2023

Application of a gene modular approach for clinical phenotype genotype association and sepsis prediction using machinApplication of a gene modular approach for clinical phenotype genotype association and sepsis prediction using machine learning in meningococcal sepsise learning in meningococcal sepsis

Asrar Rashid

Arif Anwary

Feras Al-Obeidat

Joe Brierley

Mohammed Uddin

See next page for additional authors

Authors

Asrar Rashid, Arif Anwary, Feras Al-Obeidat, Joe Brierley, Mohammed Uddin, Hoda Alkhzaimi, Amrita Sarpal, Mohammed Toufiq, Zainab Malik, Raziya Kadwa, Praveen Khilnani, M Guftar Shaikh, Govind Benakatti, Javed Sharief, Syed Ahmed Zaki, Abdulrahman Zeyada, Ahmed Al-Dubai, Wael Hafez, and Amir Hussain

Informatics in Medicine Unlocked 41 (2023) 101293

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Application of a gene modular approach for clinical phenotype genotype association and sepsis prediction using machine learning in meningococcal sepsis

Asrar Rashid [a,b,*], Arif R. Anwary [a], Feras Al-Obeidat [c], Joe Brierley [d], Mohammed Uddin [e], Hoda Alkhzaimi [f], Amrita Sarpal [g,h], Mohammed Toufiq [i], Zainab A. Malik [e,j], Raziya Kadwa [b], Praveen Khilnani [k], M Guftar Shaikh [l], Govind Benakatti [m], Javed Sharief [b], Syed Ahmed Zaki [n], Abdulrahman Zeyada [b], Ahmed Al-Dubai [a], Wael Hafez [b,o], Amir Hussain [a]

[a] Edinburgh Napier University, Merchiston Campus, 10 Colinton Road, Edinburgh, Scotland, EH10 5DT, UK
[b] NMC Royal Khalifa Hospital, Abu Dhabi, United Arab Emirates
[c] College of Technological Innovation at Zayed University, Abu Dhabi, United Arab Emirates
[d] Great Ormond Street Children's Hospital, London, UK
[e] College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences, Dubai, United Arab Emirates
[f] New York University Abu Dhabi, United Arab Emirates
[g] Weill Cornell Medicine, Doha, Qatar
[h] Sidra Medicine, Doha, Qatar
[i] The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA
[j] Mediclinic City Hospital, Dubai, United Arab Emirates
[k] Medanta Gururam, Delhi, India
[l] Royal Hospital for Children, Glasgow, UK
[m] Yas Clinic, Abu Dhabi, United Arab Emirates
[n] All India Institute of Medical Sciences, Bibinagar, Hyderabad, India
[o] Medical Research Division, Department of Internal Medicine, The National Research Centre, Cairo, Egypt

ARTICLE INFO

Keywords:
Meningococcal septic shock
Machine learning
Artificial neural network
Gene modular approach

ABSTRACT

Sepsis is a major global health concern causing high morbidity and mortality rates. Our study utilized a Meningococcal Septic Shock (MSS) temporal dataset to investigate the correlation between gene expression (GE) changes and clinical features. The research used Weighted Gene Co-expression Network Analysis (WGCNA) to establish links between gene expression and clinical parameters in infants admitted to the Pediatric Critical Care Unit with MSS. Additionally, various machine learning (ML) algorithms, including Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and Artificial Neural Network (ANN) were implemented to predict sepsis survival. The findings revealed a transition in gene function pathways from nuclear to cytoplasmic to extracellular, corresponding with Pediatric Logistic Organ Dysfunction score (PELOD) readings at 0, 24, and 48 h. ANN was the most accurate of the six ML models applied for survival prediction. This study successfully correlated PELOD with transcriptomic data, mapping enriched GE modules in acute sepsis. By integrating network analysis methods to identify key gene modules and using machine learning for sepsis prognosis, this study offers valuable insights for precision-based treatment strategies in future research. The observed temporal-spatial pattern of cellular recovery in sepsis could prove useful in guiding clinical management and therapeutic interventions.

* Corresponding author. Dr Asrar Rashid, Edinburgh Napier University, Department of Computer Science, Merchiston Campus, 10 Colinton Road, Edinburgh, Scotland, EH10 5DT, UK.

E-mail addresses: asrar.rashid@napier.ac.uk, asrar@medicalbrainbox.com (A. Rashid), A.Anwary@napier.ac.uk (A.R. Anwary), feras.al-obeidat@zu.ac.ae (F. Al-Obeidat), joe.brierley@me.com (J. Brierley), Mohammed.Uddin@mbru.ac.ae (M. Uddin), hoda.alkhzaimi@gmail.com (H. Alkhzaimi), asarpal@sidra.org (A. Sarpal), mohammedtoufiq91@gmail.com (M. Toufiq), Zainab.Malik@mediclinic.ae (Z.A. Malik), Raziya.kadwa@nmc.ae (R. Kadwa), drpraveen.k@rainbowhospitals.in (P. Khilnani), Guftar.shaikh@ggc.scot.nhs.uk (M.G. Shaikh), govind.pgi@gmail.com (G. Benakatti), javed.sharief@nmc.ae (J. Sharief), drzakisyed@gmail.com (S.A. Zaki), Abdulrahman.zey@nmc.ae (A. Zeyada), A.Al-Dubai@napier.ac.uk (A. Al-Dubai), Waeelhafez@yahoo.com, wael.hafez@nmc.ae (W. Hafez), A.Hussain@napier.ac.uk (A. Hussain).

https://doi.org/10.1016/j.imu.2023.101293
Received 9 April 2023; Received in revised form 3 June 2023; Accepted 7 June 2023
Available online 16 June 2023

1. Introduction

Sepsis is a significant global health challenge affecting individuals across socioeconomic backgrounds and countries, including low, middle, and high-income nations [1]. Based on data extrapolated from the United States of America, it is estimated that sepsis accounts for 15–19 million cases worldwide each year. Given the polymicrobial and heterogeneous nature of sepsis, studying specific clinical pathogenic states in particular age groups can provide valuable experimental benefits. One such condition is Meningococcal Septic Shock (MSS). MSS, when occurring without meningitis, is associated with a higher case fatality rate (CFR), ranging from 16% to 52% [2]. The rapid progression of the disease and the significant CFR of invasive meningococcal disease leading to MSS remains a concern, particularly among infants. Additionally, the burden of meningococcal disease is highest among young infants, with serogroup B being the most prevalent [3]. Infants may exhibit a genetic predisposition to MSS, with toll-like receptor-4 mutations being associated with invasive meningococcal disease in infants under 12 months of age [4]. In MSS, the primary focus of treating physicians is to provide critical care support that can impact the progression of the disease, particularly within the first 48 h. The exploration of clinical relationships through temporal microarray analysis can provide valuable insights into disease mechanisms relevant not only to MSS pathogenesis but also to sepsis as a whole.

Gene expression analysis has proven valuable in sepsis research, providing insights that can contribute to patient outcome prediction. Longitudinal studies have investigated the transcriptome in both children and adults, highlighting the significance of gene-expression data in sepsis prognostication [5–7]. In a study by Wong et al., microarray analysis was conducted on a pediatric sepsis cohort, employing Endotyping to classify patients into three subclasses (Endotype A, B, and C), based on underlying pathobiological mechanisms [8]. The researchers identified 100 genes that effectively differentiated between Endotypes A and B in children with septic shock [9]. They further concluded that allocation to subclass A was associated with a poorer outcome. Notably, the study also observed the concept of endotype-switching, where patients transitioned from one subgroup to another during the course of sepsis. These findings underscore the potential value of utilizing gene expression studies to develop precision medicine strategies for future sepsis management.

Time-series datasets can provide an important perspective with respect to sepsis evolution. By treating time-series gene expression data as interconnected geometric clustering networks, one can exploit the inherent interdependency of intra-patient data. Amongst various network analysis approaches, Weighted Gene Co-expression Network Analysis (WGCNA) stands out, as it clusters based on biological significance, not geometric distance, grouping genes into functional modules [10]. WGCNA also enables module stratification based on clinical parameters, aiding in gene-trait relationship studies [11]. Applications of this modular approach have demonstrated relationships between hub genes and long non-coding RNAs (lncRNAs) in sepsis models [12,13], identified key genes associated with sepsis prognosis [14], and developed gene panels for sepsis diagnosis [15]. By employing a secondary analysis of pediatric sepsis datasets, key hub genes were identified and validated through qPCR, indicating potential biomarkers for pediatric sepsis [5,16]. In a novel approach, WGCNA was followed by differential correlation analysis to uncover genes with opposing correlations in different conditions [17]. This exemplifies the evolving application of WGCNA in sepsis research.

An understanding of how sepsis evolves from a cellular perspective remains deficient. This is reflected by the lack of temporal gene expression studies in the clinical literature, especially in children. Therefore, we propose a topological modular approach using Weighted Gene Co-expression Network Analysis (WGCNA) to analyze a dataset of pediatric patients with meningococcal septic shock (MSS). We believe the study dataset employed for secondary analysis was the first multi-sampling published gene expression series in infants with septic shock [18]. As researchers with access to the original clinical data, we have a unique opportunity to correlate clinical phenotype and gene expression to infants with MSS. This dataset provides a chance to undertake a temporal analysis of dynamic clinical changes in association with changes in enriched gene function. Insights based on time-associated studies could galvanize the field of sepsis research through improved clinical application. Another challenge is that Sepsis is not a simple discrete event, but rather a complex non-linear multi-variable phenomenon, known for its heterogeneity and complexity in the transition from infection to clinical sepsis. This complexity, encompassing clinical, immunological, and pathophysiological dimensions, contributes to experimental variation making statistical analysis challenging. In light of these limitations, Machine Learning (ML) has been used to model crucial sepsis end-points, facilitating unsupervised classification and supervised labeling of datasets in sepsis [19–22]. Therefore, in this research study, as well as WGCNA, ML algorithms are employed to enhance prognostication.

2. Methods

The methodology employed in this study comprises two main components: network analysis using Weighted Gene Co-expression Network Analysis (WGCNA), as depicted in Fig. 1, and Machine Learning (ML) techniques.

2.1. Patient recruitment

Study details were previously published [18] with approval from the Nottingham University ethics committee (REC reference 05/Q2403/53). Patients presenting to Nottingham University Hospital Pediatric Critical Care (PCC) were recruited after obtaining written informed consent [Table 1]. Patients received standard clinical treatment, including appropriate antimicrobial therapy for presumed meningococcal sepsis. The children studied had no pre-existing medical conditions. Blood samples were collected on admission to PCC (designated 0 h) and at 4, 8, 12, 24, and 48 h following PCC admission.

2.2. RNA extraction

The dataset from this secondary analysis was available from the ArrayExpress dataset (E-MEXP-3850).

2.3. Microarray data analysis and weighted gene Co-expression network analysis (WGCNA)

The expression data set contains 30 samples from five patients at six different time points. Patient 4 at the 24-h time point had no expression values and was removed from further analysis, reducing the total samples to 29.33,297 probe sets from 29 Human Gene 1.0 ST Arrays were generated and compared. Using R software, the 29 Microarray gene expression sample dataset underwent WGCNA. First, a gene co-expression network was constructed after calculating the Pearson correlations between pairs of genes across all samples. Next, modules were identified using a hierarchical clustering dendrogram and dynamic tree-cut methodology. Densely interconnected gene clusters were represented by modules, according to a soft thresholding power β. A soft-thresholding power of 6 was chosen. It is the lowest power for which the scale-free topology fit index curve flattens (0.68). A clustering dendrogram was generated, assigning colors to the modules. This led to the identification of 19 modules labeled 0–18, with the number of genes associated with each gene cluster. The label 0 was reserved for genes outside of all modules.

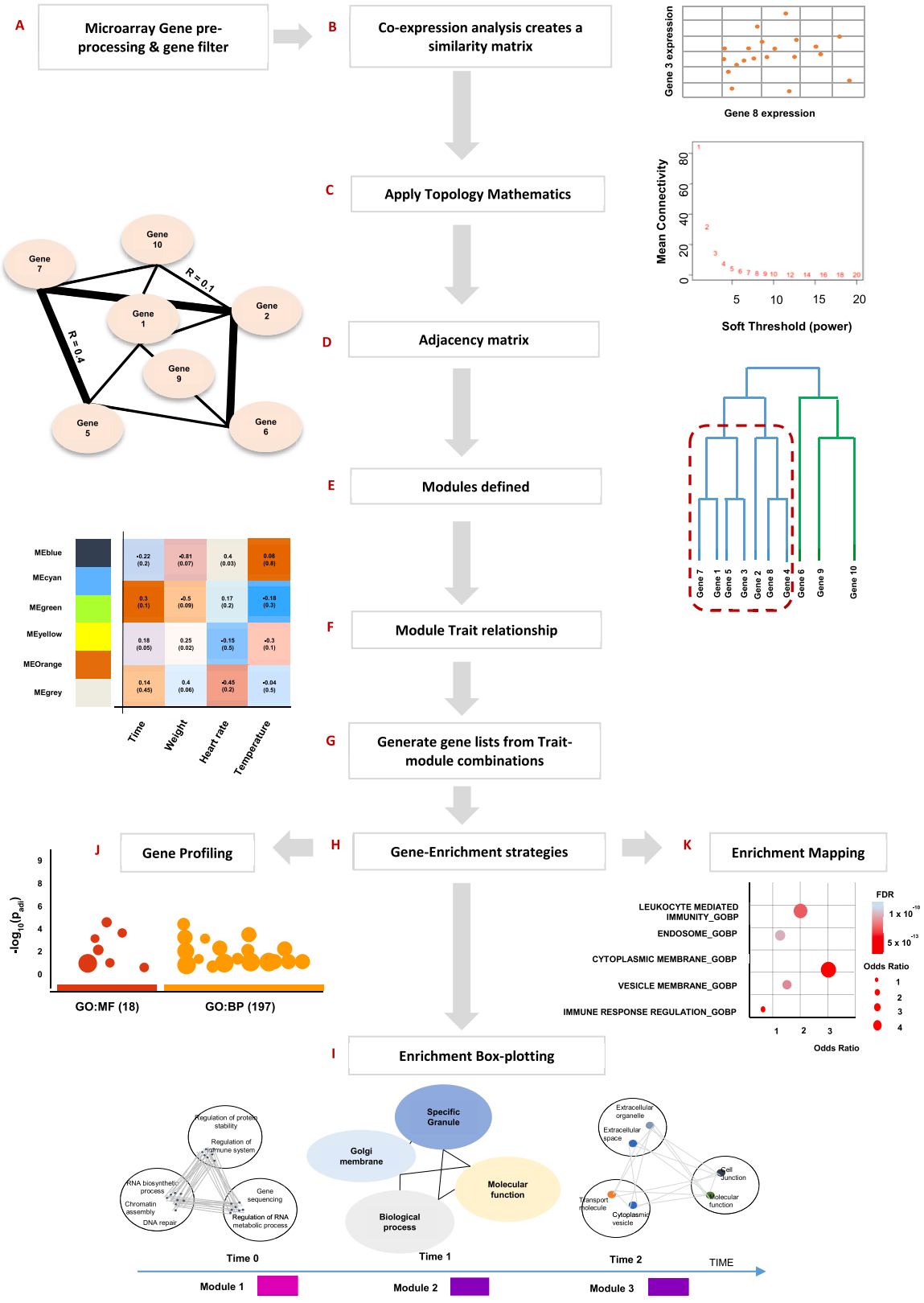

Fig. 1. **A.** Preprocessed data log2 normalized downloaded **B.** Pairwise correlation of genes undertaken for each gene-pair combination. **C.** Choosing a topological soft threshold value for the power of Beta allows the construction of a module-centric network. **D.** An adjacency network is constructed. The nodes in the network correspond to genes, and the connections are known as edges determined by the pairwise calculations in A. The edges are calculated between 0 and 1. **E.** Using a hierarchal clustering, similar genes are grouped in a tree structure with 'branches' denoted as gene modules. A module consists of a collection of highly inter-connected genes with high absolute correlation. **F.** A module-trait matrix is then generated associating traits (horizontal axis) to Module Eigenes (Vertical Axis).

2.4. Module detection

Clustering was also performed based on the module color and clinical traits of time, age, gender, mortality, and weight. Subsequently, modules were related to phenotypic data based on clinical variables. Each given module generated a first principal component, the Module Eigengene (ME). Clinical trait data were then correlated against the ME, giving a correlation coefficient. Genes from the significant modules showing high Module Membership (MM) were filtered and selected (p. MM $\leq$ 0.05).

2.5. WGCNA construction and detection of disease-associated modules

A quantitative measure of MM was defined for each module as the correlation of the ME with the gene expression profile. Modules were related to phenotypic characteristics, such as weight, age, mortality, and organ dysfunction (based on the Pediatric Logistic Organ Dysfunction score [PELOD]). An adjacency matrix was assembled, with rows corresponding to MEs and columns to clinical traits (Fig. 2). Genes from the significant modules showing high module membership were filtered and selected (the probability of module membership was $\leq$0.05). Eigengenes were formulated for each module (Module Eigenes) and correlated to phenotypic characteristics (external trait) data. Each association was color-coded by the correlation value.

2.6. Gene enrichment

WGCNA analysis generated gene lists showing significant module membership. These gene lists then underwent pathway enrichment studies. The Fisher exact test was then applied to the gene list. Using in-house R script, pathways were generated using Kyoto Encyclopedia of Genes and Genomes (KEGG) database annotation with the associated Gene Ontology (GO) terms. The subsequent enriched gene list was then imported into Cytoscape [23] and annotated using the enrichment map tool within the Cytoscape platform (Fig. 3). The Kyoto Encyclopedia of Genes and Genomes (KEGG) database provided an interpretation of the enriched gene pathways. This enriched data was then passed into the enrichment map software in Cytoscape using a p-value (0.001) and an FDR (0.01) threshold to illustrate the enriched pathways. Further, the enriched gene list, using R script, filtered using a p-value (0.001) and an FDR (0.01) threshold, generated enrichment dot plots (Fig. 2). Dot plots for PELOD 0, 24, and 48-time categories were generated using the top 25 significant ($p < 0.01$) pathways for ease of illustration (Fig. 4). Dot plots for PELOD 0, 24, and 48-time categories were generated using the top 25 significant (p-value <0.01) pathways for ease of illustration (Fig. 4). The pathways gendered according to the dot plots pertained to GO terms and terms from the Reactome database. The WGCNA-generated gene lists were also enriched by parsing through a gene profiling platform, g: Profiler. The significantly upregulated genes ($p < 0.05$) according to the adjacency matrix trait underwent functional enrichment analysis using g: Profiler. A $p < 0.05$ for statistical significance and the Benjamini-Hochberg FDR (False Discovery Rate) were used to reduce the chance of false positives. As detailed (Fig. 5), g: Profiler uses a number of client libraries to interpret gene lists from a functional enrichment point of view.

2.7. Machine learning data processing and methods

The dataset contained 29 instances of survival class (23) and non-survival class (6), which was an unequal distribution of classes. In machine learning, unequal data distribution is one of the major causes of decreasing accuracy of classification models. Due to the imbalance instances in the dataset, machine learning models could not effectively learn the patterns for survival and non-survival classes. As the non-survival class was less in number, the results generated by this class would become ineffective. To overcome this challenge, a synthetic minority oversampling technique (SMOTE) was applied to handle the imbalanced data [24]. This popular approach is often used in classification problems of imbalanced datasets. SMOTE is considered one of the most powerful, reliable, and adaptable pre-processing techniques in machine learning [25]. After balancing the dataset, it is important to identify patterns in the data series and express them so that the similarities and differences can be observed and reduce the dimensionality without losing too much information. Principal component analysis (PCA) is a multivariate technique to reduce the complexity of the input variables. This analyses extremely interrelated components in the dataset and decreases the complexity and dimension. Thus extracting the most significant information in the dataset. Therefore, PCA was applied to strip out the low-influence features from the dataset. After the preprocessing of data, six popular machine learning techniques, Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, and Artificial Neural Network (ANN), were applied to understand the impact of each technique on the classification of the given survival and non-survival datasets. SVM is a supervised machine learning algorithm that identifies different classes by separating the classes with the help of a decision boundary known as a hyperplane (a line that distinguishes two classes). DT is a classifier that uses a tree-like structure based on knowledge gained on classification. KNN is a classifier technique where the training is predicated on “how similar” one dataset is from another based on the distances between a

Table 1
Five children were recruited into the Meningococcal septic shock study. Patient one was non-surviving. Also, patient 1 was culture negative with the diagnosis of MSS on clinical grounds. All children developed DIC and required mechanical ventilation. GpB = Group B Neiserria Meningococcus.

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5
Number of samples	5	5	5	5	5
Age (months)	13	10	22	24	9
Sex	Female	Female	Female	Male	Male
Weight (Kg)	12	12.9	12	15	8
Duration of PICU admission (DAYS)	9	4	3	6	3
No. of organ(s) in failure	4	4	3	6	3
PELOD score on admission	61	31	31	12	11
PELOD Score at 24 h	52	2	22	22	2
PELOD Score at 48 h	43	2	31	12	1
Median PRISM Score at 12 h	12	11	17	14	9
Median PRISM Score at 24 h	15	7	15	13	4
Serotype	Negative culture[a]	GpB meningococcus	GpB meningococcus	GpB meningococcus	GpB meningococcus
GCS at 24 h	3	7	3	3	10
Mean Inotrope score on Day 1	38	13	112	27	9
Mortality (at 28 days)	Died	Alive	Alive	Alive	Alive
DIC	Yes	Yes	Yes	Yes	Yes
Duration of mechanical ventilation (days)	5	4	4	4	4′

[a] Presumed meningococcal sepsis based on clinical grounds.

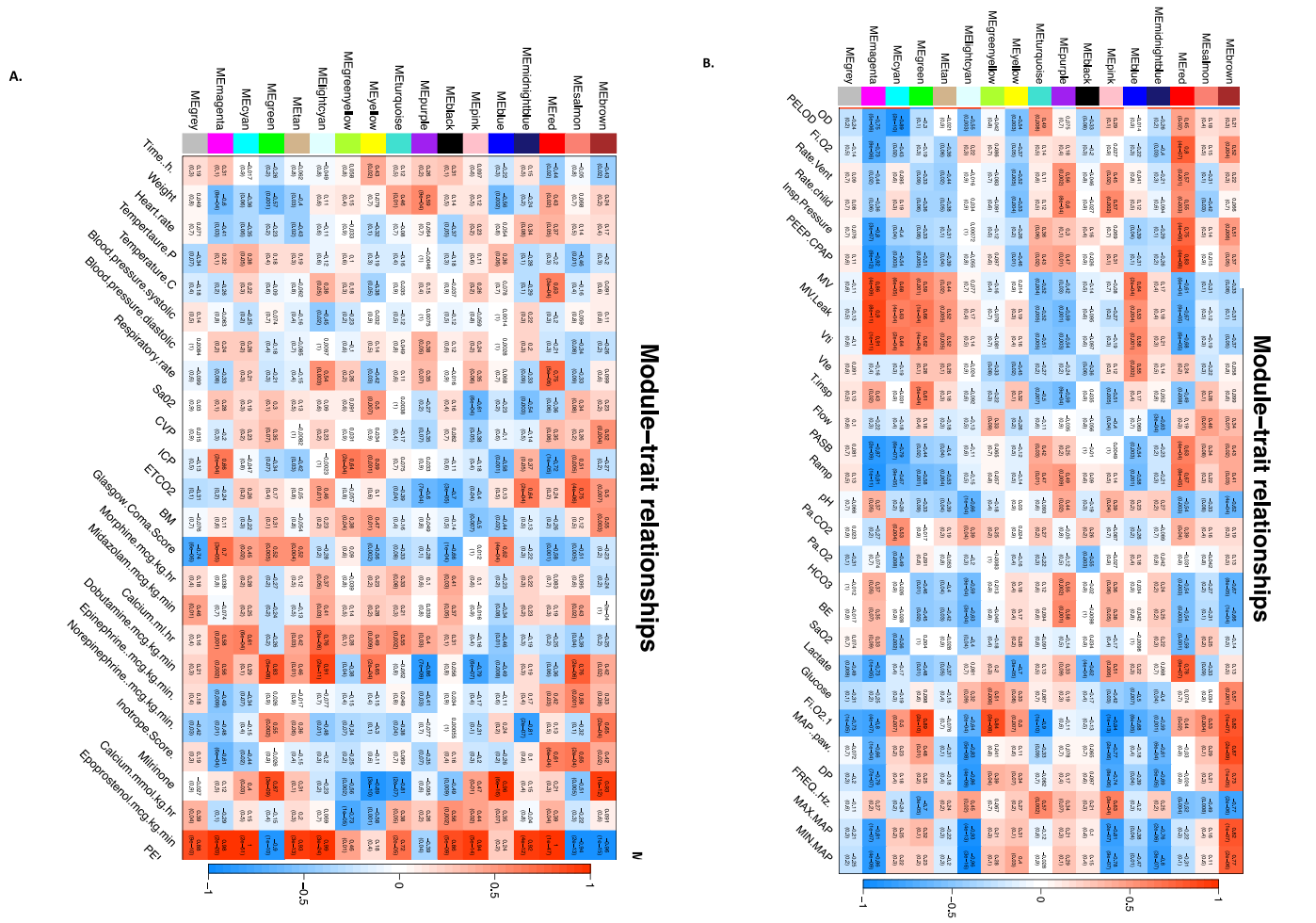

Fig. 2. Module-trait associations heat map representation of adjacencies in the eigengene network (**2A** and **2B**). The table is color-coded by correlation according to the color legend. White color represents low adjacency (low correlation), red high adjacency (positive correlation) and green represents high adjacency (negative correlation). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

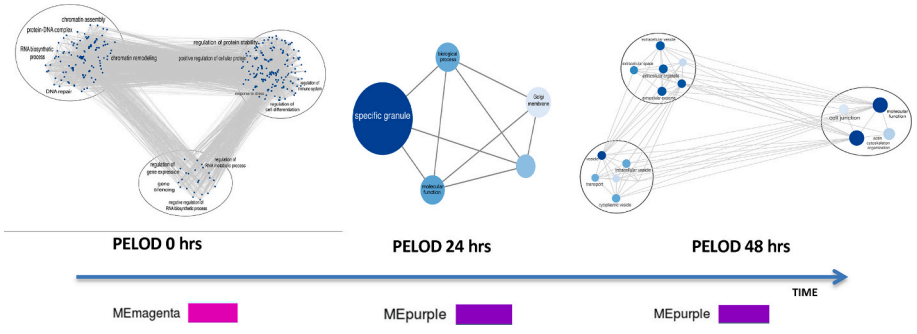

Fig. 3. Gene-set enrichment results are graphically mapped to the Enrichment Map. The enrichment score (the enrichment p-value) is mapped to the node color as a color gradient, with node size proportional to the odds ratio. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

point and all the examples within the data, selecting the required number of examples (K) closest to the point, incorporating votes for the frequent leading label. The random forest creates many trees that achieve their output through ensemble learning methods for classification. Naïve Bayes is a classification technique that uses a simple probability that applies Bayes Theorem with high independent assumptions. Bayes theorem is used in statistics to calculate the probability of a class of each attribute group present to determine which class is optimal. ANN is another classification technique that mimics the functioning of a human brain with the basic principle that a number of parameters as inputs are processed in such a way as in the hidden layer (multiplication, addition, division, etc.), then processed again in the output layer to produce an output. For these machine learning techniques, the pre-processed data were partitioned into training and testing with a ratio of 70%:30%. The training dataset is fitted to the machine learning classifier, and later predictions were obtained using the testing dataset. These six

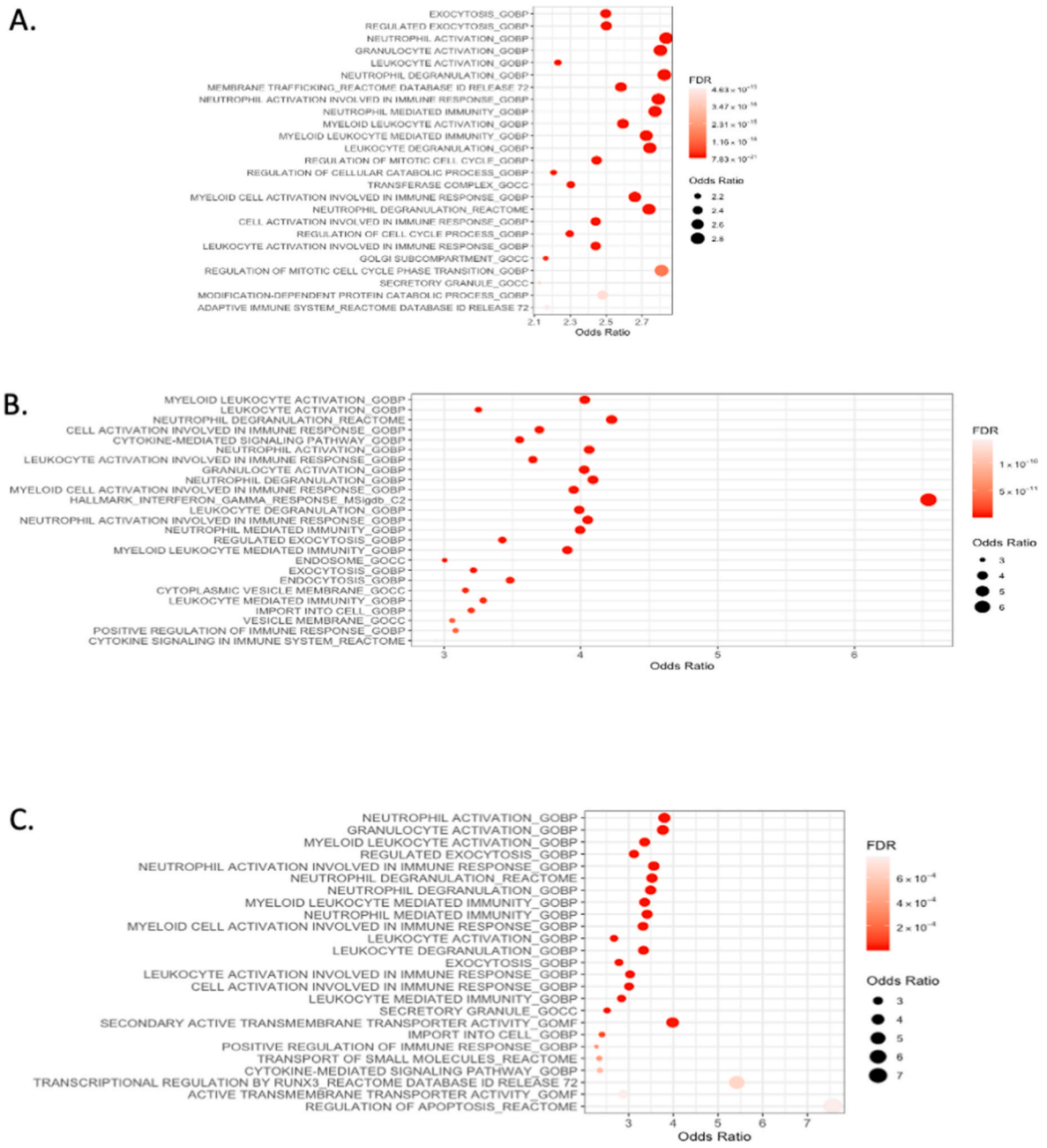

Fig. 4. Box plot enrichment box plot of significant genes from the WGCNA clusters for PELOD 0 h **A.** PELOD 24hrs. HALLMARK_INTERFERON_GAMMA_RESPONSE_MSigdb_C2 is seen to be an outlying pathway. **B.** PELOD 48 h. **C.** Enrichment results were filtered using a p-value (0.001) and an FDR (0.01) threshold. The plots display the top 25 pathways.

machine-learning techniques were applied, and the results were obtained.

3. Results

3.1. Patient demographics

All Infants demonstrated clinical phenotype consistent with severe shock and diffuse intravascular coagulation consistent with Meningococcal sepsis (Table 1). Patient 1 died and is noted as having received a Protein C infusion during treatment.

3.2. Module trait associations

An adjacency network matrix was formulated from the WGCNA of the gene expression data. The correlation between eigengenes and traits is depicted as a heat map (Fig. 2). Each row corresponds to a module eigengene, the column to a trait. Each cell contains the corresponding correlation and p-value (in parenthesis). Each row and column in the heatmap corresponds to one module eigengene (labeled by color) or weight. The highest correlation for PELOD at 0 h was with the MEmagenta module (0.83) with a highly significant p-value of 2e-08. At PELOD 24 h, MEpurple modules were the most significant, with a correlation of 0.74 and a p-value of (4e-05). With PELOD 48 h, MEpurple modules were the most significant, with a correlation of 0.95 and a p-value of (7e-17).

3.3. Pathway enrichment studies, enrichment map generation

At PELOD time 0, pathways related to cell nuclear function were seen

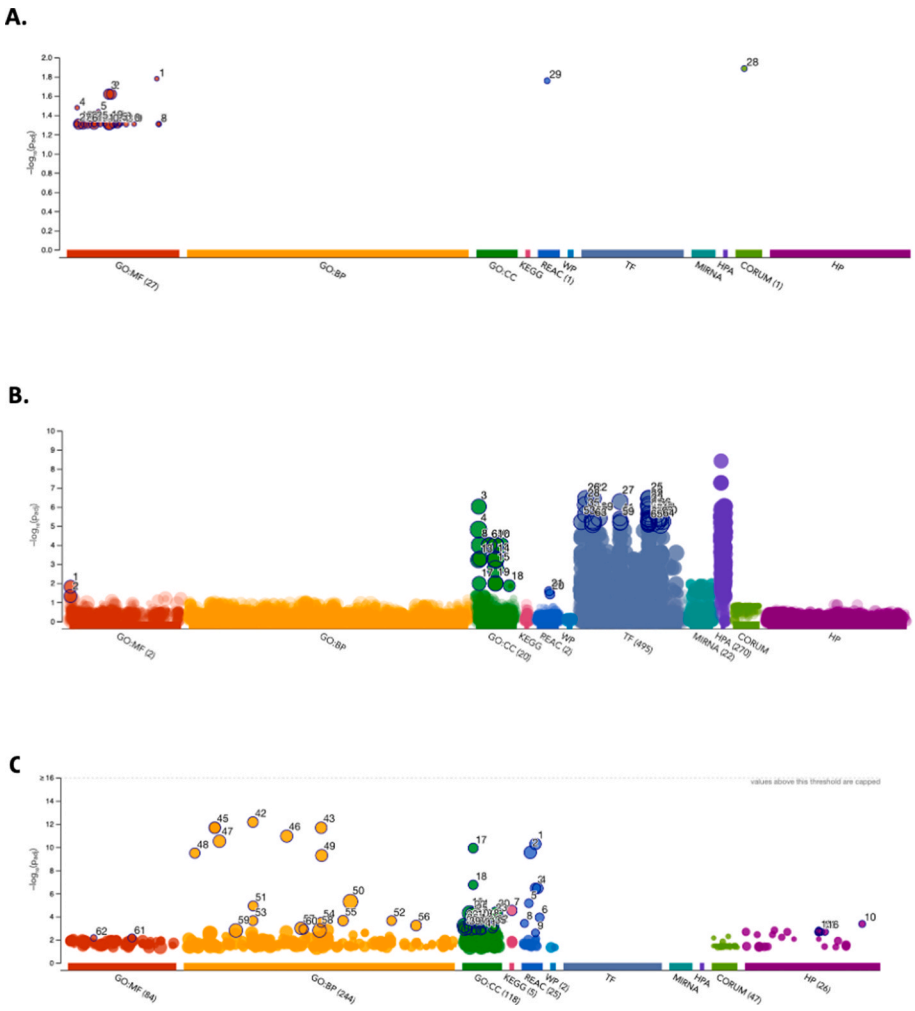

Fig. 5. WGCNA generated gene lists for **A.** PELOD 0hrs, **B.** PELOD 24hrs, and **C.** PELOD 48 h were then parsed through the g: Profile web application to show the enriched genes. Key is GO - Gene Ontology, GO: MF - Molecular Functions, GO: BP - Biological Process, GO: CC - Cellular Component, REAC: Reactome, KEGG - Kyoto Encyclopedia of Genes and Genomes, TF - Transpac, CORUM - CORUM protein complexes, HP - Human Phenotype Ontology, MIRNA – miRTarBase.

to be up-regulated (p-value 0.01 FDR 0.01); at PELOD 24 h, cytoplasmic gene function was upregulated (p-value 0.01 FDR 0.01), and finally at 48 h, extracellular gene function upregulated (p-value 0.01 FDR 0.01) (Fig. 3). Enrichment mapping through the Cytoscape application generated significant pathways at time 0, pathways related to cell nuclear function; at 24 h, cytoplasmic gene function and at 48 h, extracellular gene function. The Enrichment map node size represents the number of genes in the gene set; edge thickness is proportional to the overlap between gene sets.

3.4. Functional enrichment analysis

A graphical representation functional enrichment analysis using g: Profile software was undertaken. Data was parsed through the g: Profile platform from the WGCNA-generated gene modules for selected clinical traits. Generated Manhattan plots according to PELOD 0 h, 24 h, and 48 h are shown (Fig. 5). The x-axis represents functional terms grouped and color-coded by data sources (e.g., Molecular Function from GO is red; the sources not included in the analysis are shown in grey). The y-axis shows the adjusted enrichment p-values in the negative log10 scale. The light circles represent insignificant terms (if available). P values in the outputs are color-coded from yellow (insignificant) to blue (highly significant or smallest possible p-value). From the Manhattan plots, PELOD 48 and 24 h appear to have more enriched genes than PELOD 0 h. At PELOD 24 h, GO cellular component pathways include a nuclear lumen, intercellular cytosol, organelle, nuclear body, and chromatin activity is noted. At PELOD 48 h, the pattern of GO cellular component pathways is similar to that at 24 h.

3.5. Enrichment plots

According to the odds ratio (OR), the innate response, depicted by neutrophil-activation pathways, showed more significant expression at PELOD 24 h (p-value = 2.96e-15 OR = 4.06 FDR = 1.07e-12) and 48 h (p-value = 4.13e-12 OR = 3.80 FDR 5.35e-09) compared to PELOD 0 h (p-value = 1.27e-21 OR = 2.83 FDR = 1.80e-18)(Fig. 4). In addition, the OR at PELOD 24 h showed. For the 48-h PELOD, OR beyond 5.0, included the TRANSCRIPTIONAL REGULATION by RUNX3_REACTOME and Regulation of APOPTOSIS_REACTOME. Pathways present at PELOD 0 but not at the other time points include GO pathways related to the mitotic cycle and the Golgi sub-compartment. Pathways present at 24 h and not at PELOD 0 or 48 h included GO pathways related to the cytoplasmic vesicle membrane, endoscope and import function into the cell. Regarding cytokine signaling, no pathways were seen at PELOD 0 h, but GO pathways were present at 24 and 48 h.

3.6. Machine learning

The applied machine learning techniques are summarised (Table 2). Among the six algorithms used, ANN provided the most accurate

prediction. After preprocessing the dataset using SMOTE, PCA was employed, which included 99% variance in the dataset used for applying machine learning techniques (Fig. 6). For the ANN, a three-layer neural network was constructed. The first layer was the input layer which contained 16 neurons having "relu" as an activation function. The input layer accepts the input from the gene and forwards it to the second layer. The second layer is the inner hidden layer, which is used to construct the model, containing eight neurons with "relu" as an activation function. The parameter is mapped in hidden layers to one of the most appropriate feature classifications and ends with a predicted output. The third layer is the output layer which contains "sigmoid" as an output function. This layer differentiates total values obtained from inner hidden layers into two classes, 1 for yes and 0 for no. During this process, the loss function binary cross entropy was used, which provided the network with gene information to improve its knowledge of the input data. The Adaptive Moment Estimation (Adam) optimizer and activation function ReLU was used to improve this by changing the weights of each neuron and then trying again to improve prediction. Changing weights alters the extent to which the input neurons affect the final result; this implies that some parts of the input data may impact output variables more than others. Hyperparameter tuning was employed to reduce errors between the training and testing sets for optimal learning. The training and test scores were 100% and 100%, respectively, the same for the imbalanced and balanced datasets.

4. Discussion

Sepsis is a rapidly developing condition associated with systemic instability. Utilizing MSS as a septic shock model, a unique temporal microarray dataset from infants with MSS in PICU was scrutinized through WGCNA analysis [18]. This dataset, with six mRNA sampling points, enabled tracking of sepsis progression using WGCNA analysis. Thus allowing PELOD scores at 0, 24, and 48 h to be correlated, represented by a module-trait matrix, to gene expression data (Fig. 2). We propose that this is the inaugural integration of PELOD, a pediatric clinical scoring system, with transcriptomic data to delineate enriched gene expression modules in acute sepsis. The WGCNA analysis revealed a dynamic transition in gene function pathway enrichment, from nuclear to cytoplasmic, and finally to extracellular, associated with the PELOD times. Gene expression activity consistent with nuclear activity in sepsis was also noted by Wong et al. (2010) in pediatric polymicrobial sepsis [26]. Moreover, Walsh et al.(2016) corroborated the utility of WGCNA in unveiling gene-modular relationships in adult ICU patients over a longer period (7 days–6 months); their study demonstrated gene modular enrichment for skeletal muscle regeneration and deposition of the extracellular matrix [27]. Further, our study highlights the usefulness of time-series gene expression data, showing an augmented innate response associated with higher PELOD scores at 24 h and 48 h compared to PELOD scores at 0 h (Fig. 2). In sepsis, the dysregulated and disrupted physiological process requires the restoration of normal regulatory mechanisms. Based on our results, we propose the temporal pattern of gene function enrichment relates to the spatial recovery of essential cellular functioning. Firstly there is the correction of nuclear mechanisms to facilitate clinical recovery; genes are intricately involved in cellular regulation and enrichment of associated gene function pathways could be an important early indicator of the normalization of cellular function. The temporal analysis then leads to the next spatial layer outside of the nucleus, the cytoplasm. Rectifying dysfunction occurring in the immediate cytoplasmic area further restores normal cellular processes. The final step, as suggested by the temporal pattern in genomic function enrichment, is the restoration of the extracellular framework. These steps likely hail the normalization of severe organ dysfunction seen in patients with severe sepsis or septic shock.

Langfelder et al. (2013) compared WGCNA over standard statistical methods for differential gene expression [28]. Here Langfelder investigated the use of WCGNA for hub-gene selection, finding WGCNA as an enhancement over standard statistical approaches incorporating the p-value. However, counter to this, Langfelder also found, regarding analytical repeatability using independent data sets, that standard statistical methods were an enhancement over WGCNA. However, WGCNA methodology is advantaged by minimizing type 1 and type 2 statistical errors. Moreover, WGCNA applied to sepsis may show potential beyond traditional clinical biomarkers. For example, LONG et al. (2020) combined WGCNA with a machine learning algorithm and applied their workflow to three publicly available sepsis datasets [29]. Hereby applying artificial intelligence methods to WGCNA a diagnostic classifier was presented with the potential for early sepsis diagnosis.

We believe this study (in MSS) to be methodically advantageous over sepsis studies where the chosen pathogen is dissimilar. Wong et al. (2007) advocated a single-organism approach [30]; assuming similar changing patterns in gene expression minimized experimental variation and simplified gene analysis. Another factor affecting the host's genomic response to sepsis is age. For example, Wynn et al. (2011) studied neonates, infants, toddlers, and school-age children within 24 h of PCC admission in septic shock [31]; demonstrating that developmental age impacts the early whole-blood transcriptomic response in sepsis. Furthermore, Raymond et al. (2017) explored age effects on the transcriptome, showing infants and children being mostly similar, whereas neonates and adults were more different in their responses [32]. In our study, recruitment was restricted to infants with no previous co-morbidities, thereby minimizing extraneous effects.

Table 2 reveals that SVM, Random Forest, and DT yield high training scores but disappointing test scores, while Naïve Bayes generates comparable scores for both training and testing. KNN achieved 90% training and 89% testing accuracy on the imbalanced dataset, improved to 94% and 93% respectively after dataset balancing with SMOTE. ANN, however, achieved 100% accuracy for both training and testing, regardless of dataset balance. The principal components PC1 to PC3 exhibited the highest variances (Fig. 3). Notably, a hyperplane failed to separate the training and test data, making SVM, similar to Random Forest, DT, and Naïve Bayes, unsuitable for this classification. KNN, using Euclidean distance to separate classes, could potentially use an oval circle to separate the red dots represented by PC1 to PC3, but its train and test scores fell short compared to ANN. Thus, considering all factors, ANN, which achieved the highest accuracy (100%) for both training and testing datasets, was chosen. ANN excels in capturing non-linear relationships between input features and the target variable, which is crucial in the context of sepsis involving complex datasets with intricate interactions [33,34]. ANN also performs representation learning, automatically extracting meaningful features from the data. Additionally, ANN is adaptable and able to handle large amounts of data, having the capacity to capture the dataset complexity through its architecture. Hyperparameter tuning further enhances ANN's performance, making it superior to other classifiers in the study [35]. Further, we chose to use the SMOTE to address the issue of imbalanced data. SMOTE is a popular and effective oversampling method that can generate synthetic samples for the minority class by interpolating between existing instances. Thereby alleviating overfitting risk while increasing the representation of the minority class and improving the learning capability of the

Table 2
Results of different ML Models applied to the dataset.

	Imbalanced Dataset, PCA 99% variance		Balanced Dataset, PCA 99% variance	
	Train Score	Test Score	Train Score	Test Score
SVM	1	0.78	1	0.5
Random Forest	1	0.78	1	0.93
Logistic Regression	1	1	1	1
Decision Tree	1	0.89	1	0.93
Naive Bayes	0.97	0.78	0.97	0.79
KNN	0.9	0.89	0.94	0.93
ANN	1	1	1	1

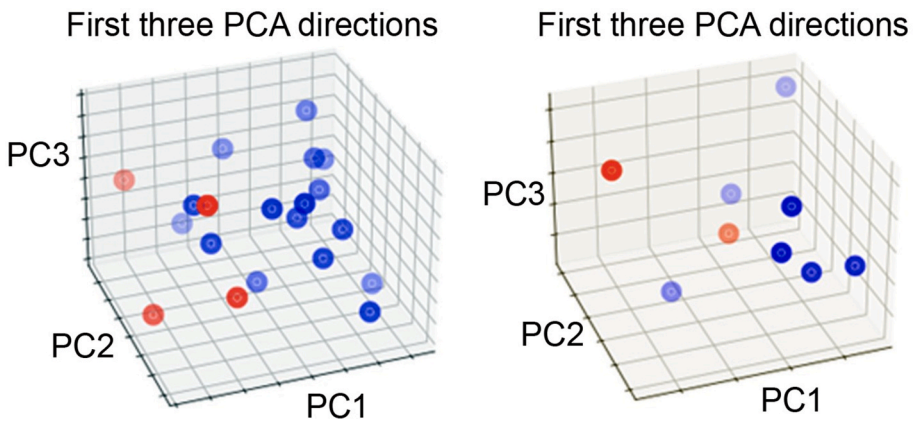

Fig. 6. PCA is applied to the dataset to strip out the low-influence features. The training dataset's first 3 principal component representation is shown (Fig. 6A), and the test dataset (Fig. 6B). The colors red and blue represent different classes. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

machine learning algorithms whilst preserving data integrity.

A challenge for sepsis studies is in establishing a correlation between clinical manifestations and cellular-level processes. This could be a reason contributing to why a significant therapeutic advancement in the field has yet to materialize. In trying to relate clinical variables to those of gene expression in the analysis, one limitation was the small sample size. However, for the application of WGCNA, a minimum of 15–20 samples are recommended, a criterion which was met in our study (30 samples) [36]. Despite this, the study advances knowledge related to sepsis transcriptomics by linking clinical parameters to gene function through the modular approach described. The small sample size also presents a challenge for ML. This was circumvented by dividing the data into separate training and testing groups, according to a ratio of 70:30, randomly selecting samples into each group. We then conducted a 5-fold cross-validation to assess model performance. This mitigates the impact of the limited sample size and ensures a robust evaluation of our ML approach. Future scope of research could include the application of explainable AI (XAI) helping to narrow down the focus on specific genes and molecular pathways, thereby enhancing the interpretability of temporal gene expression data. In addition, it is worth considering expanding the future scope of this research to encompass a comparison of alternative data reduction techniques beyond that of PCA employed in this study. Finally, the study attempted to include therapy information by presenting a trait-gene module adjacency matrix (Fig. 2). However, due to the methodology employed in this study, there were limitations in comparing different therapies and management strategies. This aspect should be the primary focus of future research. Additionally, incorporating diagnostic staging could further enhance the analysis and provide valuable insights.

The temporal aspect of this study capitalizes on the inherent value of time-related gene expression datasets. Specifically, each patient's samples constitute a temporal sequence, documenting the evolving septic process. This allows network methodologies such as WGCNA to be effectively utilized, even in studies with small sample sizes. This led to the discovery of a temporal-spatial gene expression pattern that could have future applications in assessing clinical management strategies and developing novel therapeutics. However, a temporal limitation of the study relates to the fact that although the study included six-time points, there may be other critical time points during the course of sepsis not captured. Also, the arbitrary allocation of time points along the sepsis time trajectory could affect the analysis of time-dependent changes in gene function. In this study time-labeling of patients occurred from Pediatric Critical Care Unit admission onwards, independent of disease trajectory as accurate clinical time-profiling is not possible. Further, the idea of temporal sampling and compartmentalization in sepsis is complicated by the heterogeneity of sepsis. Moreover, it is difficult to time-match gene expression series without a robust objective definition of sepsis. The unknown temporal difference between infection and symptom onset in patients, as well as sepsis heterogeneity encompassing factors such as symptom onset speed, pathogenesis rapidity, and the ability to seek medical assistance, pose numerous challenges. Nevertheless, it is important to highlight that the dataset utilized in this study represents a secondary analysis of the first published case series in infants with septic shock [18]. In this dataset, despite potential variances due to the various temporal factors, distinct time-associated patterns related to gene function were still discernible. Temporal patterns may be attributed to the therapeutic drive for physiological stability, reflecting a clinical impact on each transcriptome. Looking forwards, temporal studies of sepsis are suggested especially with regard to sepsis management at the bedside and the development of precision strategies.

5. Conclusions

This study demonstrated the value of time-related trajectory transcriptomic data and gene co-expression network analysis in understanding sepsis evolution in infants admitted to the pediatric intensive care unit. Uniquely, the application of WGCNA was shown to correlate temporal gene expression with bedside clinical data, resulting in the elucidation of a recovery pattern of temporal-spatial gene expression. The approach provided insights into the molecular trajectory of MSS, permitting visualization of treatment impacts in relation to genomic modular patterns in MSS. In parallel, we conducted a comparative study employing six machine-learning algorithms - SVM, Naive Bayes, KNN, DT, Random Forest, and ANN - for sepsis survival prediction. ANN emerged superior, offering 100% accuracy for both training and testing datasets. Future work aims to expand the training and testing datasets, augmenting the reliability of the resultant ML model. The integration of network methods for isolating biologically significant gene modules and machine learning for sepsis prognostication heralds a new era for precision therapeutic strategies. Future exploration into the temporal correlation of physiological and genomics data remains a promising avenue to enhance our comprehension of rapidly evolving sepsis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors thank the anonymous reviewers for their insightful

comments and suggestions. We are extremely grateful, to Professor Delawar Uddin, Professor Harish Vyas, Dr. David Thomas, the Charity For Lucie, and Dr. Mark Peters at the Institute of Child Health. We thank Dr. Ege Ulgen for assistance with the data pre-processing. Professor Hussain acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) - Grants Ref. EP/M026981/1, EP/T021063/1, EP/T024917/1. To the team at NMC Royal Hospital Abu Dhabi, Dr. Mouhamad Al Zoubhi, Dr. Husam Saleh, Dr. Maki Hamad, Dr. Ali Nawaz, Mr. Juju Thomas, and NMC Corporate, Mr. Frank Delisi, Dr. Alan Stewart, Ms. Kate Hoffman, and Mr. David Hadley for supporting International research at NMC Healthcare. Also, thanks to Dr. Ege Ulgen for aiding in pre-processing of data. Finally, and certainly, not least, Professor Hector Wong, whose decades-long contribution to the field of sepsis genomics remains an enduring legacy, may he rest in peace.

Abbreviations

(CFR) Case fatality rate
FDR (False Discovery Rate)
(C3AR1) Gene Complement C3a Receptor 1
(MMP9) gene Gene Matrix Metalloproteinase-9
(KEGG) Kyoto Encyclopedia of Genes and Genomes
(lncRNAs) Long non-coding RNAs
(MSS) Meningococcal Septic Shock
(ME) Module Eigengene
(MM) Module Membership
(PCC) Pediatric Critical Care
(PELOD) Pediatric Logistic Organ Dysfunction
(p.MM) p-value for Module Membership
(PqPCR) Quantitative Polymerase Chain Reaction
(WGCNA) Weighted Gene Co-expression Network Analysis

References

[1] Organization SWH. Improving the prevention, diagnosis and clinical management of sepsis. WHO; 2017. A70/13.
[2] Sadarangani M, Pollard AJ. Can we control all-cause meningococcal disease in Europe? Clin Microbiol Infect 2016;22(Suppl 5):S103–12.
[3] MacNeil JR, Bennett N, Farley MM, et al. Epidemiology of infant meningococcal disease in the United States, 2006-2012. Pediatrics 2015;135(2):e305–11.
[4] Faber J, Meyer CU, Gemmer C, et al. Human toll-like receptor 4 mutations are associated with susceptibility to invasive meningococcal disease in infancy. Pediatr Infect Dis J 2006;25(1):80–1.
[5] Wong HR, Cvijanovich N, Allen GL, et al. Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. Crit Care Med 2009;37(5):1558–66.
[6] Braga D, Barcella M, Herpain A, et al. A longitudinal study highlights shared aspects of the transcriptomic response to cardiogenic and septic shock. Crit Care 2019;23(1):414.
[7] Cazalis MA, Lepape A, Venet F, et al. Early and dynamic changes in gene expression in septic shock patients: a genome-wide approach. Intensive Care Med Exp 2014;2(1):20.
[8] Wong HR, Cvijanovich N, Lin R, et al. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. BMC Med 2009;7(1):34.
[9] Wong HR, Cvijanovich NZ, Anas N, et al. Developing a clinically feasible personalized medicine approach to pediatric septic shock. Am J Respir Crit Care Med 2015;191(3):309–15.
[10] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinf 2008;9:559.
[11] Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome 2007;18(6):463–72.
[12] Huang J, Sun R, Sun B. Identification and evaluation of hub mRNAs and long non-coding RNAs in neutrophils during sepsis. Inflamm Res 2020;69(3):321–30.
[13] Cheng L, Nan C, Kang L, et al. Whole blood transcriptomic investigation identifies long non-coding RNAs as regulators in sepsis. J Transl Med 2020;18(1):217.
[14] Xu C, Xu J, Lu L, Tian W, Ma J, Wu M. Identification of key genes and novel immune infiltration-associated biomarkers of sepsis. Innate Immun 2020;26(8):666–82.
[15] Zhou X, Wang Y, Chen J, Pan J. Constructing a 10-core genes panel for diagnosis of pediatric sepsis. J Clin Lab Anal 2020:e23680.
[16] Li Y, Li Y, Bai Z, Pan J, Wang J, Fang F. Identification of potential transcriptomic markers in developing pediatric sepsis: a weighted gene co-expression network analysis and a case-control validation study. J Transl Med 2017;15(1):254.
[17] Sheng L, Tong Y, Zhang Y, Feng Q. Identification of hub genes with differential correlations in sepsis. Front Genet 2022;13. 876514-876514.
[18] Kwan A, Hubank M, Rashid A, Klein N, Peters MJ. Transcriptional instability during evolving sepsis may limit biomarker based risk stratification. PLoS One 2013;8(3):e60501.
[19] Rodriguez A, Ruiz-Botella M, Martin-Loeches I, et al. Deploying unsupervised clustering analysis to derive clinical phenotypes and risk factors associated with mortality risk in 2022 critically ill patients with COVID-19 in Spain. Crit Care 2021;25(1):63.
[20] Boussen S, Cordier PY, Malet A, et al. Triage and monitoring of COVID-19 patients in intensive care using unsupervised machine learning. Comput Biol Med 2022;142:105192.
[21] Mueller YM, Schrama TJ, Ruijten R, et al. Stratification of hospitalized COVID-19 patients into clinical severity progression groups by immuno-phenotyping and machine learning. Nat Commun 2022;13(1):915.
[22] Wenric S, Shemirani R. Using supervised learning methods for gene selection in RNA-seq case-control studies. Front Genet 2018;9:297.
[23] Cline MS, Smoot M, Cerami E, et al. Integration of biological networks and gene expression data using Cytoscape. Nat Protoc 2007;2(10):2366–82.
[24] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57.
[25] Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 2018;61:863–905.
[26] Wong HR, Freishtat RJ, Monaco M, Odoms K, Shanley TP. Leukocyte subset-derived genomewide expression profiles in pediatric septic shock. Pediatr Crit Care Med 2010;11(3):349–55.
[27] Walsh CJ, Batt J, Herridge MS, et al. Transcriptomic analysis reveals abnormal muscle repair and remodeling in survivors of critical illness with sustained weakness. Sci Rep 2016;6:29334.
[28] Langfelder P, Mischel PS, Horvath S. When is hub gene selection better than standard meta-analysis? PLoS One 2013;8(4):e61505.
[29] Long G, Yang C. A six-gene support vector machine classifier contributes to the diagnosis of pediatric septic shock. Mol Med Rep 2020;21(3):1561–71.
[30] Wong HR, Shanley TP, Sakthivel B, et al. Genome-level expression profiles in pediatric septic shock indicate a role for altered zinc homeostasis in poor outcome. Physiol Genom 2007;30(2):146–55.
[31] Wynn JL, Cvijanovich NZ, Allen GL, et al. The influence of developmental age on the early transcriptomic response of children with septic shock. Mol Med 2011;17(11–12):1146–56.
[32] Raymond SL, Lopez MC, Baker HV, et al. Unique transcriptomic response to sepsis is observed among patients of different age groups. PLoS One 2017;12(9):e0184159.
[33] Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR; 1998.
[34] Jain S, Chalisgaonkar D. Setting up stage-discharge relations using ANN. J Hydrol Eng 2000;5(4):428–33.
[35] Jentzen A, Kröger T. Convergence rates for gradient descent in the training of overparameterized artificial neural networks with biases. 2021. *arXiv preprint arXiv:210211840*.
[36] Langfelder P, Horvath S. WGCNA package faq. 2017.