh Vu; writing—original draft preparation, Tuan Anh Vu; writing—review and editing, Hoai Bac Dang and Minh Tuan Nguyen; visualization, Tuan Anh Vu; supervision, Hoai Bac Dang; funding acquisition, Minh Tuan Nguyen. All authors have read and agreed to the published version of the manuscript.

References

[1] N. Wang, H. Huang, Y. Tan and N. Zhang, "Research progress of biomarkers for sepsis and precision medicine," *Emergency Medicine International*, vol. 2025, no. 1, p. 4585495, 2025.

[2] R. Pan, J. Mao, Y. Zheng, W. Chen, J. Guo and L. Wang, "MiR-30a-5p alleviates LPS-induced HPMEC injury through regulation of autophagy via Beclin-1," *Biocell*, vol. 48, no. 3, p. 431, 2024.

[3] K. L. Kalantar *et al.*, "Integrated host-microbe plasma metagenomics for sepsis diagnosis in a prospective cohort of critically ill adults," *Nature Microbiology*, vol. 7, no. 11, pp. 1805–1816, 2022.

[4] T. van der Poll, M. Shankar-Hari and W. J. Wiersinga, "The immunology of sepsis," *Immunity*, vol. 54, no. 11, pp. 2450–2464, 2021.

[5] K. L. Burnham *et al.*, "Shared and distinct aspects of the sepsis transcriptomic response to fecal peritonitis and pneumonia," *American Journal of Respiratory and Critical Care Medicine*, vol. 196, no. 3, pp. 328–339, 2017.

[6] Q. Su, J. Huang, Y. Zhang, Z. Liu, Z. Lv, C. Zhang, C. Ling, H. Su, L. Zhan and Z. Zhang, "AI-driven discovery of minimal sepsis biomarkers for disease detection and progression: precision medicine across diverse populations," *Frontiers in Medicine*, vol. 12, p. 1521827, 2025.

[7] 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, doi: 10.1186/s12859-023-05134-1.

[8] M. Abbas and Y. El-Manzalawy, "Machine learning based refined differential gene expression analysis of pediatric sepsis," *BMC Medical Genomics*, vol. 13, no. 1, p. 122, 2020.

[9] J. Wang, J. Cai, L. Yue, X. Zhou, C. Hu and H. Zhu, "Identification of potential biomarkers of septic shock based on pathway and transcriptome analyses of immune-related genes," *Genetics Research*, vol. 2023, no. 1, p. 9991613, 2023.

[10] Y. Yang, Y. Zhang, S. Li, X. Zheng, M.H. Wong, K.S. Leung and L. Cheng, "A robust and generalizable immune-related signature for sepsis diagnostics," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 6, pp. 3246–3254, 2021.

[11] K. H. Goh, L. Wang, A. Y. K. Yeow, H. Poh, K. Li, J. J. L. Yeow and G. Y. H. Tan, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature Communications*, vol. 12, no. 1, p. 711, 2021.

[12] W. Xiong, Y. Zhan, R. Xiao and F. Liu, "Advancing sepsis diagnosis and immunotherapy: machine learning-driven identification of stable molecular biomarkers and therapeutic targets," *Scientific Reports*, vol. 15, no. 1, p. 8333, 2025.

[13] H. Wang, L. Len, L. Hu and Y. Hu, "Combining machine learning and single-cell sequencing to identify key immune genes in sepsis," *Scientific Reports*, vol. 15, no. 1, p. 1557, 2025.

[14] W. Zhang, H. Shi and J. Peng, "A diagnostic model for sepsis using an integrated machine learning framework approach and its therapeutic drug discovery," *BMC Infectious Diseases*, vol. 25, no. 1, p. 219, 2025.

[15] 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.

[16] Z. Chen, L. Zeng, G. Liu, Y. Ou, C. Lu, B. Yang and L. Zuo, "Construction of autophagy-related gene classifier for early diagnosis, prognosis and predicting immune microenvironment features in sepsis by machine learning algorithms," *Journal of Inflammation Research*, pp. 6165–6186, 2022.

[17] Q. Zhao, N. Xu, H. Guo and J. Li, "Identification of the diagnostic signature of sepsis based on bioinformatic analysis of gene expression and machine learning," *Combinatorial Chemistry & High Throughput Screening*, vol. 25, no. 1, pp. 21–28, 2022.

[18] T. H. Zhang, M. M. Hasib, Y. C. Chiu, Z. F. Han, Y. F. Jin, M. Flores, Y. Chen and Y. Huang, "Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype pre-

dictions," *Cancers*, vol. 14, no. 19, p. 4763, 2022.
[19] A. Khan and B. Lee, "DeepGene Transformer: Transformer for the gene expression-based classification of cancer subtypes," *Expert Systems with Applications*, vol. 226, p. 120047, 2023.
[20] H. S. Shon, Y. Yi, K. O. Kim, E. J. Cha and K. A. Kim, "Classification of stomach cancer gene expression data using CNN algorithm of deep learning," *Journal of Biomedical and Translational Research*, vol. 20, no. 1, pp. 15–20, 2019.
[21] E. A. T. 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.
[22] B. Y. Al-Mualemi and L. Lu, "A deep learning-based sepsis estimation scheme," *IEEE Access*, vol. 9, pp. 5442–5452, 2020.
[23] Y. Tang, Y. Zhang, and J. Li, "A time series driven model for early sepsis prediction based on transformer module," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 23, 2024.
[24] Z. Zhai, J. Peng, W. Zhong, J. Tao, Y. Ao, B. Niu and L. Zhu, "Identification of key genes and potential therapeutic targets in sepsis-associated acute kidney injury using transformer and machine learning approaches," *Bioengineering*, vol. 12, no. 5, p. 536, 2025.
[25] E. Clough and T. Barrett, "The Gene Expression Omnibus Database," in *Statistical Genomics: Methods and Protocols*, C. E. Gondro, J. van der Werf, and B. Hayes, Eds. New York, NY: Humana Press, 2018, pp. 93–110.
[26] U. Sarkans, A. Füllgrabe, A. Ali, A. Athar, E. Behrangi, N. Diaz, S. Fexova, N. George, H. Iqbal, S. Kurr and J. Munoz, "From ArrayExpress to BioStudies," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1502–D1506, 2021.
[27] M. B. Kursa, A. Jankowski and W. R. Rudnicki, "Boruta–a system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.
[28] X. Zheng *et al.*, "scCaT: An explainable capsulating architecture for sepsis diagnosis transferring from single-cell RNA sequencing," *PLOS Computational Biology*, vol. 20, no. 10, p. e1012083, 2024.
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[31] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj and D. J. Inman, "1D convolutional neural networks and applications: A survey," *Mechanical Systems and Signal Processing*, vol. 151, p. 107398, 2021.
[32] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

Tuan Anh Vu received the degree of Engineer in Information Technology and the M.S degree in Information Systems from the Post and Telecommunications Institute of Technology (PTIT), Hanoi, Viet Nam, 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.

Hoai Bac Dang received the B.Sc. degree in Automation Engineering from Hanoi University of Technology and Science, Hanoi, Vietnam, and the M.Sc. degree in Electronics & Communications Engineering from Posts and Telecommunications Institute of Technology (PTIT), in 1997 and 2003, respectively. In 2008 he achieved the Ph.D. Degree in Electronics & Communications Engineering from PTIT. He is now the director of PTIT. His research interests include digital signal processing in communications, telecommunication transmission and 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 received the B.S. degree from the Post and 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. He can be contacted at email: nmtuan@ptit.edu.vn.