

The way for early diagnosis of sepsis: Read host immune response via pan-infection transcriptomics and transfer learning

Chuanchuan Nan, Jize Xie, Ning Zhang, Guanghao Liu, Xinlei Huang, Yuanfen Xie, Yi Chen, Yushan Qiu ^(✉), Lixin Cheng ^(✉)

Abstract Sepsis, a life-threatening condition resulting from an overactive immune response to infection, necessitates prompt diagnosis to prevent progression. Early detection remains a challenge due to its non-specific symptoms and the need for a thorough assessment to determine the infection's source and severity. Although several studies explore host transcriptome as potential tools for sepsis diagnosis, they are limited by small data volumes or data integration methods for assembling datasets from different resources. Furthermore, infections trigger sepsis and provide early-warning signal at the host immune response level. Therefore, we utilize infection transcriptome data as a source domain to enhance the performance of sepsis diagnosis as a target domain by leveraging deep transfer learning. A pan-infection dataset of 3,686 infectious samples is curated from 37 cohorts using the Pairwise Analysis of Gene Expression (PAGE) method. We train a deep neural network based on the pan-infection data to discriminate between infection and non-infection, and then fine-tune with five sepsis cohorts, constructing a sepsis diagnosis model called sepSeek. sepSeek is validated in three external cohorts, which is superior to common machine learning algorithms and sepsis diagnosis methods. Our framework demonstrates potential for accurate sepsis diagnosis, offering a significant advancement in critical care medicine and cross-disease study.

Keywords Sepsis diagnosis, host response, infection, transcriptome, transfer learning

1 Introduction

Sepsis is a potentially lethal medical condition resulting from an overreaction of the body's immune system to an infection. In response to the infection, the immune system releases a range of pro-inflammatory factors into the bloodstream, which can inadvertently trigger widespread inflammation and potentially lead to organ damage or failure [1–4]. The severity of sepsis requires rapid diagnosis and treatment, which can drastically enhance a patient's survival prospects [5]. However, the early diagnosis of sepsis is challenging, since it requires a high level of clinical suspicion and a comprehensive evaluation to identify the source and severity of infection [4–6].

Recent studies have identified gene signatures as potential biomarkers for sepsis diagnosis [7–10]. For example, SciCluna et al. developed the sNIP score composed of three genes, NLRP1, IDNK, and PLAC8 [7]. The SeptiCyt Lab used two ratio classifiers, each consisting of a ratio of two genes, PLAC8/PLA2G7 and LAMP1/CEACAM4 [8]. Our previous work utilized long non-coding RNAs (lncRNAs) as biomarkers and proposed the PAGE method for data integration, which can overcome batch effects and demonstrate high performance in sepsis diagnosis [11–14]. However, the small data volumes in these studies have hindered accurate and generalizable identification of gene signatures across different

- Chuanchuan Nan and Jize Xie contribute equally to this work.
- Chuanchuan Nan is with Institute of Advanced Interdisciplinary Studies, Chongqing University, Chongqing 400030, China. E-mail: nancsz@163.com
- Jize Xie, Yi Chen are with the Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Hong Kong, 999077, China. Email: jxiebj@connect.ust.hk, yichen@ust.hk
- Ning Zhang, Guanghao Liu, Xinlei Huang, Lixin Cheng are with the Department of Geriatrics, Guangdong Provincial Clinical Research Center for Geriatrics, Shenzhen Clinical Research Center for Geriatrics, Shenzhen People's Hospital, Shenzhen, 518020, China. E-mail: ningzhangicu@outlook.com, jinnana7926@hotmail.com, bioinfo_lgh@163.com, huan_gx1_1999@163.com, easonlcheng@gmail.com
- Yuanfen Xie, Yushan Qiu are with the School of Mathematical Sciences, Shenzhen University, Shenzhen, 518020, China. E-mail: 2310205002@e-mail.szu.edu.cn, yushan.qiu@szu.edu.cn

* To whom correspondence should be addressed.

Manuscript received: 20xx-xx-xx; revised: 20xx-xx-xx; accepted: 20xx-xx-xx

data distributions [15–17].

Transfer learning is an effective and widely-used machine learning framework for addressing the challenges posed by limited data volumes, where a model trained on one task is repurposed on another related task [18–20]. Specifically, transfer learning could manifest as a pre-trained deep neural network trained on a large-scale dataset, and then fine-tuning the network for a specific task with the smaller dataset [21, 22]. Some recent works in bioinformatics have revealed potential applications of transfer learning in biology. For example, Wang et al. developed an interpretable transfer learning model that leverages a pre-trained RNA-proteins binding prediction model to predict the subcellular localization of RNAs [23]. Wang et al. proposed a machine learning framework to effectively reduce regional effects for microbial-based cross-regional diagnosis using transfer learning [24]. These successful attempts of applying transfer learning to bioinformatics and disease diagnosis inspired us to provide a transfer solution to overcome the obstacle of scarce sepsis samples.

Sepsis often progresses from infections, which can yield inflammatory markers, vital signs, and insights into antibiotic utilization relevant to the condition [25–27]. Unlike sepsis samples with limited data volume, existing infection samples have accumulated to a considerable scale. These infection samples can be regarded as the source domain in the context of transfer learning to improve the performance of sepsis diagnosis and enable early detection.

In this work, we aimed to leverage the information obtained from the large and early-onset infection data to assist in the accurate and early diagnosis of sepsis. To this end, we curated a pan-infection host response transcriptome dataset by integrating 37 infection cohorts from nine different countries using PAGE and selected 16 discriminant gene pairs (DGPs) through LASSO analysis. Then, we trained a deep neural network as the infection diagnosis model, which was further fine-tuned with five sepsis cohorts to establish the sepsis diagnosis model called sepSeek. Finally, we evaluated and compared its performance with common machine learning algorithms and state-of-the-art biomarkers using multiple metrics on three external validation cohorts, and the evaluation results consistently demonstrated the advancement and effectiveness of sepSeek.

2 Materials and methods

2.1 Data Collection

After comprehensively searching in the Gene Expression Omnibus (GEO) database [28], we selected 37 cohorts across nine countries, including 3,686 samples with bacterial in-

fection, virus infection, or non-infection (Table 1). Samples lacking complete or consistent information, as well as non-infected samples from individuals with other diseases, were excluded. For instance, as samples in GSE61821 were sequentially profiled on the first day and the 28th day, the latter was excluded from our study. Ultimately, 3,302 samples were integrated as a pan-infection dataset, including 2,206 infected samples and 1,096 non-infected samples. The patients infected with bacteria mainly covered *S.pneumoniae*, *S.aureus*, *E faecium*, *B.pseudomallei*, and *E.coli*. The virus group was mainly infected by H1N1, RSV, and Rhinovirus.

On the other hand, we collected eight sepsis cohorts, including 1,735 sepsis and 291 control samples (Table 2). The datasets GSE26440, GSE65682, GSE66099, GSE69686, and GSE185263 were meticulously selected to constitute the discovery cohorts. Concurrently, the datasets GSE26378, GSE57065, and GSE95233 were designated as the external validation cohorts. The discovery cohorts encompassed a total of 1,469 samples from sepsis patients and 223 control samples, while the external validation cohorts included 266 samples from sepsis patients and 68 control samples.

2.2 PAGE and LASSO for feature selection

To mitigate the batch effects and integrate the data from different cohorts, we used Pairwise Analysis of Gene Expression (PAGE) to select the gene pairs that express differentially between the infected and non-infected samples from the pan-infection dataset [9]. PAGE is based on the relative expression of gene pairs, which has been validated to be more robust than the absolute expression of genes [29–33].

We used the KEGG pathway sub-collection from the Molecular Signatures Database (MSigDB), specifically the curated canonical pathway collection (C2:CP:KEGG), to stratify the detected genes into 186 pathways, and then separately performed gene pair selection [9, 29, 34]. In parallel, PAGE initially calculated the relative expression of each gene pair within each pathway for every sample. Subsequently, gene pairs demonstrating significant differences across samples were filtered out. The raw expression intensity of the i -th gene for sample k was denoted as h_i^k . The relative expression of gene i and gene j within sample k was represented by $r_{i,j}^k$. If h_i^k was greater than h_j^k , $r_{i,j}^k$ was assigned 1; otherwise, it was assigned -1. Positive numbers were represented by a Y^k value of 1, whereas negative numbers were denoted by a Y^k value of 0. Using these representations, the contingency table was calculated as follows.

	$h_i^k > h_j^k$	$h_i^k \leq h_j^k$
Positive	$\frac{1}{2} \sum_{k=1}^n (r_{i,j}^k + 1) \times Y^k$	$\frac{1}{2} \sum_{k=1}^n (1 - r_{i,j}^k) \times Y^k$
Negative	$\frac{1}{2} \sum_{k=1}^n (r_{i,j}^k + 1) \times (1 - Y^k)$	$\frac{1}{2} \sum_{k=1}^n (1 - r_{i,j}^k) \times (1 - Y^k)$



Then, we used Fisher's Exact Test with Bonferroni corrected false discovery rate ($< 1 \times 10^{-16}$) to identify gene pairs with significantly different distributions between the samples of interest and others.

Although PAGE loses part of the quantitative information in absolute expression magnitudes, this trade-off is beneficial in our setting because sepSeek is designed for cross-cohort classification rather than within-cohort expression quantification. Absolute expression values are vulnerable to platform-, batch-, and normalization-dependent variation, whereas within-sample gene-pair ordering provides a scale-free representation that is more robust across heterogeneous datasets. The selected DGPs also preserve biological interpretability because they are derived within pathways. Empirically, PAGE ratios captured the shared infection-sepsis host-response pattern more clearly than normalized individual gene expression values (see Fig. 5), supporting the suitability of this representation for transferable sepsis classification.

In high-dimensional data analysis, where the number of features far exceeds the number of samples, the Least Absolute Shrinkage and Selection Operator (LASSO) is a useful technique for reducing noisy and redundant features [35]. In this study, LASSO was used to further refine the candidate gene pairs obtained from PAGE and to select a compact set of discriminant gene pairs. Let n denote the number of samples and p denote the number of candidate PAGE-derived gene-pair features. Let X be an $n \times p$ PAGE-encoded feature matrix, where X_{ki} represents the encoded state of the i -th gene pair in the k -th sample, and let $y_k \in \{0, 1\}$ denote the phenotype label of sample k . We fitted an L_1 -penalized logistic regression model:

$$(\beta_0, \beta) = \arg \min_{\beta_0, \beta} \left[-\frac{1}{n} \sum_{k=1}^n (y_k \log(\pi_k) + (1 - y_k) \log(1 - \pi_k)) + \lambda \sum_{i=1}^p |\beta_i| \right] \quad (1)$$

where $\pi_k = \sigma(\beta_0 + X_k \beta)$. Here, X_k denotes the k -th row of the feature matrix, π_k denotes the predicted probability that sample k belongs to the infected phenotype, $\sigma(\cdot)$ is the sigmoid function, β_0 is the intercept, β is the vector of regression coefficients for the candidate gene-pair features, and λ is the regularization parameter controlling the strength of the L1 penalty. The penalty shrinks the coefficients of weak or redundant features toward zero, and gene pairs with nonzero coefficients were retained as Discriminant Gene Pairs (DGPs). The implementation was carried out using the Scikit-learn library in the Python environment [36].

2.3 Transfer learning

Transfer learning is a powerful deep learning strategy that transfers knowledge from a source domain to a target domain. Typically, for a new target domain, vanilla transfer learning utilizes a model pre-trained on a source domain as a starting point, rather than training a model from scratch. This transferring idea allows the model to leverage background knowledge from source domains to perform well in the target domain. Inspired by this, we designed sepSeek to achieve accurate sepsis prediction based on large-scale infection knowledge [15, 16]. We define the common transfer learning paradigm as the preliminary knowledge for sepSeek. Given a source domain $D_s = (x_i, y_i)_{i=1}^{N_s} \subset X_s \times Y_s$ and a target domain $D_t = (x_i, y_i)_{i=1}^{N_t} \subset X_t \times Y_t$, where X and Y represent the feature space and label space, N is the number of samples, the goal of transfer learning is to learn a prediction function $f : X_t \mapsto Y_t$ using the source domain data as a priori, such that f minimizes the prediction risk on the target domain.

Introducing the concept of structural risk minimization (in addition to adapting to training data, models have a lower Vapnik-Chervonenkis dimension), the unified framework for transfer learning is as follows:

$$f^* = \arg \min_{f \in H} \left(\frac{1}{N_s} \sum_{i=1}^{N_s} L(f(\nu_i x_i), y_i) + \lambda R(T(D_s), T(D_t)) \right) \quad (2)$$

where H is the hypothesis space, $\nu \in \mathbb{R}^{N_s}$ represents the weights of source domain samples, where $\nu_i \in [0, 1]$, N_s is the number of source samples; $L(\cdot, \cdot)$ is the loss function, measuring the difference between the prediction and the ground truth; $T(\cdot)$ denotes the feature transformation function, λ is the regularization parameter, and $R(\cdot, \cdot)$ is the regularization term, measuring the complexity of the model [37].

2.4 Construction of sepSeek using transfer learning

Equipped with the above preliminary knowledge, we proposed sepSeek to train a reliable sepsis diagnosis model based on large-scale infection data [17–19]. Specifically, we first developed a deep neural network-based model for infection prediction. The architecture of this neural network comprised five dense layers. The first layer utilized the Rectified Linear Unit (ReLU) activation function [38], while the sigmoid function was employed in the final layer. For the intermediate layers, the hyperbolic tangent (tanh) function was used. For the optimizer, we used Adam with a learning rate of 10^{-3} [39]. Our training process involved five epochs using

the infection dataset, at the end of which we saved the model along with its respective weights.

Subsequently, we applied the trained model to the sepsis dataset. We used the same DGPs to construct the sepsis discovery dataset, which was randomly partitioned into a training set and a test set with a ratio of 4:1. Then, we fine-tuned the saved infection diagnosis model using the sepsis training set. During this fine-tuning phase, no layer of the pre-trained neural network was frozen; all five dense layers were set as trainable and updated on the sepsis training set. The pre-trained infection model therefore served as an initialization rather than a fixed feature extractor. The network architecture, activation functions, optimizer, learning rate, and other training hyperparameters were kept consistent with those used in the infection pre-training stage. To address class imbalance in the sepsis cohorts, we used a class-weighted loss during fine-tuning rather than applying oversampling or under sampling. Specifically, the minority control class was assigned a higher weight than the sepsis class, with a control-to-sepsis class-weight ratio of 2:1. This cost-sensitive strategy increased the penalty for misclassifying control samples and reduced the risk that model training would be dominated by the majority sepsis class. Early stopping was applied using the held-out sepsis test set to reduce the risk of overfitting. The five sepsis discovery cohorts contained 1,692 samples in total, including 1,469 sepsis samples and 223 control samples. The DGPs-based sepsis dataset was randomly partitioned into a training set and a held-out test set with a 4:1 ratio. Specifically, 1,353 samples were used for fine-tuning/training and 339 samples were reserved as the test set for early stopping and model evaluation. The model takes the 16 DGPs as input and outputs the probability of sepsis.

2.5 Evaluation metrics

We primarily evaluated the performance of sepSeek using Area Under receiver operating characteristic Curve (AUC). In addition, sensitivity, specificity, F1-score, and accuracy were calculated for the training set, test set, and validation set as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4)$$

$$F1 = \frac{2 \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

where TP represents the number of true positive samples, TN represents the number of true negative samples, FP represents the number of false positive samples, and FN represents the number of false negative samples. High sensitivity indicates that the model is likely to correctly classify positive samples as positive, and high specificity refers to the model being less likely to classify the negative samples as positive. The F1-score ranges from 0 (worst) to 1 (best), with higher values indicating better performance in terms of both the precision and recall. For threshold-dependent metrics, including sensitivity, specificity, accuracy, and F1-score, we used a risk-score threshold of 0.5 as the default operating point because sepSeek outputs a sigmoid-transformed probability score between 0 and 1. The AUC was used as the primary evaluation metric because it is independent of any specific decision threshold. In clinical deployment, the operating threshold can be adjusted according to the intended use case: a lower threshold may be preferred for early screening to increase sensitivity, whereas a higher threshold may be used when higher specificity is required. Therefore, the threshold of 0.5 should be interpreted as a default reporting cutoff rather than a fixed clinical decision rule.

2.6 Visualization methods

We designed a novel method to visualize the association between phenotypes and the identified gene pairs named PAGE plot. The visualization includes two key panels. The first panel illustrates the expression distributions of a pair of genes among samples with different phenotypes. Scatter plot, box plot, and ridge plot are used to show the difference between two phenotypes, e.g., disease and control. Scatter plot intuitively presents the differences between two groups of samples in two dimensions, while box plot and ridge plot reveal the distribution difference of each gene between groups. The contingency table shows the statistics of the number of samples in different expression patterns ($G1 > G2$ or $G1 < G2$) and different groups (disease or control), which is used for Fisher's exact test to evaluate the association between phenotypes and a gene pair. In this study, Fisher's exact test P-value of 0.01 is used as the threshold for differential gene pair identification between the infection and non-infection phenotypes.

To better depict the expression patterns of gene pairs across different cohorts and different phenotypes, we propose the PAGE ratio of samples with specific pairwise ex-

pression pattern ($G1 > G2$ or $G1 < G2$) in each phenotype of a given cohort. PAGE ratio (PR) is defined as

$$PR = \frac{S}{N} \quad (7)$$

where N represents the sample size of a given cohort, while S represents the number of samples in a specific pattern, i.e., the expression level of $G1$ exceeds $G2$. The calculated PAGE ratios are visualized using the R “pheatmap” package, providing an intuitive representation of the pairwise gene expression patterns in a phenotype. Hierarchical clustering analysis is used to group cohorts with similar expression patterns. Additionally, we apply Principal Component Analysis (PCA) to reduce the dimensionality of the PAGE ratios in each cohort, aiming to visualize the similarities between the cohorts in an alternative way.

3 Results

3.1 Study cohorts

Since sepsis is a dysregulated immune response to infection, studying infection is essential for the understanding of sepsis. Infection occurs earlier in the development of sepsis, indicating diagnosis of infection can provide evidence for sepsis detection (Fig. 1A). Moreover, the sample size of current available infection data is much larger than that of sepsis. Therefore, we applied transfer learning regarding infection as an adaptive source domain and sepsis as a target domain (Fig. 1B).

We initiated our study by integrating the host transcriptome data of infection cohorts from the Gene Expression Omnibus (GEO) database. Totally, 3,686 host transcriptome samples across 37 infection cohorts were obtained (Table 1, Fig. 1C). Upon the exclusion of samples with incomplete or conflicting information, the final infection dataset comprised 3,302 samples from nine countries, including USA (10 cohorts), UK (6 cohorts), and Australia (4 cohorts), covered nearly all age groups of infected individuals. Specifically, the integrated pan-infection dataset included 2,278 infected samples and 1,024 non-infected samples (Fig. 1D). Furthermore, we collected eight sepsis cohorts, including 1,735 sepsis and 291 control samples, the size of which is smaller than that of the infection data (Table 2, Fig. 1E). We used these collected infection data and sepsis data as the source domain and target domain, respectively.

3.2 Selection and characterization of discriminant gene pairs

To boost the sample space for training, we integrated all host infection transcriptome cohorts using PAGE, which is a sophisticated strategy previously proposed to mitigate batch

effects and select stable features. In this study, stable features refer to the gene pairs with differential expression between the infected and non-infected samples. Since genes often work together in the same pathway, we concentrated on the pairwise relationship of genes within a pathway to retain the biological significance. Specifically, genes within a pathway were paired and statistically tested using Fisher's exact test. Finally, we identified 14,638 differential gene pairs ($P < 1e-16$, Fisher's exact test) in pathways and then employed LASSO for further refinement, resulting in a final set of 16 Discriminant Gene Pairs (DGPs). The relationship between each pairwise combination of genes and their chromosome positions are depicted in Fig. 3A.

To show the distinct expression pattern among different phenotypes, we proposed PAGE ratio to illustrate the ratio of a pair of genes with a specific pattern (e.g., $G1 > G2$) for each cohort (see Materials and Methods). DGPs reveal obviously different PAGE ratios between the infection and non-infection cohorts (Fig. 3B-C). Importantly, we observed a similar expression trend of these infection DGPs in most of the sepsis cohorts, indicating the cross-disease feature adaptation, which proves the concept of enhancing sepsis diagnosis by studying infection. For instance, RPS6KA2 expressed higher than CCNB2 in 2,052 infection samples, which is over 8.3-fold greater than that of non-infection samples (247 samples). In the sepsis cohorts, the values are 833 for the sepsis phenotype and 104 for the control phenotype, representing an approximate 8.0-fold difference (Fig. 3D).

Moreover, we designed a new visualization method named PAGE plot to clearly illustrate the association between DGPs and phenotypes (Fig. 3E, Figure S1). The scatter plot presents the sepsis and control samples in two dimensions, which are the expression value of a gene pair. Samples in the sepsis or control groups tend to cluster separately, indicating the reverse expression pattern of the gene pair is associated with sepsis. Samples of different phenotypes are distributed differently in at least one gene of a DGP (Fig. 3E, box and ridge plots). A 2×2 contingency table was also provided for Fisher's exact test to statistically evaluate whether the gene pair is associated with the specific phenotype of sepsis.

3.3 Construction of sepsis diagnostic model via transfer learning

We trained a deep neural network for infection prediction using these DGPs, based on an infection training set ($n=2,642$) and a test set ($n=660$) from the pan-infection dataset. The design of our neural network architecture is described in Materials and Methods. This process served as a pre-training of

the sepsis model to enable better understanding and early diagnosis of sepsis. Subsequently, we constructed the sepsis diagnosis model based on the pre-trained infection model. The raw sepsis data comprised eight cohorts. Five of these cohorts were used as the fine-tuning set, while the other three cohorts were used as the external validation set (see Materials and methods).

The sepsis fine-tuning dataset was partitioned into a training set and a test set with a 4:1 ratio. The fine-tuning process was instrumental in the refinement of the infection diagnosis model, ultimately leading to the creation of the sepSeek model—a specialized tool for the diagnosis of sepsis (Fig. 2). We repurposed the infection diagnosis model by retraining it on the sepsis training dataset, while maintaining all other network parameters constant. The training regimen was carefully monitored, with the termination criterion set as the convergence of loss in the sepsis test set, a method commonly referred to as early stopping. This rigorous process culminated in the development of the sepsis diagnostic model, aptly named sepSeek, which was designed with a dual focus on timeliness and precision in diagnostic accuracy.

3.4 Performance evaluation

In a rigorous assessment, we meticulously evaluated the performance of the sepSeek model against the state-of-the-art models and biomarkers, utilizing sepsis training, test, and validation datasets. The comparative analysis was conducted using two categories of benchmarks: 1) prevalent machine learning algorithms enhanced with our identified DGPs, and 2) established sepsis diagnosis models, namely sNIP and SeptiCyte. The objectives of these comparisons were twofold: to validate the efficacy of the DGPs in accurately identifying sepsis, and to establish the superiority of the transfer learning model over traditional machine learning approaches.

sepSeek achieved an AUC of 0.97 on the validation set (Fig. 4A). Other models equipped with the same features as sepSeek, including DNN (without pre-training), ElasticNet, SVM, and RF, also performed reasonably well, albeit slightly inferior to sepSeek. Moreover, sepSeek achieved the highest accuracy and F1-score, second highest specificity, and comparable sensitivity, indicating its excellent and comprehensive performance in different metrics. In contrast, the other machine learning methods generally showed lower specificity, accuracy, and F1-score. Regarding the performance of the two recently proposed gene signatures, sNIP recorded an AUC of 0.87 in the validation set, with a sensitivity of 0.962, specificity of 0.191, F1-score of 0.764, and accuracy

of 0.805. SeptiCyte yielded an AUC of 0.74, sensitivity of 0.507, specificity of 0.632, F1-score of 0.577, and accuracy of 0.530. These results clearly demonstrated the inferiority of sNIP and SeptiCyte compared to models utilizing DGPs. See Table S1 for more detailed comparison. Additionally, the risk score generated by the pre-trained infection model was distributed separately for the infection and the non-infection samples across the 37 infection cohorts with a clear threshold of 0.5 (Fig. 4B). Similarly, sepsis patients tend to have higher sepSeek scores compared to the control ones across the eight sepsis cohorts with a threshold of 0.5 (Fig. 4C).

Overall, our findings highlighted the generalization of the selected features, which outperformed those used in previous models. Moreover, the superior performance of sepSeek underscored the effectiveness of our transfer learning-based approach in leveraging the potential of these features compared to conventional machine learning methods. Besides, the consistent threshold of the risk score across different cohorts indicates the potential of sepSeek for clinic application.

3.5 Homogeneity between infection and sepsis

Changes in gene expression patterns are prominent features in the progression of disease. However, the transformation of gene pairs in infection and sepsis remains unclear. To this end, we calculated the PAGE ratios of the 16 DGPs in infection, non-infection, sepsis, and non-sepsis groups. The clustering results revealed that the infection and sepsis groups cluster together, revealing similar expression patterns between the infection and sepsis samples (Fig. 5A). In contrast, if using the normalized gene expression values of these genes, the infection and sepsis groups were separated (Fig. 5B).

PCA analysis also draws the same conclusion (Fig. 5C-D). The PAGE ratios of gene pairs effectively clustered the two control groups separately from infection and sepsis, whereas the four types of cohorts are not separated from each other when using individual gene expression. This result suggests that the homogeneity between infection and sepsis was well captured by the relative expression of gene pairs rather than the absolute expression of genes.

4 Discussion

Sepsis is a dysregulated immune response to infection, encompassing a broader condition that affects larger populations. Investigating infections can aid in understanding sepsis, as infection typically precedes sepsis in its development. Diagnosing an infection can also serve as an important indicator for the early detection of sepsis. Existing methodologies used to extract biomarkers directly from sepsis datasets,

resulting in the identified gene signatures lack the necessary robustness due to lack of a priori and limitation of samples. In this study, we aimed to overcome these challenges through transfer learning – pan-infection data were used as the source domain and its information was leveraged to develop a robust sepsis diagnosis model. Evaluation of the performance in different metrics demonstrated the superiority of sepSeek to existing methods.

Transfer learning is a popular framework in deep learning and has recently shown potential in bioinformatics [16, 22, 23]. Its ability to extract useful information from large data and transfer the information to smaller data is especially suitable for many clinical scenarios. It is intuitive to transfer information from infection to the sepsis. Both domains exhibit shared underlying characteristics, encompassing physiological parameters and clinical manifestations. This overlap in feature spaces facilitates effective knowledge transfer, as the learned representations from the source domain can be adapted and fine-tuned to the target domain, reducing the need for extensive labeled data in the target domain.

Biologically, sepsis is a dysregulated host response to infection, characterized by life-threatening organ dysfunction [40, 41]. As infections are often the precipitating events for sepsis, there is a strong temporal and causal relationship between the two domains. The pathophysiological processes involved in infections, such as inflammatory responses, immune system activation, and organ dysfunction, share commonalities with sepsis and can serve as early indicators of sepsis development. By leveraging information from infection data, the transfer learning model can potentially capture relevant patterns and biomarkers that are indicative of sepsis progression, enabling early diagnosis and timely intervention. The high performance of sepSeek is attributed to the synergistic effects of feature selection and model architecture optimization.

Beyond their statistical robustness, the selected DGPs also provide biological insights into the host-response transition from infection to sepsis. Sepsis is characterized by dysregulated innate and adaptive immune responses [42], involving excessive inflammation, immunosuppression, altered immune-cell function, and impaired immune homeostasis. Several representative high-weight DGPs are related to immune activation, antigen presentation, MAPK/stress signaling, and cell-cycle regulation, which are central components of infection-induced systemic inflammation. For example, the CD38–CD1C pair in the hematopoietic cell lineage pathway may reflect a shift between inflammatory immune activation and antigen-presentation-related immune

regulation. CD38 is an immune-associated molecule with enzymatic and receptor-like functions, and has been implicated in NAD metabolism, calcium signaling, immune-cell recruitment, cytokine responses, and inflammatory regulation [43]. CD1C belongs to the CD1 family of lipid-antigen-presenting molecules, which present lipid antigens to T-cell subsets and contribute to host immune recognition [44]. Because dendritic-cell antigen presentation and dendritic-cell dysfunction are closely involved in sepsis-induced immunosuppression [45], the relative expression switch between CD38 and CD1C may capture an imbalance between inflammatory activation and antigen-presenting immune function. Another representative DGP, RPS6KA2–CCNB2, may reflect the balance between MAPK-associated signaling and cell-cycle regulation. RPS6KA2 encodes a member of the RSK family, which functions downstream of the Ras/ERK-MAPK pathway and participates in cellular growth, differentiation, survival, and related signaling processes [46]. CCNB2 encodes cyclin B2, a cell-cycle regulator involved in the G2/M transition [47]. Therefore, the consistent relative expression pattern of RPS6KA2 and CCNB2 across infection and sepsis cohorts may reflect a host-response state in which MAPK-associated signaling is relatively prioritized over proliferative cell-cycle programs. These interpretations are hypothesis-generating and provide biological context for the diagnostic signature, while further experimental studies will be required to validate the causal roles of these gene-pair relationships in sepsis progression.

Importantly, although sepSeek was developed and validated using host transcriptomic datasets, its future clinical implementation would not necessarily require whole-transcriptome RNA-seq or other high-turnaround-time transcriptomic technologies. Once the model is established, sepSeek only requires the relative expression states of a compact 16-DGP panel. In a practical detection workflow, the genes composing these DGPs could be measured using a targeted RT-qPCR, instead of high-throughput technologies, which is exactly the advantage of PAGE. PAGE encodes relative gene-pair ordering rather than absolute expression magnitude, which may reduce the impact of platform-dependent measurement scale differences and facilitate translation to rapid targeted platforms. Thus, sepSeek could potentially be applied in ICU to provide an early host-response risk assessment and timely clinical decision-making. Nevertheless, prospective ICU validation, assay calibration, and workflow integration will be required before routine bedside implementation.

Declaration of interests

The authors declare no competing interests.

Reference

- [1] D. Annane, E. Bellissant, and J. M. Cavaillon, Septic shock, *The Lancet*, vol. 365, no. 9453, pp. 63–78, 2005.
- [2] M. Singer, C. S. Deutschman, C. W. Seymour, *et al.*, The third international consensus definitions for sepsis and septic shock (Sepsis-3), *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.
- [3] J. Cohen, J. L. Vincent, N. K. Adhikari, *et al.*, Sepsis: a roadmap for future research, *The Lancet Infectious Diseases*, vol. 15, no. 5, pp. 581–614, 2015.
- [4] Z. Li, B. Wu, J. Chen, *et al.*, WWP2 protects against sepsis-induced cardiac injury through inhibiting cardiomyocyte ferroptosis, *Journal of Translational Internal Medicine*, vol. 12, no. 1, pp. 35–50, 2024.
- [5] C. F. Duncan, T. Youngstein, M. D. Kirrane, and D. O. Lonsdale, Diagnostic challenges in sepsis, *Current Infectious Disease Reports*, vol. 23, no. 12, p. 22, 2021.
- [6] R. Gauer, D. Forbes, and N. Boyer, Sepsis: diagnosis and management, *American Family Physician*, vol. 101, no. 7, pp. 409–418, 2020.
- [7] B. P. Scicluna, M. A. Wiewel, L. A. Van Vught, *et al.*, Molecular biomarker to assist in diagnosing abdominal sepsis upon ICU admission, *American Journal of Respiratory and Critical Care Medicine*, vol. 197, no. 8, pp. 1070–1073, 2018.
- [8] L. McHugh, T. A. Seldon, R. A. Brandon, *et al.*, 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*, vol. 12, no. 12, e1001916, 2015.
- [9] X. Zheng, D. Meng, D. Chen, *et al.*, scCaT: an explainable capsulating architecture for sepsis diagnosis transferring from single-cell RNA sequencing, *PLOS Computational Biology*, vol. 20, no. 10, e1012083, 2024.
- [10] P. Huang, Y. Liu, Y. Li, *et al.*, Metabolomics- and proteomics-based multi-omics integration reveals early metabolite alterations in sepsis-associated acute kidney injury, *BMC Medicine*, vol. 23, p. 79, 2025.
- [11] X. Zheng, K. S. Leung, M. H. Wong, and L. Cheng, Long non-coding RNA pairs to assist in diagnosing sepsis, *BMC Genomics*, vol. 22, no. 1, p. 275, 2021.
- [12] X. Liu, C. Hong, Y. Jiang, *et al.*, Co-expression module analysis reveals high expression homogeneity for both coding and non-coding genes in sepsis, *BMC Genomics*, vol. 24, no. 1, p. 418, 2023.
- [13] X. Liu, Y. Xu, R. Wang, *et al.*, A network-based algorithm for the identification of moonlighting noncoding RNAs and its application in sepsis, *Briefings in Bioinformatics*, vol. 22, no. 1, pp. 581–588, 2021.
- [14] X. Liu, X. Zheng, J. Wang, N. Zhang, K. S. Leung, X. Ye, and L. Cheng, A long non-coding RNA signature for diagnostic prediction of sepsis upon ICU admission, *Clinical and Translational Medicine*, vol. 10, no. 3, e. 123, 2020.
- [15] S. Liu, W. Zhao, X. Liu, and L. Cheng, Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-cohort microbial signatures and potential therapeutic target, *The FASEB Journal*, vol. 34, no. 11, pp. 14166–14181, 2020.
- [16] R. Wang, X. Zheng, J. Wang, *et al.*, Improving bulk RNA-seq classification by transferring gene signature from single cells in acute myeloid leukemia, *Briefings in Bioinformatics*, vol. 23, no. 2, bbac002, 2022.
- [17] S. Yu, H. Yang, B. Wang, X. Guo, L. Li, and Y. Sun, Nomogram for predicting risk of mild renal dysfunction among general residents from rural Northeast China, *Journal of Translational Internal Medicine*, vol. 12, no. 3, pp. 244–252, 2024.
- [18] K. Weiss, T. M. Khoshgoftaar, and D. Wang, A survey of transfer learning, *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.
- [19] F. Zhuang, Z. Qi, K. Duan, *et al.*, A comprehensive survey on transfer learning, *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
- [20] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, A survey on deep transfer learning, in *Artificial Neural Networks and Machine Learning–ICANN 2018*, Rhodes, Greece, 2018, pp. 270–279.
- [21] N. Jin, C. Nan, W. Li, *et al.*, PAGE-based transfer learning from single-cell to bulk sequencing enhances model generalization for sepsis diagnosis, *Briefings in Bioinformatics*, vol. 26, no. 1, bbae661, 2025.
- [22] D. Meng, Y. Feng, K. Yuan, Z. Yu, Q. Cao, L. Cheng, and X. Zheng, scMMAE: masked cross-attention network for single-cell multimodal omics fusion to enhance unimodal omics, *Briefings in Bioinformatics*, vol. 26, no. 1, bbaf010, 2025.
- [23] J. Wang, M. Horlacher, L. Cheng, and O. Winther, DeepLocRNA: an interpretable deep learning model for predicting RNA subcellular localization with domain-specific transfer-learning, *Bioinformatics*, vol. 40, no. 2, btae065, 2024.
- [24] N. Wang, M. Cheng, and K. Ning, Overcoming regional limitations: transfer learning for cross-regional microbial-based diagnosis of diseases, *Gut*, vol. 72, no. 10, pp. 2004–2006, 2023.
- [25] P. Póvoa, L. Coelho, F. Dal-Pizzol, *et al.*, How to use biomarkers of infection or sepsis at the bedside: guide to clinicians, *Intensive Care Medicine*, vol. 49, no. 2, pp. 142–153, 2023.
- [26] E. L. Tsalik, L. B. Jaggars, S. W. Glickman, *et al.*, Discriminative value of inflammatory biomarkers for suspected sepsis, *The Journal of Emergency Medicine*, vol. 43, no. 1, pp. 97–106, 2012.
- [27] S. Kang and R. Zheng, Distribution of the causes of fever of unknown origin in China, 2013–2022, *Journal of Translational Internal Medicine*, vol. 12, no. 3, p. 299, 2024.
- [28] R. Edgar, M. Domrachev, and A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data

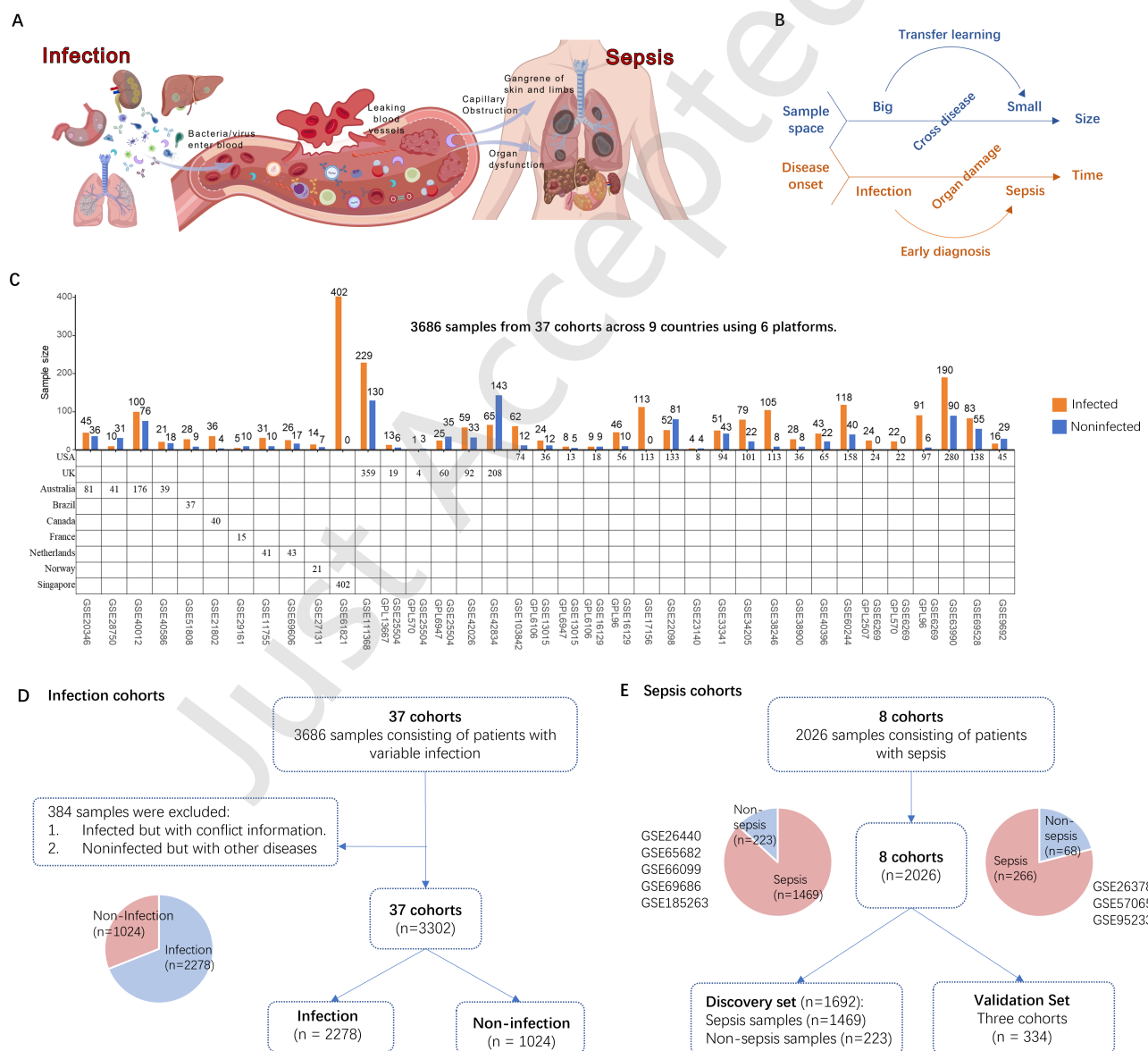
- repository, *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.
- [29] Q. Li, X. Zheng, J. Xie, *et al.*, bvnGPS: a generalizable diagnostic model for acute bacterial and viral infection using integrative host transcriptomics and pretrained neural networks, *Bioinformatics*, vol. 39, no. 3, btad109, 2023.
- [30] L. Cheng, H. Wu, X. Zheng, *et al.*, GPGPS: a robust prognostic gene pair signature of glioma ensembling IDH mutation and 1p/19q co-deletion, *Bioinformatics*, vol. 39, no. 1, btac850, 2023.
- [31] Q. Wu, X. Zheng, K. S. Leung, M. H. Wong, S. K. W. Tsui, and L. Cheng, meGPS: a multi-omics signature for hepatocellular carcinoma detection integrating methylome and transcriptome data, *Bioinformatics*, vol. 38, no. 14, pp. 3513–3522, 2022.
- [32] J. Xie, X. Zheng, J. Yan, *et al.*, Deep learning model to discriminate diverse infection types based on pairwise analysis of host gene expression, *iScience*, vol. 27, no. 6, 2024.
- [33] N. Zhang, F. Yang, P. Zhao, *et al.*, MrGPS: an m6A-related gene pair signature to predict the prognosis and immunological impact of glioma patients, *Briefings in Bioinformatics*, vol. 25, no. 1, bbad498, 2024.
- [34] A. Subramanian, P. Tamayo, V. K. Mootha, *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545–15550, 2005.
- [35] J. Friedman, T. Hastie, and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.
- [36] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, Scikit-learn: machine learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [37] J. Wang and Y. Chen, *Introduction to Transfer Learning: Algorithms and Practice*. Cham, Switzerland: Springer Nature, 2023.
- [38] V. Nair and G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in *Proc. 27th Int. Conf. Machine Learning (ICML-10)*, 2010, pp. 807–814.
- [39] D. P. Kingma, Adam: a method for stochastic optimization, *arXiv preprint arXiv:1412.6980*, 2014.
- [40] Z. Li, B. Wu, J. Chen, *et al.*, WWP2 protects against sepsis-induced cardiac injury through inhibiting cardiomyocyte ferroptosis, *Journal of Translational Internal Medicine*, vol. 12, no. 1, pp. 35–50, 2024.
- [41] L. Zhang, J. Liu, S. Mu, *et al.*, Unfractionated heparin improves coagulation in sepsis by protecting glycocalyx of endothelial cells through inhibiting heparinase, *Journal of Translational Internal Medicine*, vol. 13, no. 6, pp. 618–621, 2025.
- [42] A. M. Kashif, Y. Ouyang, Y. Li, and B. Pan, NETosis and pyroptosis of immune cells in sepsis, *Journal of Translational Internal Medicine*, vol. 13, no. 4, pp. 318–327, 2025.
- [43] Z. L. Piedra-Quintero, Z. Wilson, P. Nava, and M. Guerau-de-Arellano, CD38: an immunomodulatory molecule in inflammation and autoimmunity, *Frontiers in Immunology*, vol. 11, p. 597959, 2020.
- [44] L. Van Kaer, L. Wu, and S. Joyce, Mechanisms and consequences of antigen presentation by CD1, *Trends in Immunology*, vol. 37, no. 11, pp. 738–754, 2016.
- [45] M. Bouras, K. Asehnoune, and A. Roquilly, Contribution of dendritic cell responses to sepsis-induced immunosuppression and to susceptibility to secondary pneumonia, *Frontiers in Immunology*, vol. 9, p. 2590, 2018.
- [46] Y. Romeo, X. Zhang, and P. P. Roux, Regulation and function of the RSK family of protein kinases, *Biochemical Journal*, vol. 441, no. 2, pp. 553–569, 2012.
- [47] T. Wu, X. Zhang, X. Huang, Y. Yang, and X. Hua, Regulation of cyclin B2 expression and cell cycle G2/m transition by menin, *Journal of Biological Chemistry*, vol. 285, no. 24, pp. 18291–18300, 2010.

Table 1 The infection cohorts used in this study.

Data Accession	Cohort Details	Pathogen	Platform	Total Samples	Selected Noninfected Samples	Selected Infected Samples	Country
GSE6269	Children	Influenza A, Influenza B, S.pneumoniae,S.aureus,MSSA	GPL2507	24	0	24	USA
GSE6269	Both	Influenza A, Influenza B, S.pneumoniae,S.aureus,MSSA	GPL570	22	0	22	USA
GSE6269	Both	S.aureus,MSSA,E.coli	GPL96	97	6	91	USA
GSE9692	Children	Neisseria,Group B Strep,Group A Strep, Adenovirus	GPL570	45	29	16	USA
GSE11755	Children		GPL570	41	10	31	Netherlands
GSE13015	Adult	B.pseudomallei,E.coli,	GPL6106	67	12	24	USA
GSE13015	Adult	E.coli,B.pseudomallei	GPL6947	39	5	8	USA
GSE16129	Children	Bacteremia,Osteomyelitis	GPL6106	18	9	9	USA
GSE16129	Both	Bacteremia,Osteomyelitis	GPL96	56	10	46	USA
GSE17156	Adult	Rhinovirus	GPL571	113	0	113	USA
GSE20346	Adult	Bacteria,virus	GPL6947	81	36	45	Australia
GSE21802	Adult	H1N1	GPL6102	40	4	36	Canada
GSE22098	Adult	Mycobacterium tuberculosis	GPL6947	274	81	52	USA
GSE23140		Streptococcus pneumoniae	GPL6254	8	4	4	USA
GSE25504	Children	E.coli, virus,Group Strep,	GPL13667	20	6	13	UK
GSE25504	Children	Enterovirus	GPL570	5	3	1	UK
GSE25504	Children	Ecoli,coag neg staph, E faecium	GPL6947	63	35	25	UK
GSE27131	Adult	H1N1	GPL6244	21	7	14	Norway
GSE28750	Adult		GPL570	41	31	10	Australia
GSE29161		Endocarditis	GPL6480	15	10	5	France
GSE33341	Adult	E.coli,S.aureus	GPL571	94	43	51	USA
GSE34205	Children	RSV	GPL570	101	22	79	USA
GSE38246		DENV	GPL15615	113	8	105	USA
GSE38900	Children	RSV	GPL10558	36	8	28	USA
GSE40012	Adult	Influenza A	GPL6947	190	76	100	Australia
GSE40396	Children	Adenovirus,HHV6,Rhinovirus	GPL10558	65	22	43	USA
GSE40586		Spn	GPL6244	39	18	21	Australia
GSE42026	Children	Gram positive bacteria,H1N1,RSV	GPL6947	92	33	59	UK
GSE42834			GPL10558	356	143	65	UK
GSE51808		DENV	GPL13158	56	9	28	Brazil
GSE60244	Adult	Bacteria,virus	GPL10558	158	40	118	USA
GSE61821	Adult	H1N1	GPL10558	402	0	402	Singapore
GSE63990	Adult	Bacteria,virus	GPL571	280	90	190	USA
GSE69606			GPL570	43	17	26	Netherlands
GSE69528	Adult	B.pseudomallei,mellitus	GPL10558	138	55	83	USA
GSE103842	Children	RSV	GPL10558	74	12	62	USA
GSE111368	Adult	H1N1,H3N2,influenza	GPL10558	359	130	229	UK

Table 2 The sepsis cohorts used in this study.

Data Access	Sample number	Sepsis	Healthy	Platform	Details	Country
<i>Discovery Set</i>						
GSE26440	130	98	32	GPL570	Pediatric	US
GSE65682	802	760	42	GPL13667	Adult	Malta
GSE66099	246	199	47	GPL570	Adult	US
GSE69686	122	64	58	GPL20292	Infant	US
GSE185263	392	348	44	GPL16791	Adult	Canada
<i>Validation Set</i>						
GSE26378	103	82	21	GPL570	Pediatric	US
GSE57065	107	82	25	GPL570	Adult	France
GSE95233	124	102	22	GPL570	Adult	France

**Fig. 1** Motivation and cohorts of this study. (A) Sepsis is a serious complication of an infection. (B) Rationale of transferring information from infection data to sepsis data. (C) Infection and sepsis cohorts used in this study. (D) Inclusion criterion and data stratification of the infection samples. (E) Data stratification of the sepsis samples.

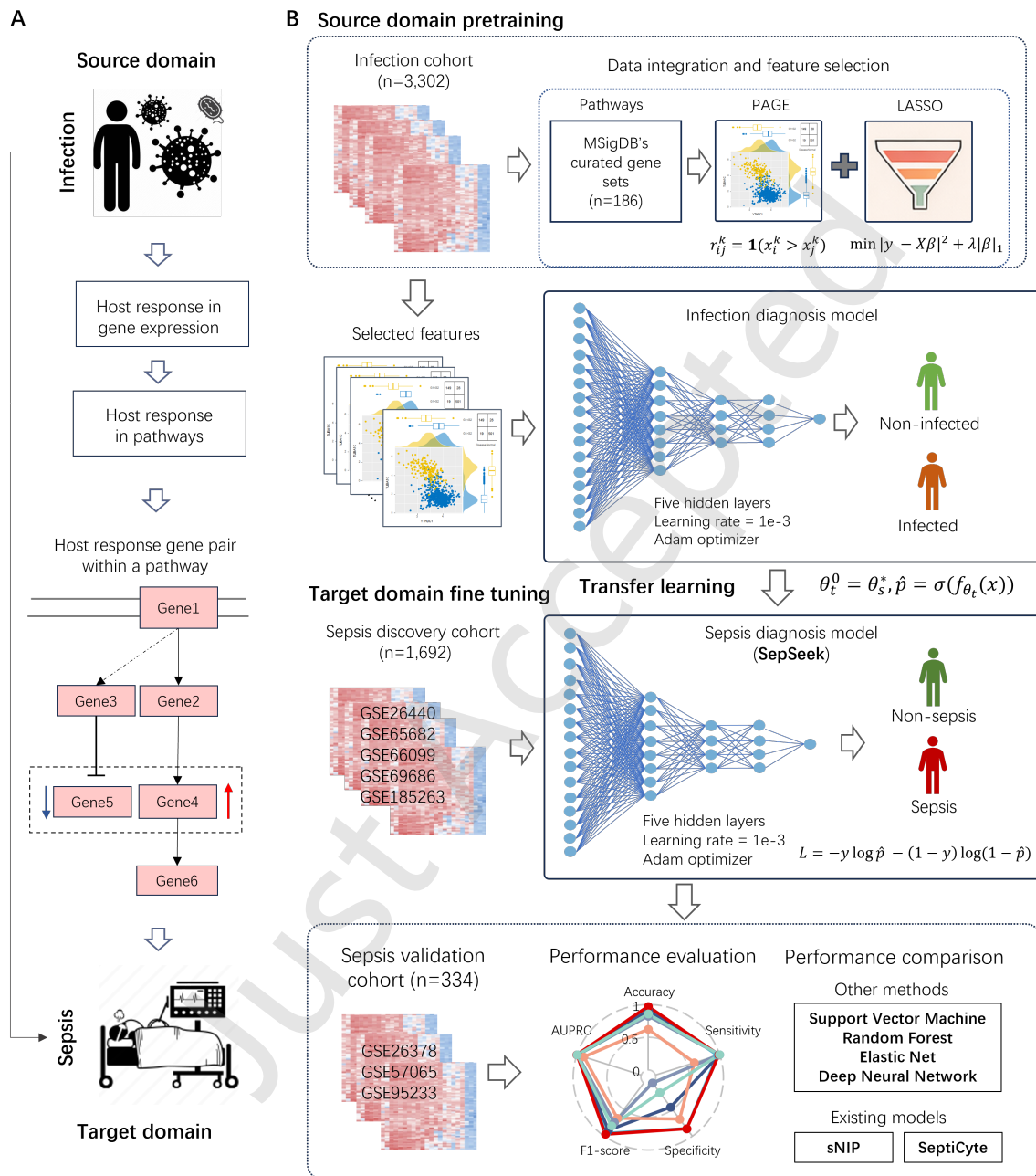


Fig. 2 Study design. (A) Biological rationale for transfer learning from infection to sepsis. Any type of infection can lead to sepsis, with dysregulated host response and the presence of organ dysfunction. (B) Transfer learning workflow. In the source domain, we integrated forty infection cohorts and used PAGE and LASSO to select DGPs. Then we trained a deep neural network with five hidden layers to discriminate infection and non-infection. In the target domain, we used the selected DGPs from the source domain to construct training set from five sepsis cohorts, and fine-tuned the trained model from the source domain. Finally, we evaluated the performance of our model in three external validation cohorts.

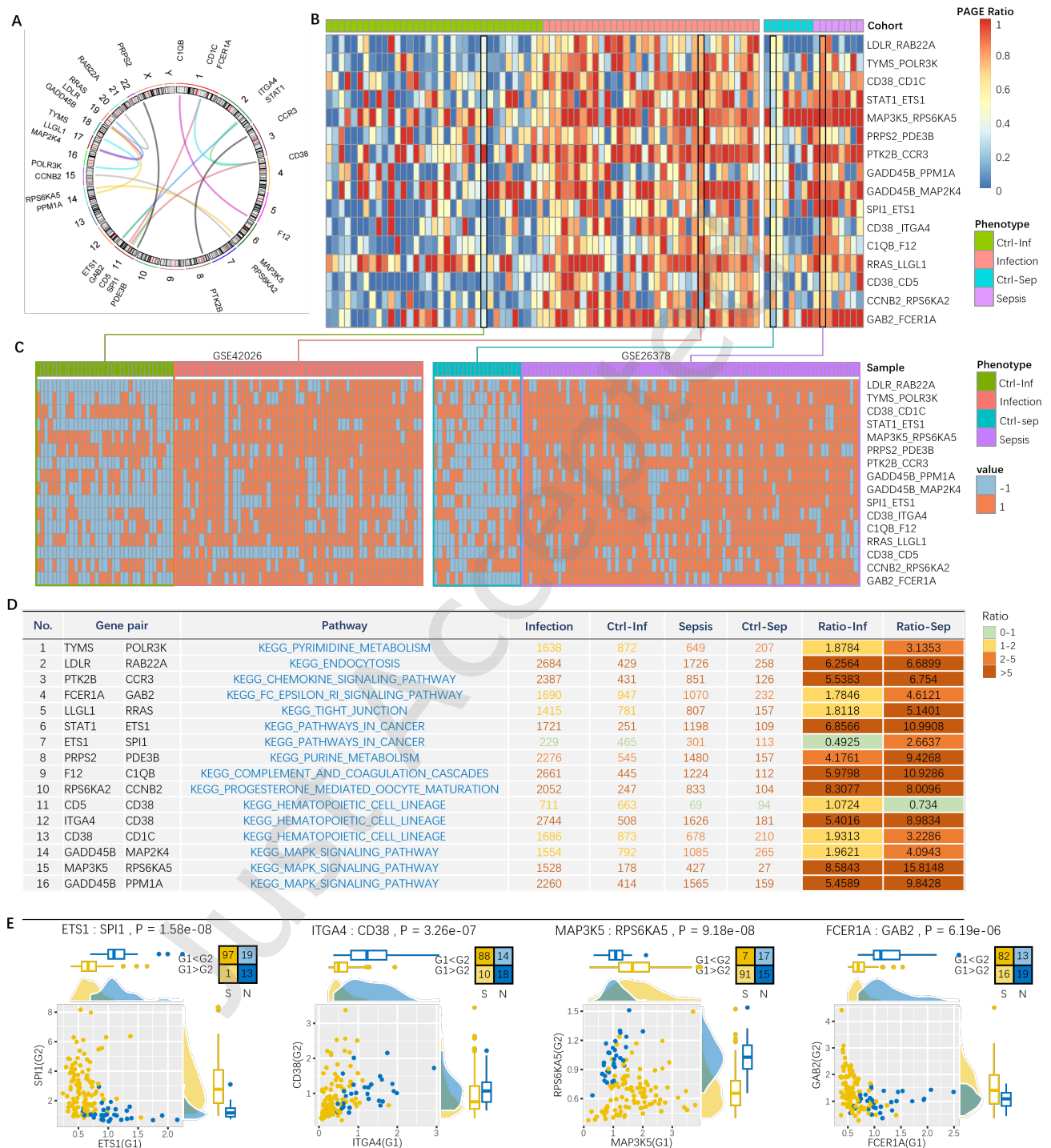


Fig. 3 Characterization of the differential gene pairs in sepSeek. (A) Genome location and the pairwise relationship of the DGPs in sepSeek. (B) Heatmap showing the PAGE ratio in the infection and sepsis cohorts. Grid color represents the ratio of a gene pair, in which the first gene expressed higher than the second one, in each cohort. (C) Heatmaps showing the gene pair expression pattern in GSE42026 and GSE26378, respectively. (D) Functional and expression characterization of DGPs. (E) PAGE plot of four gene pairs identified from GSE26440. Sepsis samples are colored in yellow while control samples are in blue.

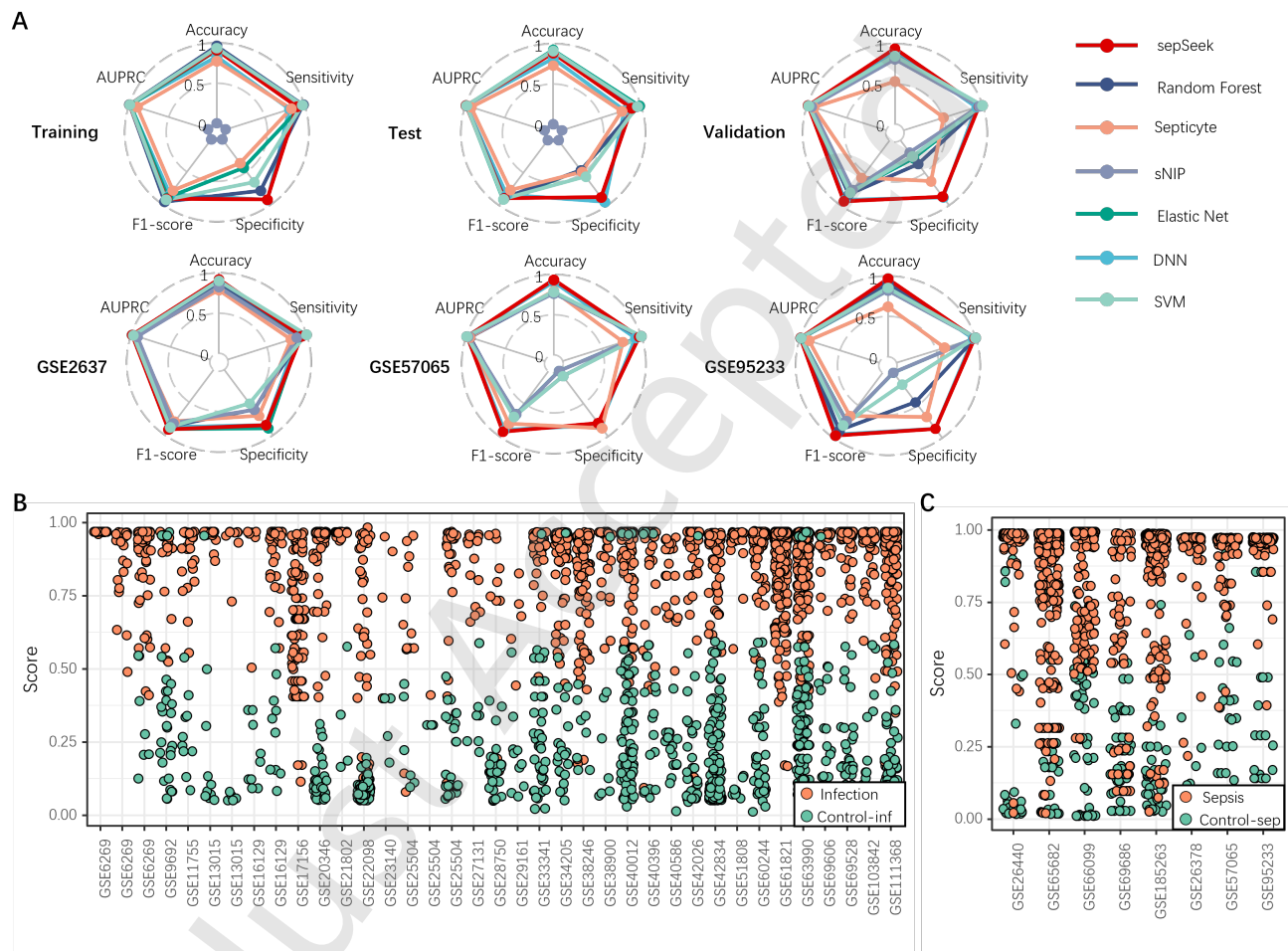


Fig. 4 Performance evaluation of sepSeek. (A) Five evaluation metrics of sepSeek was compared with common machine learning algorithms and SOTA methods in the training set, test set, and validation set, respectively (first line). The performance on the three validation cohorts GSE2637, GSE57065 and GSE95233 is shown in the second line. Machine learning algorithms include DNN, Random Forest, Elastic Net, and SVM. SOTA methods include SeptiCyte and sNIP. (B) Distributions of risk scores computed by the infection forest across the 37 infection cohorts. (C) Distributions of the sepSeek risk scores across the eight sepsis cohorts. Orange and green nodes represent case (infection or sepsis) and control samples, respectively.

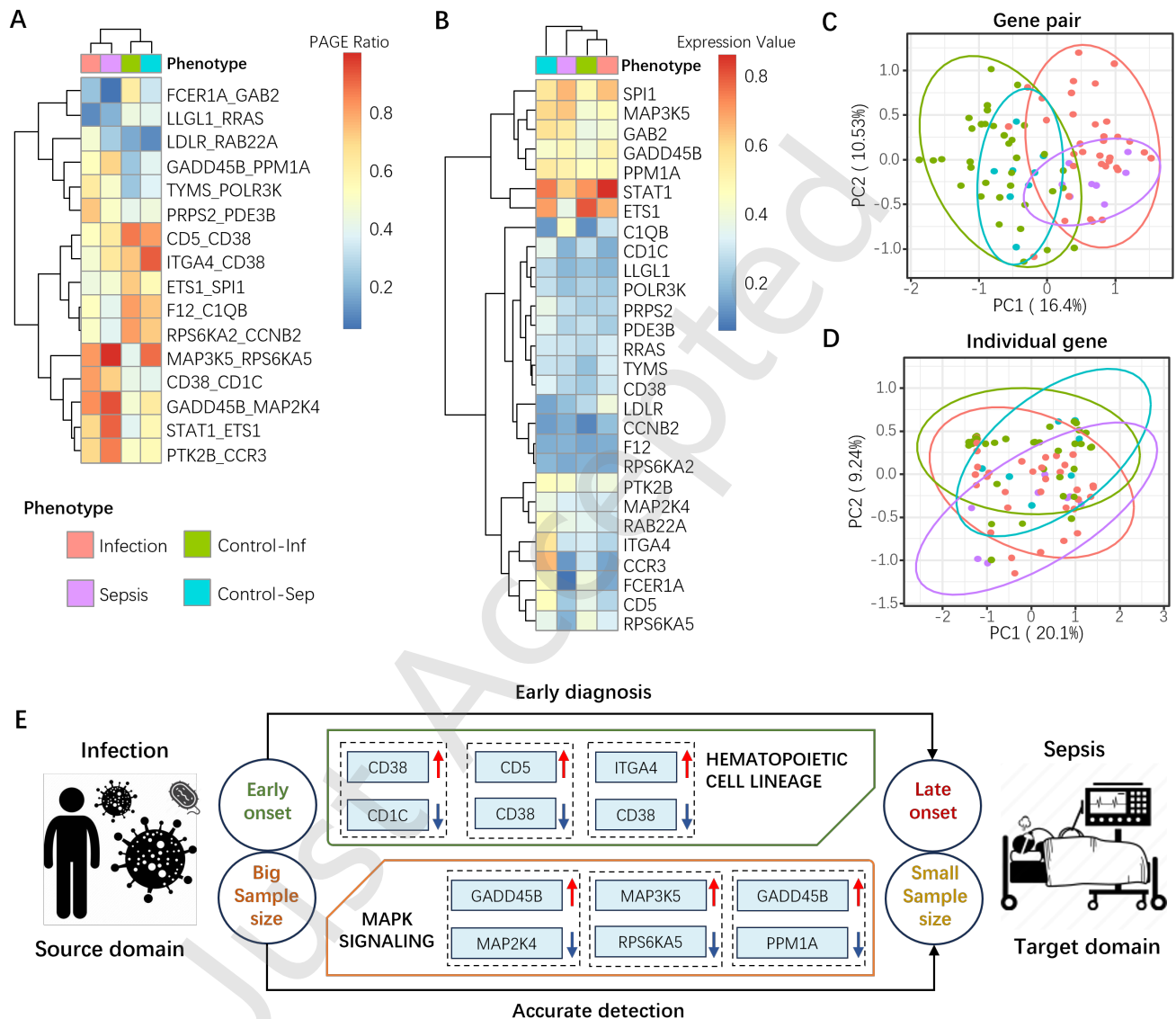


Fig. 5 Generalization of the DGPs across infection and sepsis cohorts. Heatmaps showing the infection and sepsis cohorts have similar expression pattern in gene pairs (A) but not in individual genes (B). Grid color represents PAGE ratio of samples with gene pair (G1 and G2) with $G1 > G2$ in a given phenotype in Figure A, while grid color represents average gene expression value of samples in a given phenotype in Figure B. PAGE ratio was calculated as the proportion of samples in a given category in which the expression of G1 exceeded that of G2. Hierarchical clustering was performed using Euclidean distance and the complete-linkage method. (C) PCA of gene pairs showing the infection and sepsis cohorts are close and separated from the controls. (D) PCA of individual genes showing the four types of cohorts are not separated from each other. Spots correspond to cohorts while colors correspond to states. (E) Gene pairs and pathways responsible for transfer learning from infection to sepsis.