

## Original Research Article

# A hybrid graph-based framework for sepsis prediction using patient similarity network

Sabari Giri Murugan S<sup>1\*</sup>, Sudhakaran R<sup>2</sup>

<sup>1</sup>SNS College of Engineering, <sup>2</sup>Department of Mechanical Engineering, SNS College of Engineering, Coimbatore, India

\*For correspondence: **Email:** [sabarimtech08@gmail.com](mailto:sabarimtech08@gmail.com)

Sent for review: 3 September 2025

Revised accepted: 11 April 2026

## Abstract

**Purpose:** A scalable patient similarity network was developed based on the Medical Information Mart for Intensive Care.

**Methods:** The unit consists of 40,000 patients as network nodes and 100 million weighted edges, where a connection between a patient and a doctor represents diagnoses, laboratory measurements, vital signs, and medication history. The propagation of similarity was performed using the Random Walk with Restart algorithm with a restart probability of 0.35, utilizing a hybrid system of central processing unit clusters and graphics processing unit acceleration, which achieved a speedup of 8.2 times. The network also exhibited small-world properties, with a diameter of 7 and a modularity of 0.42.

**Results:** Predictive analysis showed that the Random Walk with Restart algorithm and a hybrid Random Walk with Restart and Graph Neural Network ensemble had areas under the receiver operating characteristic curves of 0.91 and 0.94, respectively, which are better than those of the conventional models.

**Conclusion:** This has been made possible by incremental updates that enable near-real-time incorporation of new patients, providing an efficient, scalable, and clinically interpretable framework for sepsis risk prediction.

**Keywords:** Patient Similarity Networks, Hybrid graph, Heterogeneous data, Random walk with Restart

This is an Open Access article that uses a funding model which does not charge readers or their institutions for access and distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>) and the Budapest Open Access Initiative (<http://www.budapestopenaccessinitiative.org/read>), which permit unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

Tropical Journal of Pharmaceutical Research is indexed by Scopus, Chemical Abstracts, Embase, Index Copernicus, EBSCO, African Index Medicus, JournalSeek, Directory of Open Access Journals (DOAJ), African Journal Online, Bioline International, Open-J-Gate and Pharmacy Abstracts

## INTRODUCTION

Healthcare systems have digitized at an unprecedented pace, generating large volumes of patient data in the form of electronic health records (EHRs), laboratory results, imaging modalities, and monitoring devices in intensive care units (ICUs). The segmentation of this data poses a twofold problem because it is both high-dimensional and heterogeneous, comprising categorical diagnosis codes, continuous laboratory measurements, time-series measurements of vital signs, and unstructured clinical notes. Traditional machine learning

approaches may solve individual problems but ignore inter-patient relationships, which are central to precision medicine.

Similarity networks of patients connect patients as nodes and their similarities as weighted edges, allowing similar patients to be effectively clustered for disease progression and risk prediction. The objective of the study is to identify the onset of sepsis within 6 hours of prediction based on the multimodal data of intensive care unit (ICU) patients in the MIMIC-III database. In order to overcome these shortcomings of the traditional tabular models, a hybrid graph

modeling, in which inter-patient relations are represented by a Patient Similarity Network (PSN) has been suggested.

The principle findings of this study included creation of a large-scale multimodal PSN that combines diagnoses, laboratory measurements, vital signs, and medication records; the use of the Random Walk with Restart (RWR) algorithm to propagate similarities and risk diffusion; the creation of a hybrid RWR and Graph Neural Network (GNN) ensemble model to achieve a better predictive capability; and the introduction of a scalable cloud-native architecture that is capable of providing efficient computation and incremental updates to near-real-time analytics.

## METHODS

This paper develops the sepsis prediction as an early risk prediction task with multimodal patient data in the ICU. Considering past clinical data such as diagnoses, lab measurements, vital signs and prescription history, the aim is to predict the likelihood of a patient to develop sepsis within a pre-established prediction time frame of 6 hours. As an input to the prediction, only the available data at the time of prediction are taken on a patient-by-patient basis to make the clinical environment realistic. The model outputs a probability score at the end of a time period that reflects the risk of sepsis occurring at the end of the time period. A cloud-native patient similarity network architecture was created based on a large critical care dataset comprising roughly 40,000 intensive care unit patients, in which network connections were defined by clinical similarity scores, and similarity propagation was performed using a random-walk random-restart strategy to predict sepsis risk. The framework demonstrates scalability in network analysis, faster computation time through parallel processing, and enhanced predictive accuracy compared to traditional clinical frameworks, favoring patient-centric, real-time analytics in critical care units.

## Related works

Healthcare research has seen rapid growth in predictive modeling based on big data from electronic health records, driven by improvements in data accessibility, machine learning algorithms, and computing resources [1,2]. The literature related to the proposed study covers three key areas: patient similarity networks, graph-based algorithms in healthcare, and scalable cloud architectures for faster processing of health data [3-5].

## Similarity networks of patients

The patient similarity networks represent patient groups as graphs where patients are the nodes, and the relationship between the similarity of two patients forms the edges [1,6]. Such representation supports the fusion of heterogeneous clinical data, preserving interpretability and enabling cohort-level analysis [7,8]. Initial studies in this field aimed to define a reliable similarity measure for individual data modalities, such as clinical characteristics, genomic databases, and medical images [3,9]. An outstanding method for combining several data views into a single patient network is called Similarity Network Fusion, and it is effective at finding clinical subtypes, especially in oncology [3,10]. Disease comorbidity analysis and clinical risk prediction have been used in other studies to construct phenotype similarity networks using electronic health record data [5,11,12]. Altogether, the patient similarity networks offer a robust framework for processing high-dimensional, heterogeneous healthcare data by mapping each modality to a similarity-based network representation [1,8].

Although these exist, most current studies on patient similarity networks rely on sparse feature sets or one-dimensional data and are typically tested on rather small patient groups [6,11]. The overall integration of diagnoses, laboratory measurements, vital signs monitoring, and medication orders into a single large-scale net has not yet been fully studied, especially in the context of longitudinal intensive care unit data [12,13].

## Graph algorithms in medical research

Graph-based algorithms are highly applicable to healthcare analytics due to the relational complexity of patients' similarity networks [10,14]. Random walk procedures have been widely used in bioinformatics, including ranking, association discovery, and link prediction [4,6]. Random walk with restart algorithms spread information across a network with local and global connectivity, enabling the identification of nodes closely related to a target cohort [7]. Gene prioritization and drug repurposing have been successfully applied using these techniques, reinforcing the idea that they reveal dormant biological connections [4,8].

Graph-based machine learning algorithms, such as graph neural networks, have also been used more recently in clinical environments to learn inter-patient interactions and predictive accuracy on clinical tasks such as sepsis prediction

[2,9,11]. Nevertheless, diffusion techniques based on random walks are not widely used when large-scale electronic health records are used to support clinical predictions, and most studies still rely on conventional classifiers or sequence-based deep learning models [7,10,15]. Also, the use of graph neural networks in healthcare has mostly been in small datasets or disease cohort elite cases [9,11,14].

### **Large-Scale Architectures and Rapid Health Data Processing.**

Patient similarity networks of scale require scalable storage and distributed computing, as well as high-performance computing infrastructure, to build [12,13]. Cloud computing solutions offer automatic scaling, distributed data storage, and embedded analytics, facilitating end-to-end pipelines for healthcare data [12,16,18]. Hybrid cloud architecture has become a viable paradigm for operating sensitive medical data and, at the same time, facilitating large-scale computation [12,16].

New trends in graphics processing unit (GPU)-accelerated graph analytics have achieved significant efficiency gains in iterative graph algorithms [14,17]. Parallel graph processing packages have achieved significant runtime improvements over more traditional central processing unit-based implementations [17]. Hybrid central processing unit-graphics processing unit architectures are also becoming more popular, which allows efficient preprocessing of data on the central processing unit and offloads computationally intensive graph operations to graphics processing units [14,15]. Although this has been developed, only a little work has systematically integrated cloud scalability with graph analytics-accelerated similarity networks for large patient volumes in critical care units [16,18].

### **Sepsis prediction and research gap application**

Sepsis is still considered one of the most problematic conditions that is difficult to predict and treat early in critical care facilities [19]. Many studies based on electronic health record data have used time-series models, both traditional machine learning and deep learning, to predict sepsis [20]. Although these strategies achieve encouraging performance, they tend to treat patients separately and do not allow for the observation of complex inter-patient interactions [9,11].

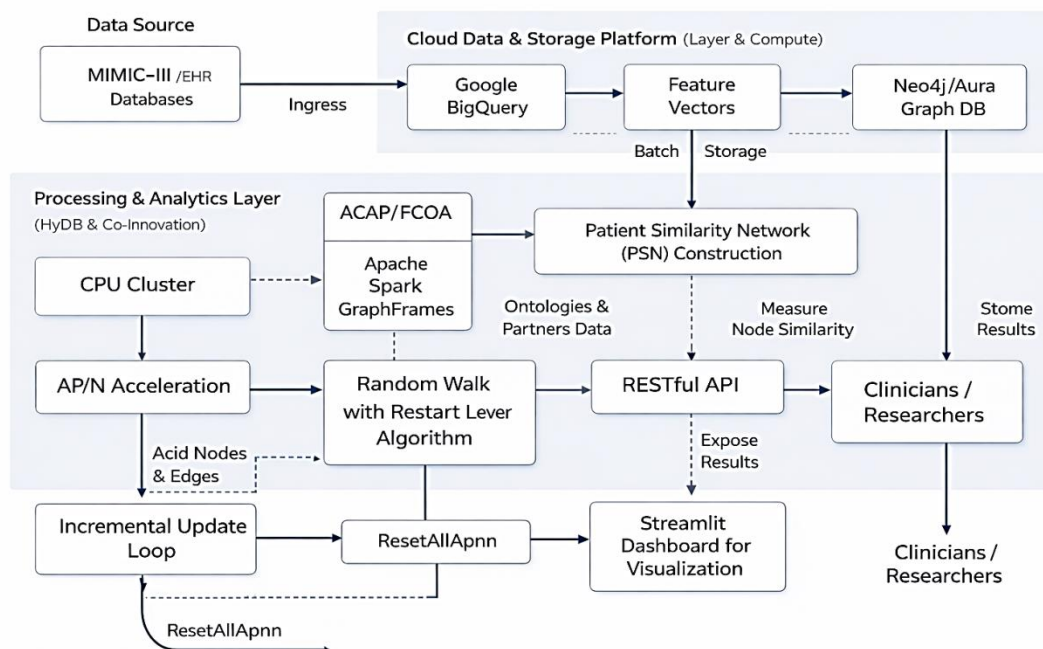
To overcome this limitation, recent studies have used graph-based methods, such as knowledge graphs and graph neural networks, to enhance prediction accuracy [9,14]. However, a large-scale multipatient similarity network with network diffusion as a sepsis risk stratification method has not yet been rigorously examined [2,10]. The current research seldom integrates different clinical modalities into a single network and discusses the computational scalability required to implement these modalities at an intensive care unit-wide level [13,16].

To conclude, despite the achieved major advancements in electronic health record-based predictive modeling [1,19], there are still gaps in multi-modal data integration [6,12], network diffusion methods applied to clinical risk prediction [7,8], and scalable implementation of the graph-based analytics with the help of cloud and accelerators of graphics processing units [14,16,18]. It is important to fill these gaps to facilitate real-time and patient-centric prediction in critical care settings [2,20].

## **METHODS**

Raw intensive care unit data from the Medical Information Mart for Intensive Care database are ingested into a cloud-based data warehouse to extract patient features, including diagnoses, laboratory measurements, vital signs, and medication records. These features are stored in columnar format for batch processing and in a graph database for interactive queries, and patients are linked within a patient similarity network using multi-modal similarity measures. Distributed graph processing is performed on CPU clusters, while GPU acceleration is used to expedite the random walk with restart diffusion and generate converged similarity scores. The resulting outputs are made available via an application programming interface and visualized in an interactive dashboard, with incremental updates that allow new patient data to be incorporated without requiring a full network recomputation.

Table 1 presents the entire algorithm for constructing the patient similarity network and applying a random walk with restart to rank patients on the Medical Information Mart in Intensive Care dataset. The dataset is initially subjected to several preprocessing steps, such as normalization, feature reduction, and imputation, to align diagnoses, laboratory measurements, vital signs, and medication records in a single feature space.



**Figure 1:** System Architecture

**Table 1:** Pseudocode for Patient Similarity Network with Random Walk (PSN-RWR)

Input: MIMIC-III dataset  $D$  {diagnoses, labs, vitals, meds}, seed set  $S$   
Output:  $p$ : Ranking vector  $p \in \mathbb{R}^N$  (patient similarity scores)

# Preprocessing:

- 1: Patient features extraction  
Mapping CCS categories to ICD-9 codes  
Z-score normalize lab values, log transform skewed  
Summarize vitals into 24h aggregates {mean, min, max}
- 2: Encode medication into therapeutic class indicators
- 3: Median imputation for missing values (within diagnosis group)
- 4: Reduce dimensionality of labs/vitals using PCA with 95% variance
- 5: Save feature matrix  $F \in \mathbb{R}^{(N \times d)}$

# construction of patient similarity networks (PSNs)

- 6: For every pair of patients  $(i, j)$ :  
 $\text{sim\_diag} = \text{Jaccard}(\text{ICD9}_i, \text{ICD9}_j)$   
 $\text{sim\_lab} \leftarrow \cos(F_{\text{lab}_i}, F_{\text{lab}_j})$   
 $\text{sim\_temp} \leftarrow \text{overlap}(\text{ICU}_i, \text{ICU}_j) / \text{max\_overlap16}$   
 $w_{ij} = 0.5\text{sim\_diag} + 0.3\text{sim\_lab} + 0.2\text{sim\_temp}$
- 7: For each patient, keep the top-k 100 neighbors  
Build sparse adjacency matrix  $W \in \mathbb{R}^{(N \times N)}$ , col- {0,1}
- 8: Random walk with restart (RWR)  
Set  $p_0$  uniform over seed set  $S$ ;  $p \leftarrow p_0$
- 9: repeat
- 10:  $p_{\text{new}} = (1 - r) \cdot W \cdot p + r \cdot p_0$
- 11: if  $\|p_{\text{new}} - p\| < \epsilon$  then; break  
end if
- 12:  $p \leftarrow p_{\text{new}}$
- 13: until convergence
- 14:  $p$  (relevance score stationary patient) return

# Cloud Implementation:

- 15: Store raw tables into Google BigQuery
- 16: Store similar edges in parquet (batch) and neo4j (interactive)
- 17: Distributed PSN construction with Spark GraphFrames
- 18: Iterate on Matrix Multiplications with RAPIDS cuGraph on GPU

# Upon admission of new patient:

- 19: Update node and get the edges contained locally
- 20: Incremental update adjacency block and RWR scores
- 21: Expose results via rest API + streamlit dashboard

This is followed by the calculation of patient-to-patient similarities across three modalities: diagnosis similarity using the Jaccard coefficient, laboratory profile similarity using the cosine similarity, and temporal overlap of intensive care unit stays. All these measures of similarity are combined to create a sparse k-nearest neighbor graph. The random walk with restart process is based on a column-normalized adjacency matrix and starts with a probability vector over a defined set of clinically relevant initial seeds, with the final result (a ranked list of patients) determined by their similarity to the initial seeds. The framework is deployed on a cloud-native architecture that combines graph data warehousing, distributed, GPU-based graph computation, and a graph database with an application programming interface and visualization dashboards. This design facilitates scalable, efficient analysis of patient similarity based on clinical interpretations.

### Dataset: MIMIC-III

The database employed by the research is MIMIC-III (Medical Information Mart for Intensive Care, version 1.4), which provides de-identified health records of more than 58,000 ICU admissions of about 40,000 patients. In this study, a sample size of 40,000 patients was used due to the availability of full longitudinal clinical data. Inclusion criteria included patients who had an ICU stay lasting more than 24 hours and who had significant modalities such as diagnoses, laboratory measurements, vital signs, and medication records. Data quality and consistency were achieved by dropping patients whose data were excessively missing or incomplete.

In order to prevent repetitive admissions, only one ICU stay was taken into account. Clinical features were derived and coded into four modalities, namely diagnosis codes (categorized into CCS, such as mean, minimum, and maximum over 24 hours), laboratory measurements (normalized continuous variable), vital signs (aggregated into statistical summaries, e.g., mean, minimum, and maximum over 24 hours), and medication records (encoded into therapeutic class indicators). Median imputation was used to address missing values when there was a similarity in clinics. Everything was standardized and the dimensionality reduction was adopted on the basis of principal component analysis (PCA), wherein 95% of the variance was retained. The sepsis labels were also determined based on the Sepsis-3 clinical criteria, which define sepsis as endangering organ dysfunction as a result of dysregulated host response to a pathogen. Namely, the appearance of sepsis

was determined as suspected infection and a rise of the Sequential Organ Failure Assessment (SOFA) score by  $\geq 2$  points. Cases where the patient met the established MIMIC-III benchmarking protocols were defined as the time of sepsis onset, with the earliest time meeting the protocols taken as the onset time. Labeling of every patient was done in a time-conscious way in accordance with the task of prediction, so that only the available information before the onset (or prediction window) was modeled.

### Data preprocessing and feature engineering

Before the construction of the models, the raw clinical data had been preprocessed to be consistent and strong. Clinical Classifications Software (CCS) categories were used to map diagnosis codes (ICD-9); thus, dimensionality was reduced and interpretability increased. The laboratory measurements were handled as continuous variables and the normalization of skewed distributions was done using the z-score normalization and log-transformed. Statistical summaries of vital signs were used to combine vital signs into fixed 24-hour windows with a mean, minimum, and maximum value. The records of medication were coded as binary indicators in terms of classes of therapeutic records.

In clinical groups of patients with similarities, median imputation was used to treat missing values to maintain distributional properties. Outliers on the laboratory values were discussed bmts.vt winsorization. All the features were then scaled to a normalized range and Principal Component Analysis (PCA) was applied to reduce dimensionality, keeping 95% of the total variance. The feature matrix it produced represented similarity computation in the Patient Similarity Network in a stable and compact form.

### Patient similarity network construction

Patient Similarity Network (PSN) is built as a weighted graph  $G = (V, E)$ , where each node  $v \in V$  denotes a patient and every edge,  $e_{ij} \in E$  is

the similarity of patients  $i$  and  $j$ . In order to make the computations scaled to  $k$ , a k-nearest neighbor (kNN) method was used. It has used  $k=100$  such that each patient is related only to the top- $k$  nearest patients of it. The patient similarity was calculated by combining three mutually complementary modalities: similarity in diagnoses, similarity in laboratory features, and co-occurring ICU stay. Similarity of diagnoses was calculated with the help of the Jaccard coefficient between sets of ICD-9 codes. Cosine

similarity was used to calculate the laboratory similarity between the normalized feature vectors. The definition of temporal similarity was given as the proportion of overlapping length of stay in the ICU to the length of stay that could have been overlapped. The last weight obtained in the convex combination of the components of the end between two patients was:

$$w_{ij} = 0.5 \cdot \text{sim}_{\text{diag}} + 0.3 \cdot \text{sim}_{\text{lab}} + 0.2 \cdot \text{sim}_{\text{temp}} \quad (1)$$

The resulting graph is first directed by the construction of the kNN. To enhance the stability and connection, symmetrization was used by keeping the maximum weight between  $w_{ij}$  and

$w_{ji}$  resulting in an undirected weighted graph.

Following a top- k neighbor per node pruning stage and symmetrization, the resulting PSN has around 40,000 nodes and about 100 million weighted edges. This sparse graph allows storage and scalable computation to be efficient and the clinically meaningful relationships are maintained.

### RWR (Random Walk with Restart)

Random Walk with Restart (RWR) was used to spread similarity through the network and to rank patients in relation to a seed cohort. Random Walk with Restart be thought of as a diffusion process where a random walker is traversing through the network, while it has a probability to go back to the seed nodes.

More formally, let  $W$  be the column-normalized adjacency matrix of the PSN,

$$W_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} \quad \dots\dots\dots (2)$$

The probability distribution  $W_{ij}$  of the nodes at

time step  $t$  was represented  $p_t$ , and the starting

probability vector was concentrated on a group of patients with sepsis who were considered to be seeds. The probability distribution was given by specifying the Random Walk with Restart iteration.

$$p_{t+1} = (1-r)w_{pt} + rp_0 \quad (3)$$

Notice that  $r \in (0,1)$  as a hyperparameter denoting the fraction of the sentence that will come from the prior ( $p_0$ ) and the remaining

fraction will be from  $w_{pt}$ , where  $W$  is the word embedding matrix, since  $w_{pt}$  is the word

embedding representation of the words in  $p_t$ . In

this study,  $r = 0.35$  is used consistently across all experiments.

where  $r$  is the chance of a restart and it regulates the trade-off between the global exploration and local exploitation. In this work choose  $r = 0.3$ , which indicates that in each step, the walker had a 35 % probability of returning home (the seed set) and a 65 % probability of continuing diffusion. Until convergence results were obtained, further information was provided with the stopping criterion, and the algorithm repeated these iterations (Eq 4).

$$\|p_{t+1} - p_t\| < 10^{-5} \quad \dots\dots\dots (4)$$

Here  $\|\cdot\|$  denotes the  $L_1$  norm. The output stationary distribution  $p^*$  puts more probability mass on patients who are closely connected to the seed set in the network.

### Cloud implementation

The dataset size was imposed, the patient similarity network was provided, and all processing was implemented in a cloud-native setup. The raw MIMIC-III data were converted into patient feature vectors via ordered SQL queries at Google BigQuery. Similarity edges were created and saved as columnar Apache Parquet files for batch processing and as a Neo4j graph database for querying. Homogeneous similarity score parallel propagation was executed on Apache Spark GraphFrames on CPU clusters and on RAPIDS Graph on cloud GPUs, using an iterative matrix product required by Random Walk with Restart. The ability to add new patients incrementally without recomputing the entire network was made possible. The resulting rankings of patient similarity were made available through a Web API, visualized in a dashboard, and allowed the exploration of clusters of patients and risk factors associated with clinical conditions (sepsis).

In order to have a realistic and unbiased assessment, strict measures were put to ensure

that no data would leak out. As input features, only clinical data that patients had before the prediction time (before the 6-hour prediction window) were considered. All the information that was recorded later than the prediction window or later than the occurrence of sepsis became part of the model.

Besides this, data were categorized into training, validation and test sets at the patient level to prevent overlap of samples from the same individual. Clinical event temporal ordering was maintained during the preprocessing and feature extraction. Such precautions will make sure that the model will not unintentionally take advantage of future information, hence giving a stable evaluation of predictive performance.

### Setup

In order to achieve a healthy and balanced assessment, the data was split into training (70), validation (15), and test (15), concentrating on the person-level in order to make sure that no individual patient would be represented in more than one split. The validation set was used in tuning hyperparameters of the models, the parameters of the restart probability and the parameters of graph construction. All experiments were repeated five times with various random splits in order to consider variability and enhance reliability and average performance and standard deviation were reported. The weighted loss functions were utilized in the training of the models to overcome class imbalance in sepsis prediction. The standard measures that were used to determine model performance were Area Under the Receiver Operating Characteristic Curve (AUROC), Area Under the Precision-Recall Curve (AUPR), and F1-score.

### Hardware configuration

Google Cloud Platform as a computational infrastructure system had 32-core CPU cluster (256 GB RAM) to run Apache Spark (distributed graph processing) and 2 NVIDIA V100 GPUs (16 GB memory) in a cloud instance with 192 GB RAM to support efficient combination of distributed and CPU-accelerated processing.

### Parameter settings

For the RWR restart probability is  $r=0.3$  [15] and the medical similarity networks used a convergence threshold of  $\|p_{i+1} - p_i\| < 10^{-5}$ , and

achieved convergence in a matter of iterations on average in 28. The same set of patients was

used to evaluate all the baseline models (Logistic Regression, Random Forest and Graph Neural Networks) to keep the evaluation fair.

## RESULTS

### Network statistics

A k nearest neighbor algorithm ( $k = 100$ ) produced a Patient Similarity Network of 40000 patients and 100 million edges, with the average degree of 2500 and the graph density of 0.0063 (Table 2). The network had a small-world effect with a diameter of about 7 and a modularity of 0.42, which demonstrated an explicit community structure and a clinically significant group, such as sepsis and cardiovascular disease.

**Table 2:** Network analysis

| Metric                       | Value                          |
|------------------------------|--------------------------------|
| Number of Patients (Nodes)   | 40,000                         |
| Number of Edges              | 100 million (sparse, $k=100$ ) |
| Average Degree               | 2,500                          |
| Graph Density                | 0.0063                         |
| Diameter (approx.)           | 7                              |
| Modularity (Community Score) | 0.42                           |

### Random walk with restart (RWR)

The random walk with restart algorithm effectively diffused through the large Patient Similarity Network with a restart probability of 0.35 and converged in 28 +/- 2 iterations, as depicted in Table 3. Execution on Apache Spark using CPU took about 3 hours, but using the RAPIDS cuGraph with GPU acceleration took only 22 minutes, resulting in an 8.2-fold speedup and enabling real-time application to different seed cohorts.

**Table 3:** Performance of random walk with restart

| Parameter                 | Value        |
|---------------------------|--------------|
| Restart Probability (r)   | 0.35         |
| Iterations to Convergence | $28 \pm 2$   |
| Runtime (CPU, Spark)      | ~ 3 hours    |
| Runtime (GPU, RAPIDS)     | ~ 22 minutes |
| Speedup (GPU vs CPU)      | 8.2x         |

Scalability of the suggested framework is tested with a runtime analysis of the CPU and GPU systems. RWR algorithm took about 3 hours on a distributed CPU-based Spark platform, but when accelerated with a GPU (rapid cuGraph), it took only about 22 minutes and was 8.2x faster. This significant decrease demonstrates the efficiency of parallel processing of graphs with large-scale patient similarity networks. Also, incremental

updates are used to revise the entire network, which incorporates new patient data without needing to recalculate the entire network, and provides near-real-time updates. These findings suggest that the selected system may be scaled to computers and used with large clinical data, but needs a more complex optimization to be deployed in real-time and strict mode in operational environments.

### Predictive evaluation measures

Table 2 summarizes structural characteristics of the Patient Similarity Network (PSN) built on 40,000 patients and about 100M edges. The value of the mean degree (2,500) and low density (0.0063) of the graph show that the network is sparse but well-connected. The small diameter ( $\approx 7$ ) makes it a small-world, which makes it efficient in information propagation. Secondly, a modularity score of 0.42 also indicates that there are meaningful clinical clusters, e.g., sepsis and cardiovascular disease clusters. The results stated here are the average of five separate runs and the standard deviation as a measure of variability. The final RWR + GNN ensemble has shown the best performance in all the measures with a low variance, which implies consistency in predictions and reliability.

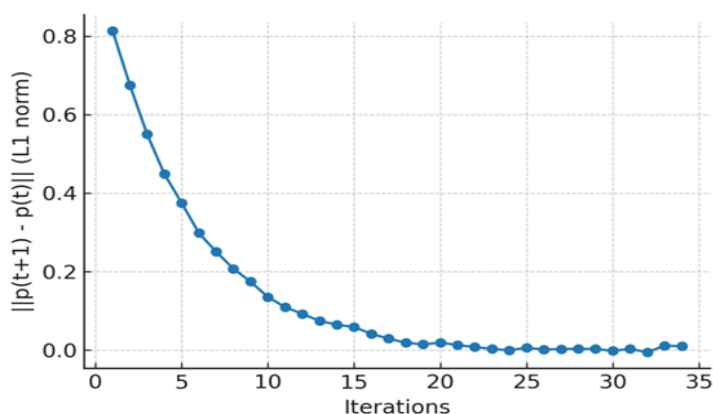
### Visual analysis

RWR rapidly converges in 25 -30 iterations, and differences in L1 norms exponentially decrease

**Table 4:** Performance analysis

| Model               | AUROC           | AUPR            | F1-score        |
|---------------------|-----------------|-----------------|-----------------|
| Logistic Regression | 0.72 $\pm$ 0.01 | 0.41 $\pm$ 0.02 | 0.52 $\pm$ 0.01 |
| Random Forest       | 0.78 $\pm$ 0.02 | 0.49 $\pm$ 0.02 | 0.58 $\pm$ 0.02 |
| GNN (GraphSAGE)     | 0.86 $\pm$ 0.01 | 0.62 $\pm$ 0.02 | 0.70 $\pm$ 0.01 |
| RWR (PSN-based)     | 0.91 $\pm$ 0.01 | 0.68 $\pm$ 0.01 | 0.74 $\pm$ 0.01 |
| RWR + GNN Ensemble  | 0.94 $\pm$ 0.01 | 0.72 $\pm$ 0.01 | 0.78 $\pm$ 0.01 |

Values are (mean  $\pm$  standard deviation)



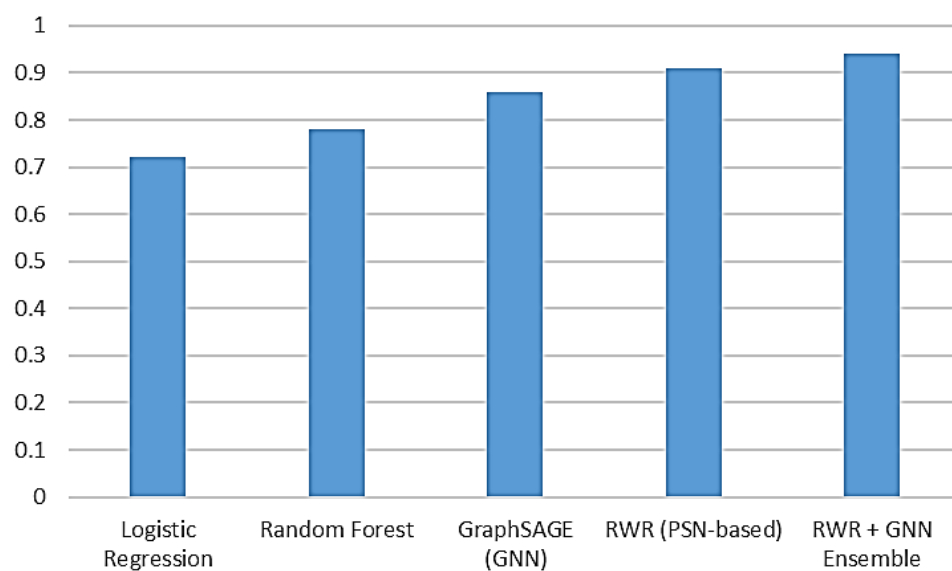
**Figure 2:** RWR Convergence

(Figure 2). The probability of restart between 0.35 exploration and exploitation enables stable diffusion on networks of 100 million edges and makes computations practicable on a CPU or GPU. The PSN demonstrated clinically meaningful clusters, including major ones (cardiovascular disease and sepsis) and minor ones (low-prevalence conditions), identified by Louvain clustering. Hub patients are placed in the middle, outliers at the edges, and a modularity of 0.42 indicates that the structure is well organized (Figure 3). The network-based approaches were superior to conventional models, as depicted in Figure 4. GraphSAGE (AUROC 0.86) and RWR alone (AUROC 0.91) performed better than the logistic regression (AUROC 0.72) and random forest (AUROC 0.78). The hybrid RWR + GNN ensemble also showed the best performance, with the highest AUROC of 0.94, higher F1 Scores, and higher AUPR, demonstrating that patient similarity diffusion, in combination with neural network learning, is valuable.

The ablation experiment assessed the value of each element in the proposed framework, as shown in Table 5. Patient Similarity Network (PSN) by itself gives a moderate performance, which implies that similarity-based representation is informative.



**Figure 3:** Community detection on network



**Figure 4:** AUROC comparison

**Table 5:** Ablation Study

| Model variant         | AUROC       | F1-score    |
|-----------------------|-------------|-------------|
| PSN (Similarity only) | 0.82        | 0.64        |
| RWR only              | 0.91        | 0.74        |
| GNN only (GraphSAGE)  | 0.86        | 0.70        |
| RWR + GNN (Proposed)  | <b>0.94</b> | <b>0.78</b> |

Random Walk with Restart (RWR) is also able to propagate similarity, providing a significant improvement to the performance. Graph Neural Networks (GNNs) also detect structural relationships but underperform slightly as compared to RWR. The hybrid RWR + GNN model is the most performing, which proves that similarity propagation based on diffusion and learning the graph representation can be used in a complementary way. The statistical significance of the performance of the proposed RWR + GNN ensemble in comparison to the baseline models is confirmed by paired statistical tests in several runs ( $p < 0.05$ ).

## DISCUSSION

The clinical similarity network of patients at scale revealed clinically significant relationships: a network of 40,000 patients, 100 million edges, with an average node degree of 2,500, a graph density of 0.0063, a diameter of 7, and a modularity score of 0.42, indicating disease-specific communities, e.g., sepsis and cardiovascular diseases. Random Walk with Restart (restart probability  $r = 0.35$ ) reached convergence in  $28 \pm 2$  iterations, with an acceleration factor of 8.2 on a GPU, reducing the runtime of this algorithm (an approximate 3 hours on CPU) to 22 minutes. Predictive performance indicated that, in the traditional models (logistic regression and random forest), the AUROC was 0.72 and 0.78, respectively, but in the graph-based models, the AUROC was higher (GraphSAGE AUROC = 0.86 and RWR AUROC = 0.91). The hybrid RWR + GNN ensemble obtained the best results on both measures of association, namely, the highest AUROC = 0.94 and F1-score = 0.78, which proves that integration of diffusion-based patient similarity with the graph representation learning can be used to predict sepsis risks better and to provide scalable, patient-centered, clinical decision support. Although the proposed framework has a high predictive power, other explanations are to be put into consideration. The improvement observed is partly due to the structure of the MIMIC-III data, the selection criteria of the cohort, or inherent data correlates rather than necessarily to the success of the suggested method. Besides, the multimodal nature and the graph-based representation result in data-specific biases. Thus, the external validation on other datasets and in other clinical environments is required to establish the overall applicability of the findings.

### Limitations of the study

This study is limited in a number of ways. First, it

is an analysis of retrospective data contained in the MIMIC-III database that could reduce the ability to generalize to different populations and clinical settings. Second, though several structured clinical modalities were incorporated, the unstructured data of clinical notes and medical images were not taken into account. Third, the definition of sepsis is based on labels derived from Sepsis-3 criteria, which creates a source of uncertainty because of possible differences in clinical interpretation and onset. Fourth, the study comprises patients with relatively complete records, which creates selection bias and restricts such use to the real world, where data are missing. Fifth, even with the precautions in place, there is still a possibility of leakage of the data in retrospective observational studies. Also, calibration analysis is not included in the study to determine the reliability of probability. Lastly, the model has been tested on one dataset without external validation, and, thus, its generalization and strength in other healthcare facilities are yet to be determined.

## CONCLUSION

This paper introduces a graph-based sepsis prediction framework that is scalable and that is based on a Patient Similarity Network built on multimodal ICU data. The suggested method is a combination of Random Walk with Restart (RWR) and Graph Neural Networks (GNNs) that would help retrieve similarity diffusion and structural relationships among patients. In experiments on the MIMIC-III dataset, it has been shown that the hybrid RWR + GNN model consistently has better predictive performance than the traditional machine learning and graph-based models. Although the findings suggest that patient-centric risk modeling has potential, the analysis is conducted with the help of retrospective data and should be further tested on external data and in the prospective clinical environment. The further work will be aimed at the introduction of more modalities of data, the enhancement of the model calibration, and the analysis of the possibility of real-time deployment in clinical settings.

## DECLARATIONS

### Acknowledgement/Funding

None.

### Ethical approval

None provided.

## Use of Artificial intelligence/Large language models

All intellectual contributions, study analyses, and interpretations reported in the manuscript are the original work of the authors.

## Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Conflict of interest

No conflict of interest is associated with this work.

## Contribution of authors

We declare that this work was done by the author(s) named in this article, and all liabilities pertaining to claims relating to the content of this article will be borne by the authors.

## REFERENCES

1. Brown SA. Patient similarity: emerging concepts in systems and precision medicine. *Front Physiol* 2016; 7: 561.
2. Zhang S, Tong H, Xu J, Maciejewski R. Graph convolutional networks: A comprehensive review. *Computational Social Networks* 2019; 6(1): 1-23.
3. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods* 2014; 11(3): 333–337.
4. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Molecular Syst Biol* 2011; 7: 496–496.
5. Huang Y, Wang N, Liu H, Zhang H, Fei X, Wei L, Chen H. Study on patient similarity measurement based on electronic medical records. *Stud Health Technol Inform* 2019; 264: 1484-1485. doi: 10.3233/SHIT190496.
6. Tong H, Faloutsos C, Pan JY. Fast random walk with restart and its applications. In *Proceedings of the IEEE Int Conference Data Mining* 2006; 613–622.
7. Yin T, Chen S, Wu X, Tian W. GenePANDA - A novel network-based gene prioritizing tool for complex diseases. *Scientific Reports* 2017; 7(1): 43258.
8. Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems* 2021; 33(2): 494-514.
9. Boll HO, Amirahmadi A, Ghazani MM, de Moraes WO, de Freitas EP, Soliman A, et al. Graph neural networks for clinical risk prediction based on electronic health records: A survey. *J Biomed Informatics* 2024; 151: 104616.
10. Feng Z, Wang R, Wang T, Song M, Wu S, He S. A comprehensive survey of dynamic graph neural networks: Models, frameworks, benchmarks, experiments and challenges. *IEEE Transactions on Knowledge and Data Engineering* 2025 Oct 14.
11. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. *Nature Biomed Engine* 2022; 6(12): 1353-1369.
12. Pino C, Di Salvo R. A survey of cloud computing architecture and applications in health. In *the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*. Atlantis Press 2013; 1649-1653.
13. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Med* 2018; 1(1): 18.
14. Cui H, Lu J, Xu R, Wang S, Ma W, Yu Y, et al. A review on knowledge graphs for healthcare: Resources, applications, and promises. *J Biomed Informatics* 2025: 104861.
15. Al-Khafaji M, Ramaha NT. Hybrid deep learning architecture for scalable and high-quality image compression. *Scientific Reports* 2025; 15(1): 22926.
16. Hricik T, Bader D, Green O. Using RAPIDS AI to accelerate graph data science workflows. In *2020 IEEE High Performance Extreme Computing Conference (HPEC) 2020 Sep 22 (pp. 1-4)*.
17. Jagani N, Jagani P, Shah S. Big data in cloud computing: a literature review. *Int J Engine Applied Sci Technol* 2021; 5(11): 185-191.
18. 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* 2021; 8: 607952.
19. Liu S, Gong L, Hu W, Cai J, Yang Y, Chen S, et al. Machine learning for prediction of in-ICU mortality in sepsis: an observational study using the MIMIC IV database. *Arch Med Sci* 2026; 22(1): 270-281.
20. Fleuren LM, Klausch TL, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* 2020; 46(3): 383-400.