Shock, Publish Ahead of Print
DOI: 10.1097/SHK.0000000000002227

OPEN

Advancing the Understanding of Clinical Sepsis Using Gene Expression-Driven Machine Learning to Improve Patient Outcomes

Asrar Rashid[*1], Feras Al-Obeida[2], Wael Hafez[2,3], Govind Benakatti[4], Rayaz A Malik[5,6], Christos Koutentis[7], Javed Sharief[8], Joe Brierley[9], Nasir Quraishi[10], Zainab A Malik[11], Arif Anwary[1], Hoda Alkhzaimi[12], Syed Ahmed Zaki[13], Praveen Khilnani[14], Raziya Kadwa[8], Rajesh Phatak[15], Maike Schumacher[16], Guftar Shaikh[17], Ahmed Al-Dubai[1], Amir Hussain[1]

1. School of Computing, Edinburgh Napier University. Edinburgh, UK.

2. The National Research Centre, Egypt

3. College of Technology, Zayed University, Abu Dhabi, UAE

4. Yas Clinic, Abu Dhabi, UAE

5. Institute of Cardiovascular Science, University of Manchester. Manchester, UK.

6. Weill Cornell Medicine-Qatar, Doha, Qatar

7. Department of Anesthesiology, SUNY Downstate Medical Center

8. NMC Royal Hospital. Abu Dhabi, UAE.

9. Great Ormond Street Children's Hospital, London, UK.

10. Centre for Spinal Studies & Surgery, Queen's Medical Centre. The University of Nottingham. Nottingham, UK.

11. College of Medicine, Mohammed Bin Rashid University of Medicine and Health Sciences. Dubai, U.A.E.

12. New York University, Abu Dhabi, UAE.

13. All India Institute of Medical Sciences. Hyderabad, India

14. Medanta Gururam, Delhi, India.

15. Pediatric Intensive Care, Burjeel Hospital, Najda, Abu Dhabi

16. Sheikh Khalifa Medical City, UAE

17. Endocrinology, Royal Hospital for Children. Glasgow, UK.

*Corresponding author Dr. Asrar Rashid asrar.rashid@napier.ac.uk

Conflict of Interest Declaration

The authors declare that they have no affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

ABSTRACT

Sepsis remains a major challenge that necessitates improved approaches to enhance patient outcomes. This study explored the potential of Machine Learning (ML) techniques to bridge the gap between clinical data and gene expression information to better predict and understand sepsis. We discuss the application of ML algorithms, including neural networks, deep learning, and ensemble methods, to address key evidence gaps and overcome the challenges in sepsis research. The lack of a clear definition of sepsis is highlighted as a major hurdle, but ML models offer a workaround by focusing on endpoint prediction. We emphasize the significance of gene transcript information and its use in ML models to provide insights into sepsis pathophysiology and biomarker identification. Temporal analysis and integration of gene expression data further enhance the accuracy and predictive capabilities of ML models for sepsis. Although challenges such as interpretability and bias exist, ML research offers exciting prospects for addressing critical clinical problems, improving sepsis management, and advancing precision medicine approaches. Collaborative efforts between clinicians and data scientists are essential for the successful implementation and translation of ML models into clinical practice. ML has the potential to revolutionize our understanding of sepsis and significantly improve patient outcomes. Further research and collaboration between clinicians and data scientists are needed to fully understand the potential of ML in sepsis management.

KEYWORDS: Machine Learning, Sepsis, Gene- Expression, Septic Shock

Introduction

Sepsis is a global health challenge affecting individuals of all ages and underlying diseases in low-income, middle-income, and high-income countries[1]. Based on data from the United States, it is estimated that the global annual incidence of sepsis ranges from 15 million to 19 million cases. Despite significant morbidity and mortality associated with sepsis, a comprehensive understanding of its pathophysiology remains elusive. This complexity arises from the interplay between host response, pathogen virulence, and health system response. Existing knowledge gaps, particularly those tied to the disease's heterogeneity and multivariate data types associated with sepsis, pose a significant obstacle in creating systematic best practice guidelines for its management. Hypothesis-driven clinical studies, typically used to direct clinical practice, demand a pre-established framework for interfacing clinical data pertaining to specific questions. However, the inherent variability of sepsis data complicates traditional system modeling approaches, leading to a lack of precision and impeding systematic representation of sepsis. Since sepsis results from the body's immunological response to pathogens, a deeper understanding of the immune response mechanisms in sepsis is indispensable. However, this has been hampered by the difficulty in accurately modeling the disease. Therefore, novel research strategies are necessary, such as Machine Learning (ML), a branch of Artificial Intelligence, to improve our understanding of sepsis and reduce morbidity and mortality (Figure 1).

An international consensus proposed modifications to the 2005 adult sepsis definitions, characterizing sepsis as a "life-threatening organ dysfunction resulting from a dysregulated host response to infection." Septic shock was defined as "a subset of sepsis in which particularly profound circulatory, cellular, and metabolic abnormalities are associated with a greater risk of mortality than sepsis alone"[2]. Despite these revisions, persistent ambiguity has

hampered the development of guidelines and protocols for the management of clinical sepsis (Figure 2). This ambiguity persists despite decades of immune system research focusing on the host-pathogen response in sepsis. This could involve understanding the timing of infection initiation, discrepancies in infection load, type of organism, and variations in the age of the animal model among other factors. Researchers have attempted to control for sepsis heterogeneity by simplifying study effects, such as by reducing the complexity of study effects. However, such a reductionist approach may limit the applicability of these findings to the clinical context. For example, research model simplification may be counterproductive for adequately capturing the complexity and heterogeneity of sepsis, which may be essential for developing a wider application of therapies across the sepsis spectrum. In vitro studies allow for enhanced control over disease heterogeneity; however, they can complicate the process of back-extrapolation to the clinical context. The simplification of in vivo investigations can be initiated and implemented in complex biological systems in diverse ways; however, these methodologies hinge on bio-statistical methods, which can be resource intensive. Such scientific approaches are yet to yield the radical therapeutic advances required to affect global sepsis-related mortality. Therefore, new approaches are required to address sepsis and its heterogeneity to develop specific research criteria, milestones, and endpoints. There are still significant gaps in our understanding of the immunological, biochemical, molecular, and cellular changes that occur during sepsis, particularly those relating these factors to patients at the bedside.

Omics methodologies, including lipomics, proteomics, and transcriptomics, are broad-scale data-intensive techniques that offer a holistic view of biological systems. Langston et al. (2023) reviewed leukocyte phenotyping in sepsis using omics, functional analysis, and silicon modeling[3]. Omics provides a system-level view through simultaneous analysis of multiple

biological pathways. This approach has the potential to provide a comprehensive understanding of sepsis pathogenesis. However, the employment of 'omics' strategies to analyze sepsis data using statistical methods necessitates pre-established interpretative frameworks. Thus, relating omics data to clinical parameters noted in sepsis using traditional analytical techniques may not be straightforward. For example, temporal analysis of sepsis has suggests that mRNA gene expression techniques may not be viable for biomarker discovery[4]. However, using an ML model with the same data yielded predictive benefits [5]. Thus, ML is an alternative method to model data that does not require a predetermined understanding of either the data structure or variable relationships, thereby circumventing past statistical limitations. ML approaches may be useful for the early detection of sepsis, as suggested by Stolarski et al. (2022) in murine models, showing that it was possible to determine different sepsis phenotypes 6 and 24 h after infection[6].

In essence, for a computer to learn from the input data, it must be taught to identify sepsis and, ideally, do so promptly. For example, Akram et al. (2021) adopted ML using bedside physiological markers following temporal changes to predict early sepsis[7]. However, omics studies are a proxy for cellular processes and may enhance sepsis modeling, with high-throughput gene expression used to track changes in biological functions. Hence, to understand the importance of gene expression information in the context of ML, this narrative review first explored the issue of defining sepsis. Subsequently, the focus shifts towards sepsis identification and the importance of timing in this process. Finally, the application of ML algorithms to sepsis is discussed. The application of ML was demonstrated based on the transcriptomic sepsis literature selected using a systematic search strategy (Figure 3) and tabulated (Tables 1-3).

As part of this narrative, the temporal dynamics of sepsis are considered, which are crucial in modeling the sepsis trajectory from the initial detection. Within this narrative, it is essential to consider the temporal dynamics of sepsis as they play a critical role in modeling the trajectory of sepsis from its initial detection. By understanding the changing patterns and progression of sepsis over time, ML algorithms can be optimized to provide accurate and timely predictions, aiding in early intervention to improve patient outcomes.

Types of ML for the Computational Modelling of Sepsis, Bridging the Gap in Clinical Extrapolation

Machine learning (ML) algorithms can be broadly classified based on whether they utilize labeled or unlabeled data, leading to categories such as supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning leverages a labeled dataset to make predictions, which is common in classification and regression tasks. By contrast, unsupervised learning employs unlabeled data to decipher the data structure, which is frequently utilized in clustering tasks. These techniques have applications in managing sepsis, as detailed in Table 1A (unsupervised) and Table 1B (supervised). Semi-supervised learning leverages labeled and unlabeled data when the latter are abundant.

The categories described in this section cover many machine learning approaches. However, many other variations and hybrid methods have been developed.

Sepsis Definitions Quandaries and ML Workarounds

Shehab et al. (2022) provided a comprehensive review of Machine Learning (ML) and its applications in the medical field[8]. ML can be useful in addressing the limitations of traditional approaches in modeling sepsis complexity, which includes the important issue of a

suitable definition. Defining any condition is central to the progress of research and the application of statistical or ML approaches. However, ambiguity in the definition of sepsis represents a challenge, causing various disconnects, such as the misreporting of adult sepsis mortality rates [2]. Sepsis committees tasked with developing protocols for adult and pediatric sepsis face a dearth of evidence in many clinical areas, primarily because of the lack of a clear definition of sepsis. Multiple revisions have been implemented to improve adult sepsis guidelines, including those pf 1991, 2005, and 2016 [9,10]. The next iteration of the pediatric sepsis guidelines and definitions is keenly awaited. Moreover, a suitable definition that adequately encompasses all age groups remains elusive. This latter point could be related to the fact that sepsis research seldom crosses age boundaries set by medical specialists.

An optimal sepsis definition should encapsulate practical and objective insights concerning biological changes to quantify the alterations under investigation. Nonetheless, characterization of immunological shifts in sepsis in a universally applicable manner remains a scientific quandary. To deepen our understanding of the clinical practice paradigm, in which the patient (host) is the central figure affected by sepsis, one potential approach could be to depict the biological transformation of sepsis, either as an internal or external manifestation (Figure 4A).

Clinical scoring is an established part of hospital practice, and is useful for risk stratification. However, machine-learning models predict sepsis more accurately than clinical scores. The use of machine learning (ML) to connect to the complex and diverse temporal sepsis datasets, including gene expression data and clinical scores, is shown (Figure 5). Including temporal genomic data in ML modeling may provide a cellular perspective that would otherwise be lacking when using clinical data alone.

The absence of a clear definition of sepsis the paradoxical question of how ML can accurately interface with critical sepsis characteristics. To navigate this quandary, ML specialists have made assumptions in framing machine learning models, thereby circumventing the issue of a nebulous definition. For example, changes indicative of sepsis can be utilized in an ML approach by labeling events during the progression and characterization of clinical sepsis. This workaround becomes feasible as the application of ML to clinical sepsis primarily focuses on the endpoint, bypassing the necessity for forward hypothesis testing, typically demanded by statistical analysis. Currently, when designing a machine learning risk prediction model, ML operators must understand how to best define the clinical event (sepsis) to be predicted. In the future, ML algorithms that are pivotal in feature classification could prove invaluable in the pursuit of a pragmatic definition. The following section on temporal modeling underscores the value of ML in the context of sepsis.

Temporal Considerations of Sepsis Pathogenesis

Early intervention in sepsis is crucial for favorable outcomes, and delayed management is a significant prognostic risk factor. Consequently, the immediate administration of antibiotics and fluids upon clinical suspicion alone is strongly advocated by expert consensus. Timely diagnosis of sepsis is, therefore of utmost importance, and the ability to identify the condition several hours before its onset could potentially be lifesaving. Therefore, for ML to effectively impact sepsis outcomes necessitates early prediction is required during the clinical interactions [11].

However, sepsis heterogeneity has clinical implications, resulting in variations in treatment, timing of interventions, and differential host responses. Unfortunately, sepsis cannot be reduced to a simple, discrete phenomenon from clinical, immunological, and

pathophysiological viewpoints. The progression from infection to clinical sepsis is complex. To fully comprehend the evolution of sepsis, it is essential to understand its temporal dynamics from clinical and laboratory perspectives, including its ability to track the condition over time (Figure 6). Given its highly nonlinear and complex multivariable nature, sepsis is an ideal target for machine learning (ML) approaches.

As noted by Lauritsen et al. (2021), effective machine learning (ML) models for sepsis prediction require close collaboration between clinicians and data scientists [12]. Another challenge is the lack of a suitable immune biomarker for close temporal tracking, which limits the modeling precision. Temporal predictions in ML hinge on the specification of distinct time points, which can be facilitated by employing a range of time windows to construct temporal ML models (Figure 4C). Several studies have selected time windows of 48 h before and 24 h following Suspicion of Infection (SI) events, that is, when sepsis is initially suspected, whereas others have chosen a time frame of 24 h before and up to 12 h after SI as the window[13-15]. However, such temporal configurations have not yet been applied to sepsis using time-associated gene sequence information. Incorporating gene expression information could enrich a system-wide perspective by serving as a surrogate for cellular alterations at the molecular level.

An approach based on time windows has enabled sepsis researchers to make predictions without knowing the time of the infection onset. However, the original Adult Sepsis definitions were based on the ICD-9 and utilized subjective labels and definitions, thus not allowing for relevant time signposting for ML. In 2016, the definition of adult sepsis was modified to incorporate a temporal component, as reflected in the Sepsis-3 definition, with respect to the change in the SOFA score[2]. An increased SOFA score of greater than 2, a proxy

for a change in physiological state, was used to define multi-organ dysfunction syndrome, a feature of severe sepsis. As with the Sepsis-3 definition, parameterization should be incorporated into future enhancements of the sepsis definitions. The enhanced (Sepsis-3) definition provides a temporal milestone, a feature that allows the comparison of patient trajectories. However, the current sepsis-3 definition lacks objectivity of immunological or biochemical features, reflecting a gap in our understanding of the dynamics of sepsis pathophysiology.

Using Artificial Neural Networks for Enhanced Sepsis Biomarking and Temporal Analysis

ANNs are a subset of machine learning algorithms inspired by the structure and function of the human brain. They are proficient in recognizing patterns, interpreting sensory data, and identifying patterns in large and complex datasets. Artificial Neural Networks (ANN) simulate and solve distinct problems, particularly in pattern recognition and prediction tasks.

Artificial Neural Networks (ANNs) have been utilized to analyze sepsis microarray experiments, providing the advantage of working with small sample sizes (Table 2). However, most machine learning (ML) frameworks require determining the time of sepsis onset (T_0) or the initiation of a period when patterns are consistent with sepsis as a distinct entity. Kim et al. (2022) recently adopted this approach [16]. Dale et al. (2020) implemented ANNs using Long Short-Term Memory and multi-layer perceptrons for sepsis prediction [17]. The study used five time points with 11 simulated cytokine concentrations to forecast prospective cytokine trajectories, with the multilayer perceptron performing best when using 24-hour post-infection data. However, biomedical systems are stochastic, and incorporating randomness into clinical modeling is crucial for their validity. Zhang et al. (2016) attempted a

"back in time" approach in a primate study utilizing a mathematical cluster modeling technique known as "nearest neighbor" [18]. Two pig models were used to validate the methodology: one with surgically induced peritonitis and the other using an LPS infusion-induced approach. This study did not incorporate other immunological information in addition to biomarker assay levels. Combining temporal vital sign monitoring with a single biomarker measurement resulted in a highly accurate estimate of infection onset time. Additionally, the study assumed that the timing of the onset of infection was aligned with physiological changes indicative of sepsis.

Additional types of ANNs include recurrent neural networks (RNNs) and Convolutional Neural Networks (CNNs). Their potential usefulness in gene expression sepsis studies suggests promising avenues for exploration, which will be discussed here. However, the lack of existing studies citing the use of such ANNs in combination with gene expression profiling suggests the potential for future work in harnessing such algorithms. RNNs were designed to process the input sequential data. Because they use their internal (hidden) state to process input sequences, they possess a form of memory; thus, RNNs are a suitable choice for sequential datasets. Bedoya et al. (2020) applied dual multi-output Gaussian Processes (MGPs) with RNNs, also known as MGP-RNNs, to both dynamic and static clinical data related to sepsis prediction[19]. The likelihood of developing sepsis was computed within four hours of each marked time point. Different ML methods were compared using the same datasets, and MGP-RNN was found to be superior. Sheetrit et al. (2019) focused on time intervals instead of strict time points in their RNN methods[20]. An interval approach may be more desirable because clinical data are often multivariate and originate from different sources. This approach allowed the discovery of frequently repeated temporal patterns within the datasets, thereby creating a probabilistic distribution model of temporal patterns. A

Temporal Probabilistic Profile (TPF) was developed, allowing the prospective classification of new data and outcome prediction. Large benchmark clinical datasets were used to apply TPF, demonstrating improved sepsis prediction and enhanced performance compared to other machine learning models. Convolutional Neural Networks (CNNs) are another type of ANN designed to process data with a grid-like topology, which can be viewed as a 2D grid of pixels. Because of their proficiency in extracting spatial features, they are commonly used in image and video processing tasks. They are particularly adept at spatial hierarchies, and CNNs may also be useful in temporal modeling. Using temporal CNNs, Kok et al. (2020) developed an automated sepsis prediction tool that involved per-time-step metrics. This tool showed a high predictive capability for the development of sepsis (AUROC 98%)[21].

Deep Learning in Sepsis - Enhanced Modelling in Sepsis Using Transcript Information

Deep learning (DL) is a powerful subfield of machine learning (ML) that uses multilayered artificial neural networks to learn complex representations of data that are useful for dimensional reduction in genomic studies and for predictive modeling of large and complex biological datasets[22]. DL can be unsupervised, (semi)-supervised, or reinforcement learning-based[23]. Reinforcement learning is based on an agent gaining environmental feedback using a reward or penalization system. It is ideal for large, complex biological datasets and is not discussed within this article due to the paucity of sepsis studies using this method. DL models can be applied throughout the data processing pipeline, from data acquisition to gene and pathway enrichment, thereby improving biological data analysis. For example, DL has been used to re-examine images generated from previous microarray experiments, revealing imaging defects in many studies. For example, using a DL, Qin et al. (2021) re-examined the images generated from previous microarray experiments [24]. This study analyzed microarray-generated images based on fluorescent signals from previously published studies. The results

showed the presence of imaging defects in 26.73% of the microarray studies analyzed. In addition, DL may be a useful adjunct in the quality control process for testing the accuracy of data capture. Schaack et al. (2021) performed a meta-analysis of publicly available data series extracted from NCBI Gene Expression Omnibus and EMBL-EBI ArrayExpress to create a comprehensive meta-expression set[25]. They compared various ML methods against the traditional technique of differential gene expression (DGE) analysis. Consequently, Deep Learning was the most resilient among the tested methods, including DGE, random forest, support vector machine, and decision tree analysis. The DL classifiers allowed for the differentiation of patients with and without sepsis. They found that deep-learning neural networks performed the best, especially when the data were noisy or incomplete, highlighting the efficacy of DL models in facilitating sepsis modeling. Yuan et al. (2021) applied DL to sequential single-cell RNA data using a supervised method [26]. Gene interactions were predicted using 3D tensors and trained convolutional and recurrent neural networks (RNNs). The model accurately identified regulatory and causal gene-gene interactions and new gene function assignments.

The deduction of gene relationships predicated on differential gene expression necessitates an array of computational frameworks, extending from Pearson correlation to undirected graphical modeling. Nevertheless, such a stratified approach presents challenges for Deep Learning (DL) because unsupervised processing can mistakenly identify noise-associated genes as significant. To address these complications, Yuan et al. (2019) devised a DL convolutional neural network for co-expression applied to single-cell RNA data [27]. This innovative approach offers a methodology for inferring gene relationships from image-like objects produced from expression data, thereby facilitating the identification of causality, gene-disease predictors, and functional assignments.

A significant challenge associated with DL models is their black-box nature, which could hinder the future incorporation of ML in clinical contexts. This becomes particularly problematic when clinicians are expected to take responsibility for decisions influenced by ML and the underlying methodology is obscured owing to its inherent modeling complexity. To mitigate this black-box issue, Hanczar et al. (2020) proposed a DL approach grounded based on Layer-wise Relevance Propagation (LRP) [28]. LRP, a gradient method for neural network interpretation, identifies the most critical neuronal network responsible for predicting and pinpointing gene sets that activate the same neuron. Significant neurons and genes are subsequently mapped onto translational databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and the Disease Ontology Annotation List (DOLite), thereby offering a biological context. This methodology surpasses classical Machine Learning, which typically measures neuronal effectiveness using the weighted average of output connections. Nonetheless, the biological interpretation rendered is that of the model, which may not consistently align with actual biological parameters. Additional limitations include the fact that DL models primarily search for correlations between inputs and outputs rather than causality, and reliance on databases may introduce biases.

Various platforms are available for studying gene expression. Unfortunately, earlier platforms, such as Microarray and RNA-seq, produced averaged gene expression results because RNA is derived from many cells. Conversely, single-cell RNA sequencing (scRNA-seq) offers multiple expression profiles at the expense of generating substantial data, while maintaining the capability to concentrate on specific cell types. Large datasets obtained by scRNA-seq have introduced new computational challenges. However, they also make it an ideal application for Deep Learning Neural Networks, which require large amounts of data for effective learning. Deep-learning neural networks have been used to interpret mRNA gene

expression. For instance, solely from the gene sequences, Deep Convolutional Networks can predict 60-80% of human RNA abundance variation[29,30]. By applying Deep Learning principles, Magnusson et al. (2022) studied the impact of transcription factors on gene regulation [31]. This methodology is called the 'advancing past the black box' machine learning model. This allowed the prediction of the relationship between transcription factors and the target gene network, thus providing a mechanistic understanding of disease processes. Because the training was constrained, the derived predictive models were interpretable. In a related study, Yuan et al. (2021) used temporal scRNA-seq data and trained Deep Learning models, including recurrent neural networks and convolutional models, to identify regulatory and causal gene relationships and assign new functions to genes [26]. These studies demonstrate the potential of Deep Learning Neural Networks in understanding gene expression regulation. The authors suggest that this is the first step towards developing fully interpretable white-box models.

Deep Learning algorithms have also shown promise in analyzing medical time-series data despite the challenge of dealing with sensor- and noise-based errors [16]. However, small sample sizes can lead to overfitting, which can be addressed using self-supervised learning, transfer learning, or data augmentation [32]. Kim et al. (2022) applied a recurrent neural network to a time-series dataset using a neural architecture search method to optimize the architecture and a genetic algorithm approach to balance the computational resources and search efficiency[16]. An auto-encoder was also employed to denoise the data and improve the learning process. The model outperformed the standard clinical scores (SOFA, qSOFA, and SAPS II) and LSTM, and its performance decreased with extended prediction times, as indicated by the lower sensitivity, specificity, and AUROC values. Rafie et al. (2021) used a combined Deep Learning approach with LSTM and convolutional and fully connected layers

to improve the earliest time of sepsis prediction, achieving better AUROC values than other methods[33]. For ML to achieve wider application in clinical sepsis, its interpretability may be a happy middle ground. Another option is to combine ML algorithms with clinical inputs, as described in the next section on Ensemble ML.

Ensemble Learning Techniques - To Improve Clinical Modelling

Ensemble learning is a powerful computational technique that combines multiple models such as experts or classifiers to solve complex problems. This approach has been shown to improve the accuracy of predictions and can incorporate a clinician's expertise to guide the development and validation of machine-learning techniques. Ensemble techniques have broad applicability across various domains, including studying biological systems and analyzing sepsis as a disease process (Table 3). Ensemble machine learning aims to decrease the variance and bias associated with single models by incorporating multiple machine learning algorithms into a combined predictive model.

Ensemble techniques have been applied to various problems, including predicting cellular dynamics in biological systems and analyzing sepsis as a disease process. Ensemble methods combine processes to make sense of data inputs, including different neural networks with inherent strengths and weaknesses. These models may be structured in parallel or sequentially and can incorporate weightage or 'learning' from different models based on averaging or regression. Ensemble models may be useful for supervised tasks related to classification and unsupervised tasks related to clustering. One particularly novel approach in ensemble learning is the "expert in the loop" ensemble method, which incorporates the expertise of a clinician to guide the development and validation of machine learning techniques. This approach combines deductive analysis with inductive (data-driven) learning

and is particularly useful for optimal feature selection, error correction, and incremental learning tasks. Additionally, certain algorithms, such as the long short-term memory neural network (LSTM), are well-suited for capturing temporal relationships and can remember the sequencing of datasets over long periods. Additionally, the process of transforming weak learners into strong learners by forming a different algorithm for classifying rules is known as 'boosting.' Boosting is a key aspect of ensemble learning, and has been shown to improve model performance by reducing bias and variance.

Discussion

The discussion section of this paper illuminates key findings and potential implications of using Machine Learning (ML) techniques in the study and management of sepsis. These encompass potential opportunities, inherent challenges, and various ML applications in augmenting our understanding and handling of this complex condition. Sepsis is multifaceted and characterised by diverse etiologies and heterogeneous clinical manifestations. Conventional methods have limitations in deciphering the complex interplay between the host response, pathogen virulence, and various clinical factors. However, ML, given its ability to analyze extensive and varied datasets, including gene expression data, has emerged as a promising approach to addressing this complexity. Notably, ML algorithms such as neural networks and deep learning can discern patterns and relationships within data, facilitating more accurate predictions and insights into the pathogenesis of sepsis.

A persistent challenge is the ambiguity in the definition of sepsis, which complicates the modeling and analysis of sepsis data. However, ML models, offer a workaround by prioritizing endpoint prediction over a predefined understanding of the disease. The capability of these models to learn effectively from labeled and unlabeled data enables them

to discern sepsis patterns and categorize patients based on their clinical characteristics and gene expression profiles. Despite the lack of a definitive sepsis definition, ML models present a feasible solution by harnessing the predictive power of gene expression data.

An essential aspect of sepsis research is understanding its temporal dynamics for early detection and intervention. ML models can incorporate time-related information by considering specific time windows and data sequences. This is particularly relevant in sepsis, in which disease progression and host response evolve. Furthermore, gene expression data is instrumental in recording the cellular and molecular changes asscociated with sepsis. ML models that integrate this information can provide a holistic perspective of sepsis pathogenesis and aid in detecting biomarkers for early disease identification and monitoring [34].

One major challenge with ML models, particularly those based on deep learning algorithms, is their 'black-box' nature, which makes it difficult to interpret their decision-making processes. This issue could potentially obstruct the integration of ML models into clinical practice. However, ongoing endeavors are to enhance the interpretability of ML modules by leveraging techniques such as Layer-wise Relevance Propagation (LRP) and ensemble learning.

Despite the promising potential of ML in sepsis research, several limitations of this study need to be addressed. Small sample sizes, data representation bias, identifying causal relationships are important considerations. Future research should focus on developing interpretable ML models, validating the efficacy of ML in sepsis management through prospective studies, and integrating clinical expertise with ML algorithms using ensemble

techniques. The successful implementation and translation of ML models into clinical settings hinge on collaboration between clinicians, data scientists, and researchers.

Conclusion

ML offers a substantial potential for transforming our comprehension and management of sepsis. Although the intricate nature of sepsis poses considerable challenges, ML techniques propose pioneering solutions that amalgamate clinical data with gene expression data. These techniques facilitate the prediction, classification, and temporal analysis of sepsis. Notably, ML in sepsis research overcomes the lack of a universally applicable definition of sepsis, shifting the focus toward predicting endpoints and classifying patients. Despite the interpretability challenges owing to the black-box nature of ML algorithms, efforts are in progress to develop understandable "white-box" models. Further research is needed to fully understand the causal relationships in sepsis and develop more interpretable models.

By integrating gene expression data and temporal analysis, ML models aid in early disease detection and improve patient outcomes. Temporal considerations play a crucial role in sepsis management, and ML excels at capturing temporal relationships in complex datasets. In particular, deep learning neural networks have shown promise in analyzing temporal sequences and predicting sepsis outcome. ML models can facilitate early detection and intervention by integrating gene expression data and temporal analysis, ultimately improving patient outcomes. Although ML offers tremendous opportunities, challenges remain. Small sample sizes, overfitting, and the search for correlations rather than causality in deep learning models must be addressed using self-supervised learning, transfer learning, and data augmentation techniques. Ensemble learning techniques offer a powerful approach for enhancing the performance and robustness of ML models in sepsis research. By combining

multiple models and incorporating the expertise of clinicians, ensemble methods can improve prediction accuracy, feature selection, and incremental learning.

ML has immense potential to enhance our understanding of sepsis and boost patient outcomes by integrating clinical data, gene expression, and temporal analyses. Collaboration between clinicians and data scientists is crucial for the successful implementation of ML models in clinical practice, potentially leading to more personalized sepsis management strategies. With further advancements, ML can substantially contribute to revolutionizing sepsis care, expanding the implementation of precision medicine in sepsis, and reducing the devastating impact of this condition.

Pathogen Considerations: Sepsis is a complex disease initiated by various pathogens, either singularly or in combination (commensals or newly invading organisms). Different pathogens may have differing immunological host effects. Also, if the host is infected by one organism, this may weaken the host's immune system and ease secondary bacterial infection. This is known as the two-hit or, when multiple organisms are involved, a multi-hit hypothesis.

Host-Pathogen Interactions: The transformation from initial bacteremia or viremia to clinical infection is contingent upon host-pathogen interactions and the pace of progression through various stages, leading from infection to sepsis and eventually septic shock. Timely treatment is essential to reverse the infection trajectory. Delayed intervention and host susceptibility significantly influence treatment response and the patient's ability to recover from infection and respond to sepsis.

Port of Entry: The port of entry from infection to sepsis, including the invading entry site, is consequential to sepsis progression. In some individuals, pathogens may be commensals, but at the same site, in other individuals, the same organisms may be pathogenic. For example, Neisseria Meningitides resides in the throat as a commensal in some cases, whereas in others, it is responsible for invasive Meningococcal Sepsis and Meningitis.

Host Factors: Transformation of the initial bacteremia or viremia progressing to clinical infection depends on host-pathogen factors and the rate of progress through the different stages. Age may play an important role in sepsis in neonates, infants, and the elderly, causing higher morbidity and mortality. Host co-morbidities and genetic variation may also be key factors in determining disease progression.

Sepsis Treatment: There are only a few immunomodulators used in sepsis, with the mainstay of treatment being early intervention, according to the principles of 'Sepsis-6. ' Early antibiotic treatment and stabilization of bacterial sepsis remain cornerstones of treatment. Steroids may have value in septic shock as part of various treatments to maintain hemodynamics and treat Diffuse Intravascular Coagulopathy (DIC).

Treatment variation: Despite protocolized approaches, several factors, such as late patient arrival, clinical inexperience, and treatment inequities related to location or patient demographics may result in delayed sepsis treatment. Furthermore, aligning patients according to the onset of infection can be complex, leading to heterogeneity in the clinical presentation.

Genetic Implications for both the Host and Pathogen: The genetic profiles of both the host and pathogen significantly affect the development, progression, and treatment response of sepsis. Host genetics influence susceptibility to infection, response to treatment, and prognosis, largely due to gene variations associated with the immune response and drug metabolism. Pathogen genetics determine virulence, antimicrobial resistance, and adaptability, with certain genetic elements enabling evasion of the host's immune system, resistance to antibiotics, and survival under varying conditions. Understanding these genetic influences offers valuable insights for personalized sepsis management, although further research is required to translate these findings into clinical practice.

Implications for mRNA research: Owing to the heterogeneity of sepsis mRNA studies. A notorious challenge in high-throughput mRNA technologies is the portability of insights from one mRNA-based study to another. There can be issues with experimental variation owing to the platform itself and the data variance. This may be caused by differences in experimental techniques or external factors. Transcriptomic endotyping has been applied to sepsis for disease classification, particularly in complex disorders, to categorize patients into homogeneous subgroups based on the underlying biological or pathophysiological mechanisms that drive sepsis.

Sepsis Definition: The definition of sepsis has been adapted to mirror advances in the clinical field. However, such definitions are heavily dependent on clinical interpretation rather than on the immunological patterns of the disease. Differences in definitions can affect patient selection, diagnosis timing, and input and output characteristics.

Though High-throughput gene sequencing has been applied to sepsis, the heterogeneity of sepsis may lead to complex and diverse genetic data, challenging interpretation, and meaningful applications. Sophisticated bioinformatics and statistical approaches are often required to analyze and interpret these data. The advantage of ML is in the ability to model a complex, multivariable process, without a detailed understanding of disease mechanism.

Acknowledgments

The authors thank the anonymous reviewers for their insightful comments and suggestions. To Dr. David Inwald, PICU Addenbrookes Hospital, Cambridge, for reading and providing feedback on the article. Prof Hussain and Prof Al-Dubai acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) funded COG-MHEAR Programme (EPSRC Grant Reference: EP/T021063/1). Prof Hussain also acknowledges the support of UK EPSRC Grant No. EP/T024917/1. For Dr. Binu George, for keeping us on track with the study goals. Finally, and certainly, not least, Professor Hector Wong, whose decades-long contribution to the genomics of sepsis remains an enduring legacy.

References

1. Organization SWH. Improving the prevention, diagnosis and clinical management of sepsis. *WHO.* 2017(A70/13).
2. Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA.* 2016;315(8):801-810.
3. Langston JC, Yang Q, Kiani MF, Kilpatrick LE. LEUKOCYTE PHENOTYPING IN SEPSIS USING OMICS, FUNCTIONAL ANALYSIS, AND IN SILICO MODELING. *Shock.* 2023;59(2):224-231.
4. 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.
5. Rashid A, Anwary AR, Al-Obeidat F, et al. Application of a gene modular approach for clinical phenotype genotype association and sepsis prediction using machine learning in meningococcal sepsis. *Informatics in Medicine Unlocked.* 2023:101293.
6. Stolarski AE, Kim J, Nudel J, Gunn S, Remick DG. Defining Sepsis Phenotypes—Two Murine Models of Sepsis and Machine Learning. *Shock.* 2022;57(6):268-273.
7. Mohammed A, Van Wyk F, Chinthala LK, et al. Temporal Differential Expression of Physiomarkers Predicts Sepsis in Critically Ill Adults. *Shock.* 2021;56(1):58-64.
8. Shehab M, Abualigah L, Shambour Q, et al. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine.* 2022;145:105458.
9. Gül F, Arslantaş MK, Cinel İ, Kumar A. Changing Definitions of Sepsis. *Turk J Anaesthesiol Reanim.* 2017;45(3):129-138.
10. Obonyo NG, Schlapbach LJ, Fraser JF. Sepsis: Changing Definitions, Unchanging Treatment. *Frontiers in Pediatrics.* 2019;6.

11. Lin P-C, Chen K-T, Chen H-C, Islam MM, Lin M-C. Machine Learning Model to Identify Sepsis Patients in the Emergency Department: Algorithm Development and Validation. *Journal of Personalized Medicine.* 2021;11(11):1055.

12. Lauritsen SM, Thiesson B, Jorgensen MJ, et al. The Framing of machine learning risk prediction models illustrated by evaluation of sepsis in general wards. *NPJ Digit Med.* 2021;4(1):158.

13. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Front Med (Lausanne).* 2021;8:607952.

14. Desautels T, Calvert J, Hoffman J, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Med Inform.* 2016;4(3):e28.

15. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Crit Care Med.* 2018;46(4):547-553.

16. Kim JK, Ahn W, Park S, Lee SH, Kim L. Early Prediction of Sepsis Onset Using Neural Architecture Search Based on Genetic Algorithms. *Int J Environ Res Public Health.* 2022;19(4).

17. Larie Dale AG, Cockrell Chase. Artificial neural networks for disease trajectory prediction in the context of sepsis. *Cornell University* 2020; https://arxiv.org/abs/2007.14542, 2020.

18. Zhang LA, Parker RS, Swigon D, et al. A One-Nearest-Neighbor Approach to Identify the Original Time of Infection Using Censored Baboon Sepsis Data. *Critical care medicine.* 2016;44(6):e432-e442.

19. Bedoya AD, Futoma J, Clement ME, et al. Machine learning for early detection of sepsis: an internal and temporal validation study. *JAMIA Open.* 2020;3(2):252-260.

20. Sheetrit E, Nissim N, Klimov D, Shahar Y. Temporal Probabilistic Profiles for Sepsis Prediction in the ICU. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019; Anchorage, AK, USA.

21. Kok C, Jahmunah V, Oh SL, et al. Automated prediction of sepsis using temporal convolutional network. *Computers in Biology and Medicine.* 2020;127:103957.

22. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. *Nature Genetics.* 2019;51(1):12-18.

23. Mahmud M, Kaiser MS, McGinnity TM, Hussain A. Deep Learning in Mining Biological Data. *Cognit Comput.* 2021;13(1):1-33.

24. Qin Y, Yi D, Chen X, Guan Y. Deep learning identifies erroneous microarray-based, gene-level conclusions in literature. *NAR Genom Bioinform.* 2021;3(4):lqab089.

25. Schaack D, Weigand MA, Uhle F. Comparison of machine-learning methodologies for accurate diagnosis of sepsis using microarray gene expression data. *PLoS One.* 2021;16(5):e0251800.

26. Yuan Y, Bar-Joseph Z. Deep learning of gene relationships from single cell time-course expression data. *Brief Bioinform.* 2021;22(5).

27. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci U S A.* 2019.

28. Hanczar B, Zehraoui F, Issa T, Arles M. Biological interpretation of deep neural network for phenotype prediction based on gene expression. *BMC Bioinformatics.* 2020;21(1):501.

29. Agarwal V, Shendure J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* 2020;31(7):107663.

30. Zrimec J, Borlin CS, Buric F, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat Commun.* 2020;11(1):6141.

31. Magnusson R, Tegner JN, Gustafsson M. Deep neural network prediction of genome-wide transcriptome signatures - beyond the Black-box. *NPJ Syst Biol Appl.* 2022;8(1):9.

32. Couckuyt A, Seurinck R, Emmaneel A, et al. Challenges in translational machine learning. *Hum Genet.* 2022.

33. Rafiei A, Rezaee A, Hajati F, Gheisari S, Golzan M. SSP: Early prediction of sepsis using fully connected LSTM-CNN model. *Comput Biol Med.* 2021;128:104110.

34. Abbas M, El-Manzalawy Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. *BMC Med Genomics.* 2020;13(1):122.

35. Burnham KL. Shared and distinct aspects of the sepsis transcriptomic response to fecal peritonitis and pneumonia. *Am J Respir Crit Care Med.* 2017;196.

36. Scicluna BP, van Vught LA, Zwinderman AH, et al. Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study. *Lancet Respir Med.* 2017;5(10):816-826.

37. Long Q, Li G, Dong Q, Wang M, Li J, Wang L. Exploration of the Shared Gene Signatures between Myocardium and Blood in Sepsis: Evidence from Bioinformatics Analysis. *Biomed Res Int.* 2022;2022:3690893.

38. Chen Y, Wang X, Wang J, Zong J, Wan X. Revealing novel pyroptosis-related therapeutic targets for sepsis based on machine learning. *BMC Med Genomics.* 2023;16(1):23.

39. Li M, Huang H, Ke C, et al. Identification of a novel four-gene diagnostic signature for patients with sepsis by integrating weighted gene co-expression network analysis and support vector machine algorithm. *Hereditas.* 2022;159(1):14.

40. Tang BM, McLean AS, Dawes IW, Huang SJ, Lin RC. The use of gene-expression profiling to identify candidate genes in human sepsis. *Am J Respir Crit Care Med.* 2007;176(7):676-684.

41. Saraiva JP, Oswald M, Biering A, et al. Fungal biomarker discovery by integration of classifiers. *BMC Genomics.* 2017;18(1):601.

42. 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):1561-1571.

43. Ghalwash MF, Ramljak D, Obradović Z. Patient-specific early classification of multivariate observations. *Int J Data Min Bioinform.* 2015;11(4):392-411.

44. Mayhew MB, Buturovic L, Luethy R, et al. A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections. *Nat Commun.* 2020;11(1):1177.

45. Li B, Zeng Q. Personalized identification of differentially expressed pathways in pediatric sepsis. *Mol Med Rep.* 2017;16(4):5085-5090.

46. Chen G, Han N, Li G, et al. Prediction of feature genes in trauma patients with the TNF rs1800629 A allele using support vector machine. *Comput Biol Med.* 2015;64:24-29.

47. Wu Y, Zhang L, Zhang Y, Zhen Y, Liu S. Bioinformatics analysis to screen for critical genes between survived and non-survived patients with sepsis. *Mol Med Rep.* 2018;18(4):3737-3743.

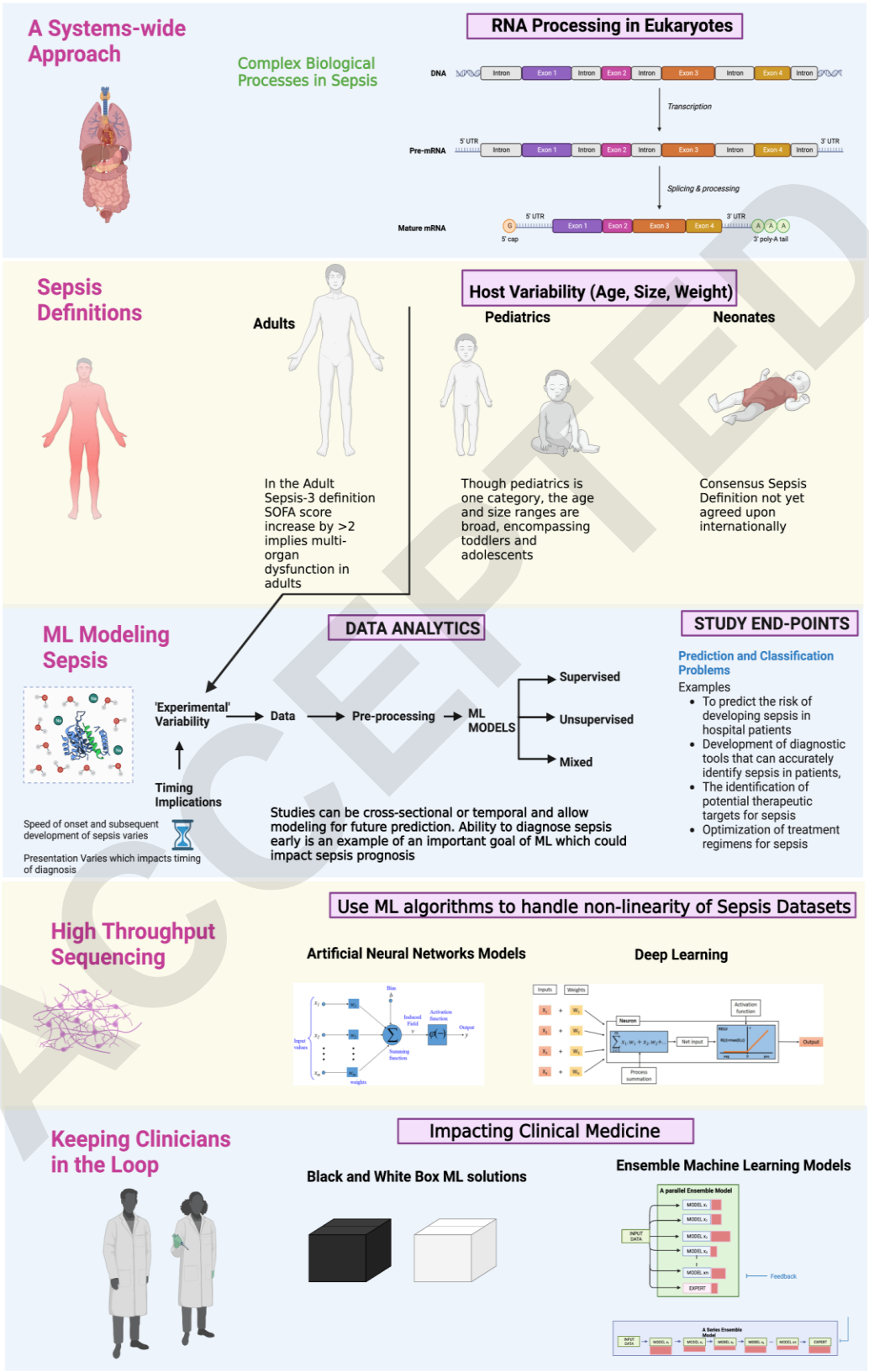
48. Chen Z, Zeng L, Liu G, et al. Construction of Autophagy-Related Gene Classifier for Early Diagnosis, Prognosis and Predicting Immune Microenvironment Features in Sepsis by Machine Learning Algorithms. *J Inflamm Res.* 2022;15:6165-6186.

49. Tong DL, Kempsell KE, Szakmany T, Ball G. Development of a Bioinformatics Framework for Identification and Validation of Genomic Biomarkers and Key Immunopathology Processes and Controllers in Infectious and Non-infectious Severe Inflammatory Response Syndrome. *Front Immunol.* 2020;11:380.

50. de Jong TV, Guryev V, Moshkin YM. Estimates of gene ensemble noise highlight critical pathways and predict disease severity in H1N1, COVID-19 and mortality in sepsis patients. *Sci Rep.* 2021;11(1):10793.

51. Bandyopadhyay S, Lysak N, Adhikari L, et al. Discovery and Validation of Urinary Molecular Signature of Early Sepsis. *Crit Care Explor.* 2020;2(10):e0195.

Figure 1. A Gene Expression-Based Machine Learning Approach for Systems-Level Analysis of Sepsis.



i. Sepsis is a complex process where a system-level approach provides an overview of multiple disease mechanisms. This can be applied as a part of an omics strategy, such as using gene expression information - a transcriptional approach. Data can be handled through various methods that are either statistical or machine learning (ML)-based, or a variation of the two. This enables the representation of many genes to be analyzed and modeled for a system-wide interpretation. ML can be ideal for handling scientific complexity where a suitable framework for understanding and objectively interpreting sepsis data is lacking. Gene expression studies have focused on ribonucleic acid (RNA), a template for an associated deoxyribonucleic acid (DNA) code. Such an approach, using ML, may be useful for sepsis prediction and classification problems.

ii. Sepsis analysis and modeling depend upon an accurate definition. For adults, the definition of sepsis was revised in 2016 (Sepsis-3). This provides more objectivity to the definition than earlier definitions, such as those based on ICD-9. Unlike earlier definitions, Sepsis-3 involves the quantification of severe organ dysfunction associated with a decrease in SOFA score of greater than 2, which implies a definite physiological change. Although a definition for pediatric sepsis exists, it differs from that of adult sepsis, and for neonates, no single definition has been agreed upon.

iii. Gene expression datasets can be modeled using different approaches. The classical ML approach includes supervised, unsupervised, and mixed methods. A gene-centric approach is useful because it delays the mapping stage until the end, thereby preserving the gene expression data structure as much as possible. Certain ML have adopted methods related to sepsis prediction and classification endpoints. One example is the use of machine-learning algorithms to predict the risk of developing sepsis in hospitalized patients. Further, Machine

learning algorithms can be trained using data from electronic health records and other sources to identify the patterns and risk factors that predict sepsis. This can allow healthcare providers to intervene early and potentially prevent sepsis progression. Other applications of machine learning in sepsis include the development of diagnostic tools that can accurately identify sepsis, identify potential therapeutic targets, and optimize treatment regimens for sepsis. Machine learning is useful for modeling large sepsis datasets and assisting with prediction and classification problems.

iv. High-throughput gene sequencing techniques such as microarray, RNA-seq, and sc-RNA-seq generate large datasets. Such datasets are ideal for artificial neural networks (ANN), which use the concept of DL to solve classification and prediction problems.

v. The challenge in applying ML to sepsis is that the solutions are mainly black-box because the 'workings' are hidden from the clinician. The black-box nature of ML may be a reason for the paucity of ML in prospective clinical trials. Interpretable (white-box) solutions can also be formulated using ML, which may improve the palatability of ML when applied to sepsis. Another option is to use clinician-in-the-loop combined with ML algorithms in what is known as an Ensemble ML model.

Annotation created with BioRender.com

Figure 2. Challenges of Genomic Research in Sepsis in Relation to Host-pathogen Factors and Implications for High-throughput Gene Experiments.

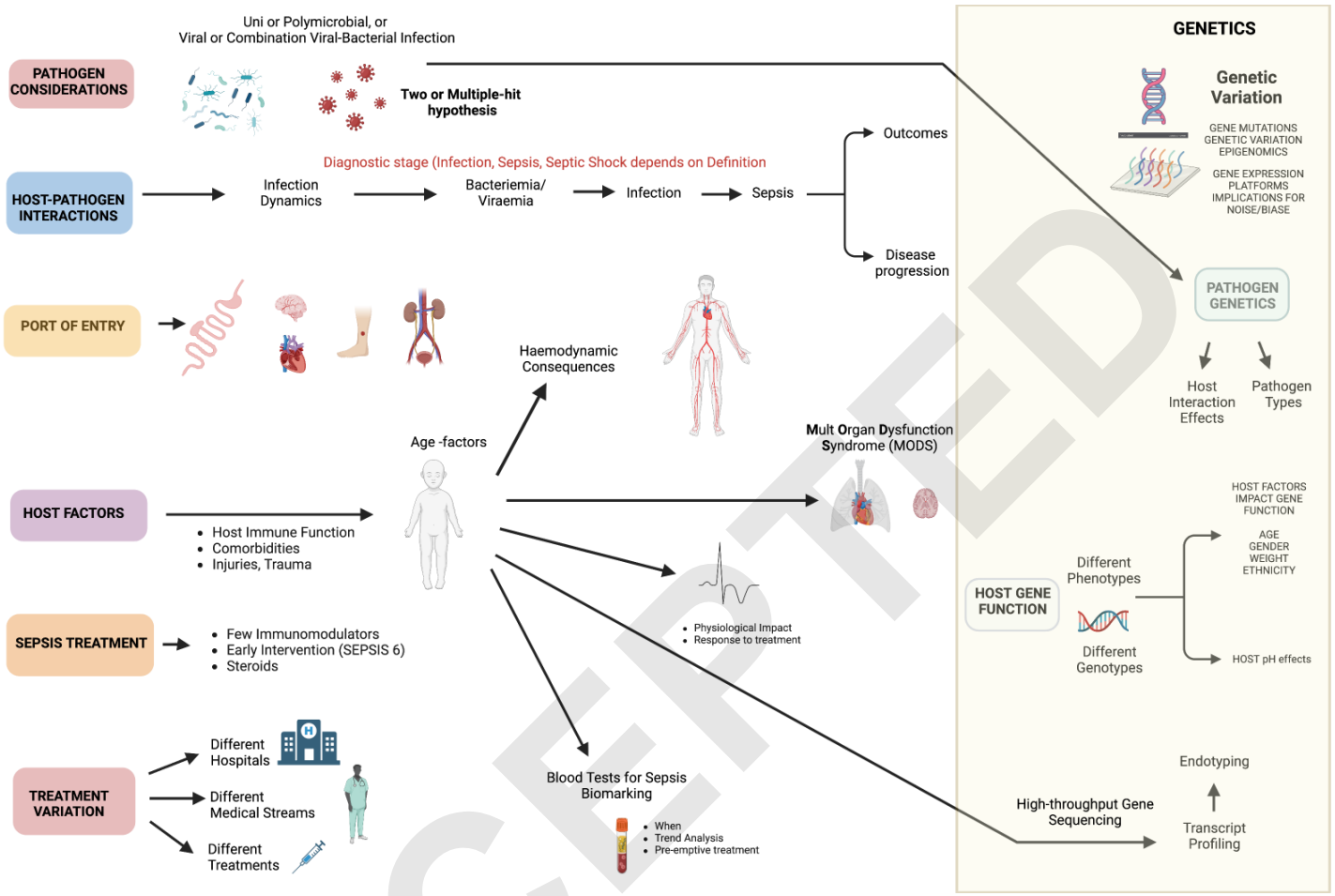


Figure 2. This figure outlines the 'Challenges of Sepsis Research,' including the application of genomic analysis. The heterogeneity caused by host and pathogen factors, as illustrated, impacts the portability of sepsis research across different sepsis studies.

Figure 3: Study Search Criteria

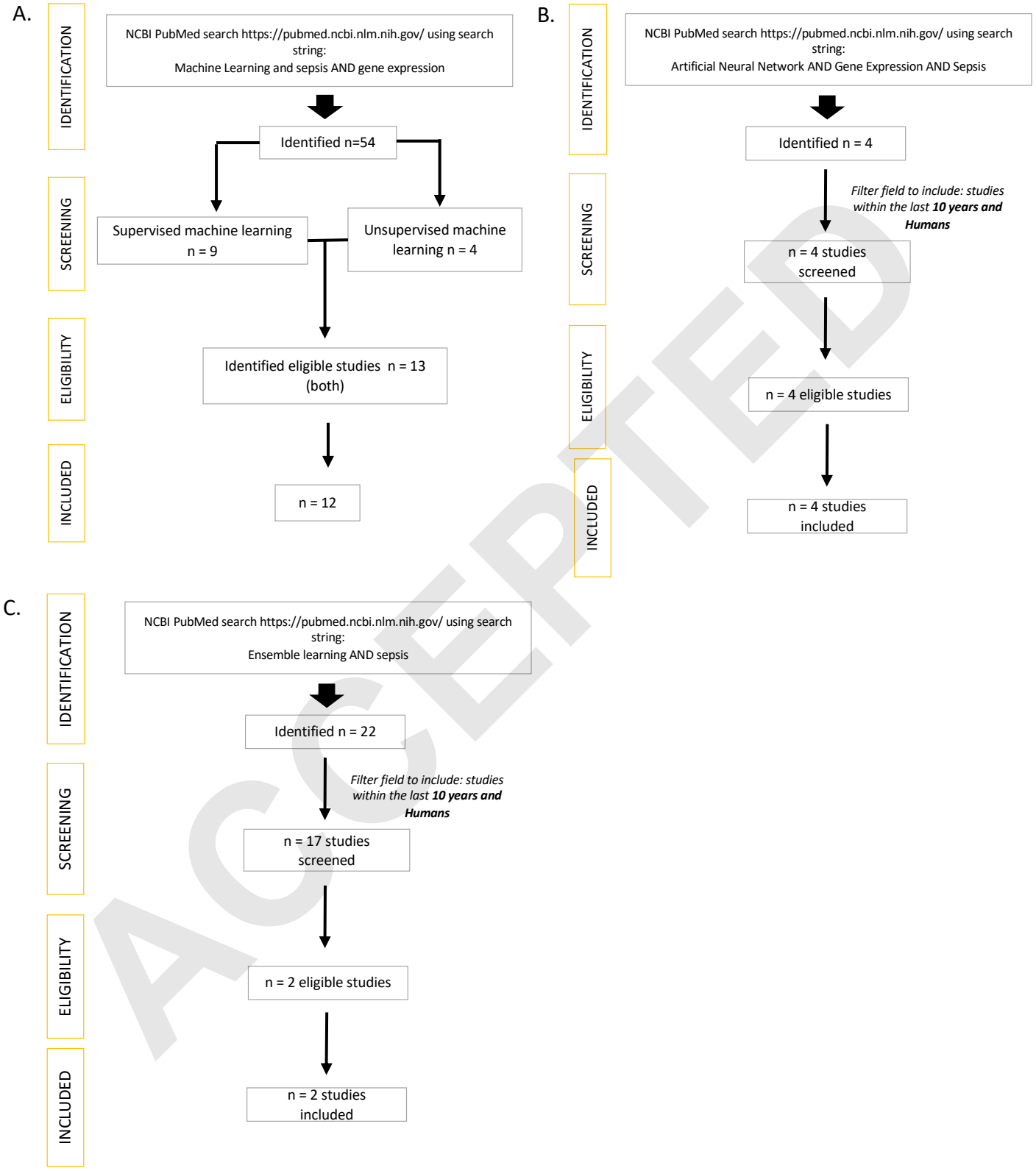


Figure 3. This paper adopted a comprehensive search strategy for the narrative review. Studies were identified and screened for keywords using the search on the PubMed website

maintained by the National Center for Biotechnology Information (NCBI). The eligibility criteria were to include studies in the last 10 years in human subjects, with the final inclusion criteria selecting research related to gene expression. Supervised and Unsupervised Machine learning studies were selected after the search as shown (**2A),** identifying 53 studies**.** **'**Supervised' or 'Unsupervised' was added to the start of the search term, allowing filtering into eligible studies consisting of Supervised (n=9) and Unsupervised groups (n=3). The NCBI search was also undertaken for Artificial Neural Network Studies identifying 4 studies that were all eligible to be included in this paper (**2B**). Ensemble Machine learning studies were selected from the past 10 years, with 18 studies being identified, of which 10 were from the last 10 years (**2C**); 7 studies were deemed eligible for inclusion.

Figure 4. Enhancing ML-based Prediction by Closing the Gap Between Internal and External Features Associated with the Patient Sepsis Journey

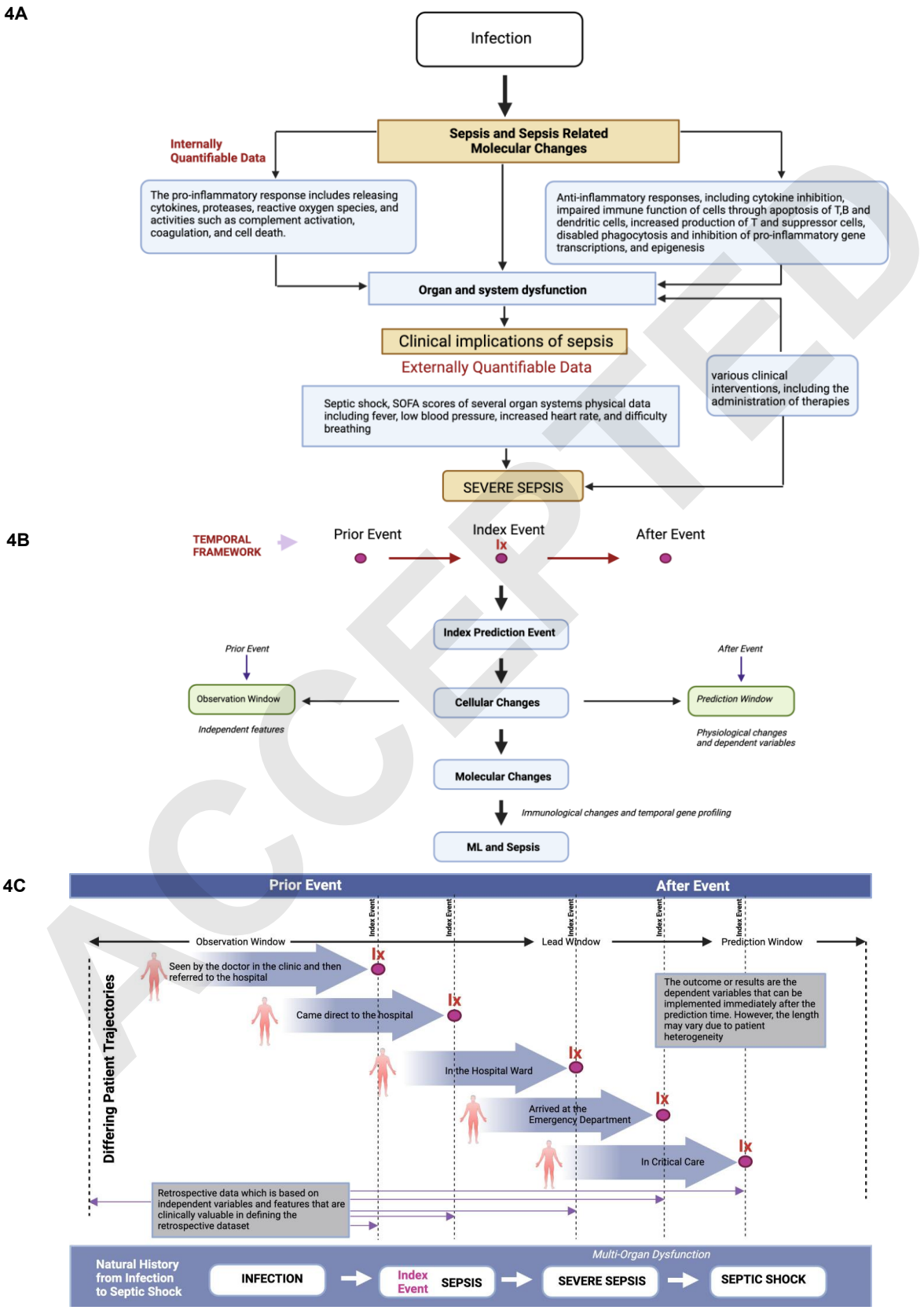


Figure 4. Challenges of Genomic Research in Sepsis in Relation to host-pathogen factors and Implications for High-throughput gene experiments.

Block figure showing internal and external features. Internally Quantifiable data (IQ-data) include molecular, DNA, mRNA, cellular, lipid, or proteomic signals. External data (EQ-data) encompass clinical descriptors, such as physiological and clinical scores, and physical textual data. Data were categorized as Internal and External from the patients' perspectives. Internal data were quantified through blood testing and external data were obtained through various data collection processes, such as clinical observations (Figure **4A**).

ML specialists have defined time windows to aid in understanding temporal dynamics. The observation window is a retrospective period before the index event, based on independent variables or features. By contrast, the prediction window samples the outcome or event, a dependent variable from which the outcome is derived. This allows the construction of various temporal framework structures using the chosen machine learning models. In addition, a discussion of the different components of the temporal framework is included. This incorporates the index prediction event (sepsis) and the timeframe windows before and after the event. The prediction time was calculated using the model, which, in this case, involved the time of sepsis diagnosis. The observation window is a retrospective period before the index event and relies on independent variables or features. The prediction window then samples the outcome of the event. This period was the dependent variable from which outcomes were derived. The prediction window starts after the prediction time or is delayed. When there is a prediction delay, this is called the lead window, also known as the gap window. This allowed various temporal framework structures to be constructed using the selected ML models (Figure **4B**).

Labeling key events can help derive key milestones in the patient journey. Red indicates the index event (I), which was the diagnosis of sepsis. The timing of the index event varies according to the different patient journeys, and as outlined in the diagram, patients can present with sepsis in different clinical contexts. Patient variability is governed by patient help-seeking behaviors and the dynamics of the health system. ML can be applied at three locations in a hospital setting: the emergency department, ward, or critical care area. The pathogenesis from infection to sepsis and then to septic shock varies according to multiple factors. Thus, the relationship between sepsis diagnosis and ensuing complications can be highly variable (Figure **4C**).

Annotation created with BioRender.com

Figure 5. Sepsis Time and Gene Expression

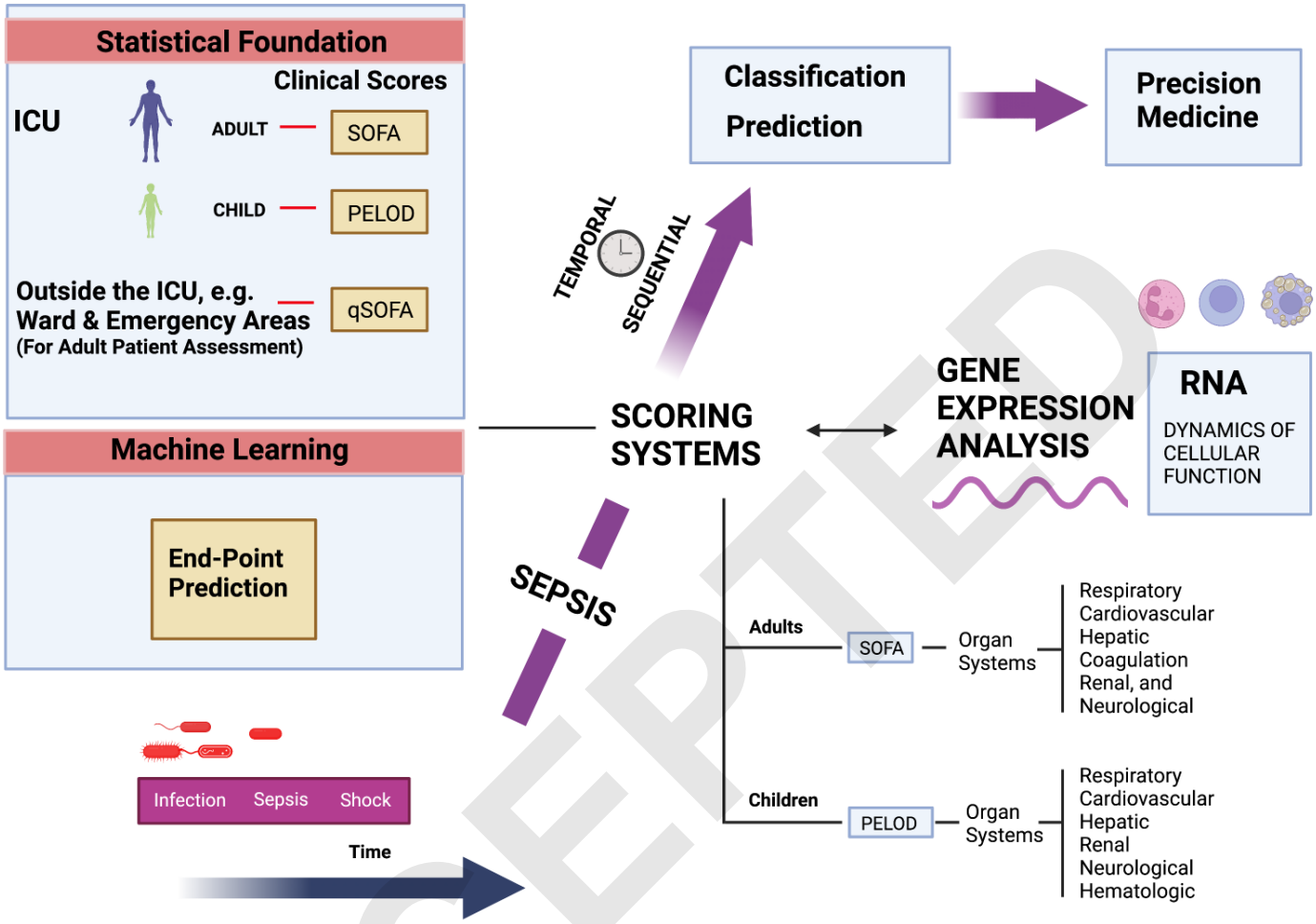


Figure 5. This figure illustrates the application of transcriptomics in correlating cellular changes with critical clinical parameters in sepsis research. Clinical scoring systems such as the SOFA, qSOFA, and PELOD scores are used to stratify patients' risk and monitor disease progression. The SOFA score is employed in adult ICUs to evaluate six organ systems and the severity of a patient's illness, with higher scores indicating greater dysfunction and mortality risk. The qSOFA score is a rapid assessment tool for identifying sepsis risk outside the ICU based on low blood pressure, high respiratory rate, and altered mental status. The PELOD score, designed for pediatric ICUs, assesses the severity of organ dysfunction in critically unwell children, with higher scores signifying severe dysfunction and increased risk of mortality death. These scores can be usefully linked to cellular function based on the

proxies provided by gene expression data. Gene expression information is often derived from the peripheral white cells of patients with sepsis, drawn from patient blood sampling. Temporal sequential data may then be used to understand underlying disease processes or to classify and predict sepsis-associated phenomena.

Annotation created with BioRender.com

Figure 6. Sepsis and the Concept of Time

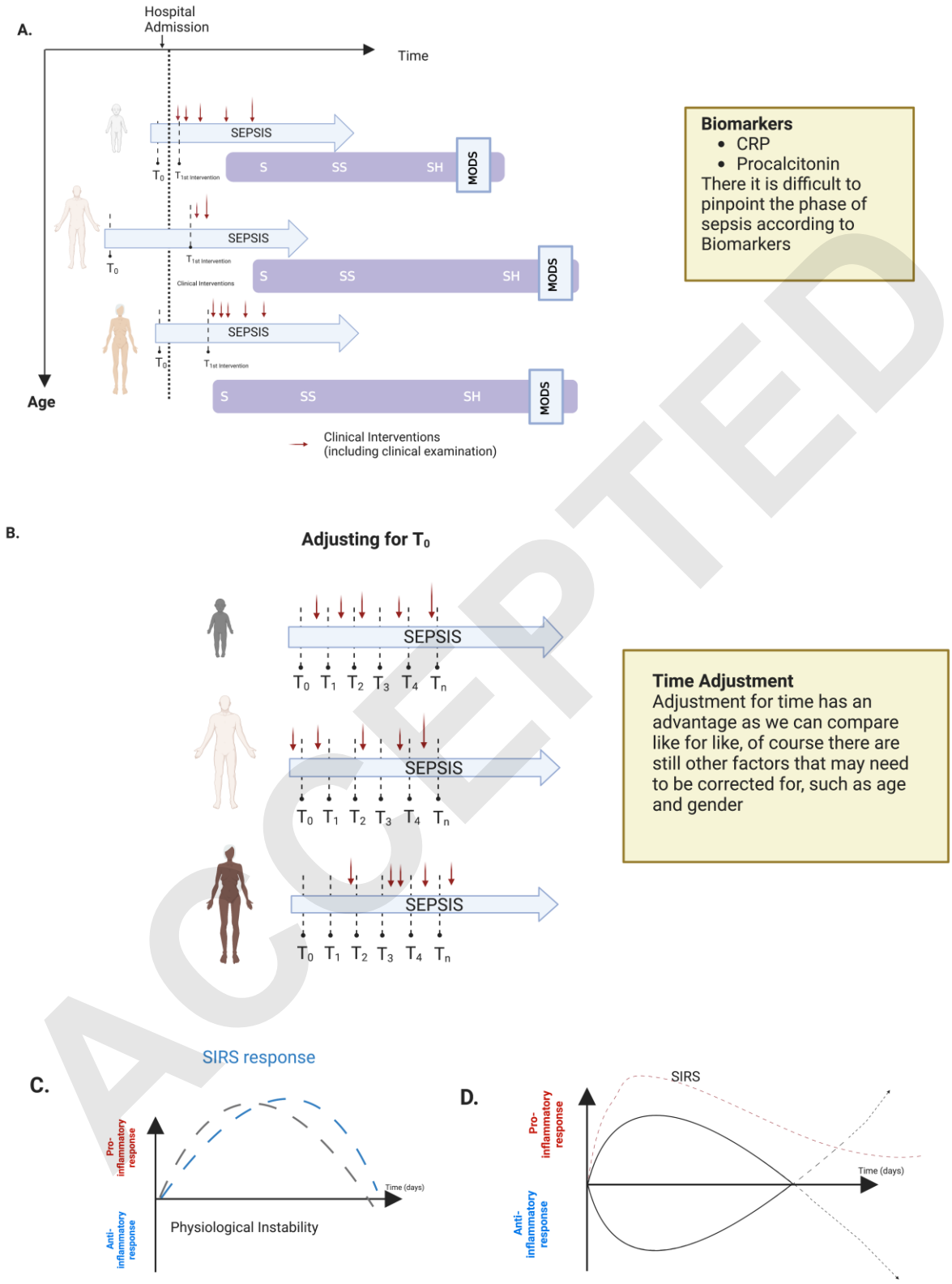


Figure 6. A Temporal appreciation of Sepsis is illustrated. Acute sepsis presents a complex challenge due to host variability, as depicted in **Figure 6A**. Further, treating neonates, infants, and the elderly presents additional difficulties, as they may display physiological instability during the early phase of sepsis. The diagram illustrates that patients have differing time characteristics in relation to arriving in the hospital setting. Temporal variations can occur due to age differences, comorbidities, and other factors. Significant delays in the patient journey can result in adverse and terminal outcomes. A blue arrow in the diagram represents the sepsis process and its evolution, with patients arriving at different time points along this trajectory. One key issue is that it is impossible to signpost the progression from bacteremia to shock against immunological milestones. Sepsis can progress from sepsis (**S**) to severe sepsis (**SS**) and then to septic shock (**SH**). The blue arrow depicts the sepsis process and its evolution. To better understand the immunopathological aspects of sepsis, some researchers have conducted time-course experiments using peripheral blood sampling. In such experiments, a sequence of blood collections is obtained from individual patients (e.g., Patients 1, 2, or 3 in Figure A), starting from the initial sample collected at admission (labeled 'time zero' or T_0). Subsequent samples can then be compared to the T_0 sample to identify differential gene expression. In some cases, a control sample may be nominated from any of the patient's samples, such as when a patient is physiologically stable; that sample is considered akin to a control. However, control samples can only be used for that patient, as each patient will have their own control. However, as illustrated, timing gene expression experiments in clinical sepsis is challenging, as patients arrive at various time points along the sepsis trajectory. Moreover, the inability to accurately time or categorize sepsis from an immunopathological perspective in the clinical setting adds to the complexity. Differences in therapies and experimental sites and the potential influence of clinical treatment on disease trajectory and transcriptomic profiles can also impact gene expression results. Researchers

can focus on a specific pathogen, age group, and treatment plan to mitigate these potential confounding factors. However, this approach may limit the sample size. Further research is needed to develop effective interventions and protocols that can improve patient outcomes while accounting for the heterogeneity of the patient population and potential confounding factors in gene expression experiments. **Figure 6B.** To effectively mitigate the impact of experimental design in clinical sepsis studies, it is crucial to consider the temporal dynamics of the disease. Sequential sampling is essential since patients may transition between different phases of sepsis during their illness. For example, up to 50% of patients have been shown to exhibit endotype switching within the first five days of ICU admission [35]. However, determining the optimal timing for sample collection presents a significant challenge. Addressing the challenge of Time Zero (T_0) through standardization could be useful from a temporal sepsis research perspective. Despite efforts to capture changes through regular sampling, constructing temporally-resolved clinical studies is beset by numerous challenges. A systemic inflammatory response syndrome was described because of the predilection of acute sepsis to cause physiological instability (**Figure 6C**). The different trajectories are shown related to host factors, such as host age, the timing of diagnosis, etc. As illustrated, sepsis was initially thought to be solely related to an inflammatory component, with anti-inflammation not featuring in early disease models. However, transcriptomic work has suggested that both components may co-exist (**Figure 6D**). The eventual summated trajectory (red dashed line) could vary according to the degree and timing of the pro and anti-inflammatory components.

Annotation created with BioRender.com

Table 1A. Unsupervised Machine Learning

STUDY	DESCRIPTION	RESULTS	REF*
Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study.	This study identified biologically relevant molecular endotypes in patients with sepsis. The study involved consecutive patients admitted to two intensive care units in the Netherlands and 29 ICUs in the UK. Genome-wide blood gene expression profiles were generated and analyzed using unsupervised consensus clustering and machine learning.	Four molecular endotypes designated Mars1-4, were identified in the discovery cohort and were associated with 28-day mortality. The worst outcome was found for patients classified as having a Mars1 endotype, with 39% of 90 people dying at 28 days.A 140-gene expression signature reliably stratified patients with sepsis to the four endotypes. Only Mars1 was consistently significantly associated with 28-day mortality across the cohorts. A biomarker was derived for each endotype, and BPGM and TAP2 reliably identified patients with a Mars1 endotype. This study provides a method for molecular classifying patients with sepsis to four different endotypes upon ICU admission, potentially aiding in personalized patient management and trial selection.	36
Exploration of the Shared Gene Signatures between Myocardium and Blood in Sepsis: Evidence from Bioinformatics Analysis.	This study using bioinformatics and machine learning methods, identified 1,049 genes commonly changed in the blood and myocardium of septic patients. Up-regulated genes were related to inflammation pathways, while down-regulated genes were related to mitochondrial and aerobic metabolism. The study divided 468 sepsis patients into two groups based on mortality-related commonly	A six-gene model was obtained, which performed well in classifying groups and predicting mortality. The study highlighted the potential of genes as biomarkers for septic cardiomyopathy and the potential impact of co-occurring pathological processes on sepsis prognosis.	37

	changed genes.		
Revealing novel pyroptosis-related therapeutic targets for sepsis based on machine learning.	This study aimed to uncover pyroptosis genes associated with sepsis and provide early therapeutic targets for treatment. The GSE134347 dataset was used to mine sepsis-related genes, and a protein-protein interaction (PPI) network was constructed. Unsupervised consensus clustering of sepsis patients was performed, and machine learning prediction models were used to identify PRGs mostly associated with sepsis. The prolactin signaling pathway and IL-17 signaling pathway were the primary enrichment pathways.	Unsupervised consensus clustering of sepsis patients was performed, and machine learning prediction models were used to identify PRGs mostly associated with sepsis. The prolactin signaling pathway and IL-17 signaling pathway were the primary enrichment pathways. NLRC4, the PRG most strongly associated with sepsis, was considered a potential target for treatment. The ceRNA network around NLRC4 could serve as a further research direction to uncover the deeper pathogenesis of sepsis.	38

TABLE 1A. Above showed are examples of unsupervised ML approach in sepsis research as data labelling is not required. That is, allowing analysis without a pre-conceived understanding of a disease mechanism. This methodology can be helpful in sepsis management at various stages in the patient journey. Once the analysis is generated, one can engage with the mapped to understand clusters, groupings, or the utility of prediction models.

*REF is the Reference citation in the literature

Table 1B. Supervised Machine Learning

STUDY	DESCRIPTION	RESULTS	REF
Identification of a novel four-gene diagnostic signature for patients with sepsis by integrating weighted gene co-expression network analysis and support vector machine algorithm.	This study aimed to identify sepsis-related diagnostic genes using integrated analysis, weighted gene co-expression network analysis, and gene regulatory networks. Results showed a significantly lower immune score in patients with sepsis compared to normal samples.	The identified genes were associated with functions like neutrophil degranulation, activation, and immunity. The study also identified a four-gene signature, including hub genes LCK, CCL5, ITGAM, and MMP9, which could be used to diagnose patients with sepsis.	39
The use of gene-expression profiling to identify candidate genes in human sepsis.	A genomewide study examined gene-expression profiling of neutrophils to identify signature genes and pathways in sepsis clinical syndrome. The study used oligonucleotide microarrays on peripheral blood samples of 94 critically ill patients.	The molecular signature of sepsis was generated from 44 samples and validated in 50. The diagnostic performance was high, regardless of age, comorbidities, or antibiotic treatment. The study found that genes involved in immune modulation and inflammatory response had reduced expression in patients with sepsis, with the activation of the nuclear factor-kappaB pathway reduced and its inhibitor gene, NFKBIA, significantly up-regulated.	40
Fungal biomarker discovery by integration of classifiers.	The study utilized Mixed Integer Linear Programming (MILP) classifiers to generate a gene signature for distinguishing fungal and bacterial infected samples. Combining classifiers increased the consistency of the biomarker gene list, with a 43% increase in pairwise overlap.	The refined gene list ranked 19 genes based on consistency in expression, most linked to the ERK-MAPK signaling pathway. The method achieved an average accuracy of 83% on unseen datasets.	41

A six gene support vector machine classifier contributes to the diagnosis of pediatric septic shock.	A study using four microarray datasets (GSE26378, GSE26440, GSE13904, and GSE4607) from the Gene Expression Omnibus database explored the mechanisms of pediatric septic shock (PSS). The MetaDE package screened consistently differentially expressed genes (DEGs) in the datasets, while the WGCNA package identified disease-associated modules and genes. The caret package selected optimal feature genes, and a support vector machine (SVM) classifier was built using the e1071 package.	The study found 2,699 consistent DEGs across the four datasets, and four stable modules were enriched with consistent DEGs. These modules selected six optimal feature genes, and an effective SVM classifier was constructed based on the six optimal genes. This classifier can potentially improve early PSS diagnosis accuracy and suggest molecular intervention targets.	42
Patient-specific early classification of multivariate observations.	The Early Classification Model (ECM) is a novel approach for early, accurate, and patient-specific classification of multivariate observations. It combines the widely used Hidden Markov Model (HMM) and Support Vector Machine (SVM) models. ECM has shown promising results in datasets, outperforming baseline models that required full-time series classification. In experiments involving Multiple Sclerosis patients, ECM used only an average of 40% of a time series, outperforming some baseline models.	In sepsis therapy datasets, ECM outperformed standard threshold-based methods and state-of-the-art methods for early multivariate time series classification.	43
A generalizable 29-mRNA neural-network classifier for acute bacterial and viral infections.	A generalizable 29-mRNA neural network classifier has been developed for acute bacterial and viral infections. The classifier uses training data from 18 retrospective	The IMX-BVN-1 AUROCs are 0.86 for bacterial infections and 0.85 for viral infections. In patients enrolled within 36 hours of hospital admission, the IMX-BVN-1 AUROCs are 0.92 for bacterial infections and 0.91	44

	transcriptomic studies and has a bacterial-vs-other AUROC of 0.92 and a viral-vs-other AUROC of 0.85. The classifier, inflammatix-bacterial-viral-noninfected-version 1 (IMX-BVN-1), is applied to an independent cohort of 163 patients.	for viral infections. With further study, IMX-BVN-1 could provide a tool for assessing patients with suspected infection and sepsis at hospital admission.	
Personalized identification of differentially expressed pathways in pediatric sepsis.	Identifying core pathways in an individual is crucial for understanding septic mechanisms and applying custom therapeutic decisions.	A study using samples from a control group and a pediatric sepsis group identified 277 enriched pathways as attractors, with 81 pathways with $P<0.05$ and 59 with $P<0.01$. The individualized pathway aberrance score (iPAS) was calculated to distinguish differences. Cluster analysis of pediatric sepsis using the iPAS method identified seven pathway clusters and four sample clusters, indicating that core pathways can be detected in most pediatric sepsis samples. This novel procedure identifies dysregulated attractors in individuals with pediatric sepsis, potentially improving the personalized interpretation of disease mechanisms and potentially useful in the era of personalized medicine.	45
Prediction of feature genes in trauma patients with the TNF rs1800629 A allele using support vector machine.	This study predicted feature genes in trauma patients with the TNF rs1800629 A allele using a support vector machine (SVM) classifier. The study used 58 gene expression data sets from Gene Expression Omnibus to predict the TNF rs1800629 A allele in trauma patients. The SVM classifier model was applied, combined with the leave-one-out	Functional annotation revealed that HMOX1 and RPS7 were mainly enriched regarding cell proliferation and the ribosome. HMOX1 and RPS7 may be key feature genes associated with the TNF rs1800629 A allele, playing a crucial role in the inflammatory response in trauma patients. The cell proliferation and ribosome pathways may contribute to the progression of severe trauma.	46

	cross-validation method. Functional annotation of feature genes was carried out to study their biological function. A total of 133 feature genes were screened out, and the SVM classifier peaked in predictive accuracy with a 100% correct rate in the training set and 86.2% in the testing set.		
The study investigates the mechanisms of sepsis, a systemic inflammatory response syndrome induced by infection in the lungs, abdomen, and urinary tract.	The expression profiles of E-MTAB-4421 and E-MTAB-4451 leukocytes were downloaded. Differentially expressed genes (DEGs) were identified and performed with hierarchical clustering analysis. A protein-protein interaction (PPI) network was constructed using the BioGRID database and Cytoscape software. A total of 384 DEGs were screened in the survival group. The PPI network was divided into four modules, involving 11 DEGs, including microtubule-associated protein 1 light chain 3 alpha (MAP1LC3A), protein kinase C-alpha (PRKCA), metastasis associated 1 family member 3 (MTA3), and scribbled planar cell polarity protein (SCRIB). Functional enrichment demonstrated that MAP1LC3A in module D was enriched in autophagy vacuole assembly.	The SVM classifier correctly identified the samples in E-MTAB-4451. In conclusion, DEGs such as MAP1LC3A, PRKCA, MTA3, and SCRIB may be implicated in sepsis progression and require further confirmation.	47

TABLE 1B. Research studies using a Supervised ML approach are shown. Supervised techniques are useful for predictive purposes when distinct groups are known and thereby already classified.

*REF is the Reference citation in the literature

Table 2. Artificial Neural Network Applied to Sepsis

STUDY	DESCRIPTION	RESULTS	REF*
Using machine learning algorithms, developing an autophagy-related gene classifier for early diagnosis, prognosis, and prediction of immune microenvironment features in sepsis.	The study focuses on the model developed using a systematic search in ArrayExpress and Gene Expression Omnibus cohorts from 2005 to May 2022. The ARG classifier was analyzed using multi-transcriptome data and correlated with immunological characteristics, including immune cell infiltration, immune and molecular pathways, cytokine levels, and immune-related genes.	The model exhibited excellent diagnostic values (AUC > 0.85) and superior differentiation of sepsis from other critical illnesses. The identified hub ARGs were significantly associated with immune cell infiltration, immune and molecular pathways, and cytokine levels. The ARG classifier exhibited superior diagnostic performance compared to procalcitonin and C-reactive protein in patients with sepsis. The ARG classifier can assist clinicians in diagnosing sepsis and identifying high-risk patients, guiding personalized treatment, and facilitating personalized counseling for specific therapy.	48
Studying a bioinformatical framework for the identification and validation of biomarkers in SIRS, sepsis, and septic shock patients	A methodologic framework for identifying and validating gene biomarkers in sepsis, sepsis, and septic shock patients was described, using a 2-tier gene screening and ANN data mining technique.	Eight key hub markers were identified, which could delineate distinct core disease processes and inform underlying immunological and pathological processes. These markers do not show enough fold change differences between different disease states to be useful as primary diagnostic biomarkers but were instrumental in identifying candidate pathways and other associated biomarkers for further exploration.	49

Table 2. Above are examples of Artificial Neural Networks (ANNs) applied to Sepsis. ANNs are machine learning algorithms which use interconnected nodes or neurons in a layered structure that resembles the human brain.
*REF is the Reference citation in the literature

Table 3. Ensemble Learning Technique in Sepsis Research

Ensemble Algorithm Types	Study Details	REF
WGCNA and Gene Ensemble Noise Reduction	Training a gradient-boosted regression tree model to classify Covid-19 severity. They argued that genes do not function in isolation but act as ensembles representing biological pathways and protein complex subunits. Further, they suggested that an imbalance in the expression of the gene ensemble results in disease pathology. Then they inferred that variance in gene expression of the ensemble or 'gene ensemble noise' is related to gene alteration. The model accurately predicted patients with mild and severe Covid-19. Using gene ensemble noise versus WGCNA demonstrated equal accuracy.	50
Four ML algorithms (random forest, recursive feature elimination using support vector classifier, logistic regression with lasso, and Boruta	A machine learning ensemble approach was used to analyze the gene expression data identifying 239 genes in urine, which effectively classified septic patients from those with other chronic conditions.	51

Table 3. The use of an Ensemble Machine Learning approach applied to sepsis is shown. The ensemble machine learning methods combine the insights from multiple learning models to improve the accuracy of decisions.
*REF is the Reference citation in the literature