Tensor Learning of Pointwise Mutual Information from EHR Data for Early Prediction of Sepsis

Naimahmed Nesaragi[a], Shivnarayan Patidar[a,*], Vaneet Aggarwal[b]

[a]*Dept. of Electronics & Communication Engineering, National Institute of Technology, Goa*
[b]*School of Industrial Engineering, Purdue University*

Abstract

Early detection of sepsis can facilitate early clinical intervention with effective treatment and may reduce sepsis mortality rates. In view of this, machine learning-based automated diagnosis of sepsis using easily recordable physiological data can be more promising as compared to the gold standard rule-based clinical criteria in current practice. This study aims to develop such a machine learning framework that demonstrates the quantification of heterogeneity within the tabular electronic health records (EHR) data of clinical covariates to capture both linear relationships and nonlinear correlation for the early prediction of sepsis. Here, the statistics of pairwise association for each hour-covariate pair within the EHR data for every 6 hours window-duration with selected 24 covariates is described using pointwise mutual information (PMI) matrix. This matrix gives the heterogeneity of data as a two-dimensional map. Such matrices are fused horizontally along the z-axis as vertical slices in the xy plane to form a 3-way tensor for each record with the corresponding Length of Stay (L). Tensor factorization of such fused tensor for every record is performed using Tucker decomposition, and only the core tensors are retained later, excluding the 3 unitary matrices to provide the latent feature set for the prediction of sepsis onset. A five-fold cross-validation scheme is employed wherein the obtained 120 latent features from the reshaped core tensor, are fed to Light Gradient Boosting Machine Learning models (LightGBM) for binary classification, further alleviating the involved class imbalance. The machine-learning framework is designed via Bayesian optimization, yielding an average normalized utility score of 0.4519 as defined by challenge organizers and area under the receiver operating characteristic curve (AUROC) of 0.8621 on publicly available PhysioNet/Computing in Cardiology Challenge 2019 training data.

The proposed tensor decomposition of 3-way fused tensor formulated using PMI matrices leverage's higher-order temporal interactions between the pairwise associations among the clinical values for early prediction of sepsis. This is validated with improved risk prediction power for every hour of admission to the ICU in terms of utility score, AUROC, and F1 score. The results obtained show a significant improvement particularly in terms of utility score of $\sim 1.5 - 2\%$ under a 5-fold cross-validation scheme on entire training data as compared to a top entrant research study that participated in the challenge.

Keywords: Sepsis, Machine Learning, Tensor Factorization, Pointwise Mutual Information, Early Prediction, Model-based diagnosis, Electronic health records,Medical informatics.

1. Introduction

Recently evolved medical informatics have led to rapid surge in both quantity and variety of EHRs. This consequently hinders the capability and extent of matrix-oriented data representations and many contemporary analysis algorithms [1]. The advent of tensor-factorizations assist to extend the 2-D view of matrix analysis to higher-order multiple modalities and support compressed representations with dimensionality reduction methods [2, 3, 4, 5, 6]. It further identifies latent information space within the data that adds to meaningful summarizations in terms of both features and instances [7, 8, 9]. The modest literature on the application of tensor-analysis to EHRs of various biomedical fields including critical care, phenotyping, and genotyping reflects that - tensor analysis may serve as an effective tool to probe the frequent updates in the medical domain based on the continued evolution of clinical and scientific evidence. Tensor factorization encourages extensive experimental studies of the EHRs to tackle various associated challenges by including proper choice of design in terms of factorization scheme, scalability of the algorithm under study, and integration of involved temporality [10, 11, 12, 13]. These advantages of tensor analysis make it a promising and reliable modeling approach for early prediction of diseases using the EHR data, which is mostly sparse-multidimensional data and needs advice from domain experts to ensure the validity of the models [3, 14, 15].

Early risk identification facilitates early clinical intervention and disease management and thereby mitigates the economic burden, morbidity, and mortality rates in acute and chronic diseases [16]. However, this can be appreciated only if there is a widespread adaption of EHRs for predictive analytics [17, 18, 19]. The disease onset prediction by harnessing EHRs has evolved as an emerging area of investigation for precision medicine, critical care, and medical informatics [20, 21, 22]. Predictive analytics-based models employing EHRs help clinicians in the future risk assessment of an individual towards the various chronic conditions and also identifies the likelihood of acquiring new diseases. Moreover, such models have advantages in terms of costs/risks compared to manual diagnosis for preventive measures.

Early identification of infectious progression to sepsis in ICU patients may help with appropriate clinical triage and decision-making, assisting in early intervention for improved outcomes [23, 24]. Sepsis suffers from poor prognosis due to various comorbidities resulting from excessive risk factors. Early prediction of sepsis, for beneficial health, is an even more challenging task. However, for faithful early detection and control, a promising direction is to conduct regular monitoring of the risk factors of sepsis onset in high-risk ICU patients. It is also worth to focus the screening for sepsis onset in patients experiencing chronic conditions related to various comorbidities, although it is less likely to confirm the prior detection of sepsis in such cohorts and seems to be a more cumbersome task. In short, it conveys that

*Dr.Shivnarayan Patidar

Email addresses: naimahmed.nesaragi@gmail.com (Naimahmed Nesaragi), shivnarayan.patidar@nitgoa.ac.in (Shivnarayan Patidar), vaneet@purdue.edu (Vaneet Aggarwal)

early detection of sepsis faces challenges of insufficient medical histories of positive (septic) patients and complicated risk factors from a data perspective. To address these challenges, we need an effective prediction model that can better extract insightful knowledge from insufficient medical histories with dominant missing values in EHR data. This calls for more effective implementation of machine learning prediction models to address these issues.

Machine-learning (ML) techniques are data-driven and provide a comprehensive approach towards diagnosis and prognostication, unlike traditional statistical methods[23, 25]. Previously, ML methods have been used in various clinical scenarios for the detection of sepsis [26, 27, 28, 29, 30, 31]. Even various deep learning techniques have tried to gain insight into clinical data of sepsis by sequence modeling, but have not yielded satisfactory performances [32, 33, 34, 35, 36]. The reason is that apart from sequence modeling, there is a need for better feature representation for involved complicated risk factors in the early detection of disease onset.

In this paper, we consider an innovative tensor-based multi-modal signal processing and machine learning method on tabular structured EHR data, the first of its kind to apply mutual information tensor analysis for the early prediction of sepsis. Based on the extent of risk factors involved in the assessment of patients' likelihood for the sepsis onset, it becomes very critical to learn the embeddings of the clinical associations to measure the latent similarity among covariates and sepsis onset. In this paper, PMI is used as an embedding map that characterizes the pairwise clinical associations among covariates, further, tensor-factorization is employed on such PMI matrices to produce a latent feature space that summarizes the interactions between covariates for sepsis onset.

In brief, a 3-way tensor depicting the clinical PMI among given clinical variables for the considered window duration of sepsis onset of 6 hours, and maintaining a stride of 1-hour between such windows, equal to the length of stay is formed for each record. Such a fused tensor for every record undergoes Tucker decomposition with an appropriate choice of rank. The core-tensor resulting after decomposition is retained to form the latent feature set while the 3 other unitary factors obtained after decomposition are excluded. This feature set is fed to the LightGBM machine-learning models designed via Bayesian optimization for clinical decision making.

The contributions of this work can be summarized as follows:

1. This work presents an innovative method that applies tensor-based multi-modal signal processing and machine learning technique for the early prediction of sepsis. This work is first of its kind to apply mutual information tensor analysis on tabular structured EHR data for detection of sepsis onset.

2. The proposed tensor factorization of the PMI matrices based fused tensor deciphers the higher-order latent information that serves as embeddings among the considered clinical covariates for early prediction of sepsis. PMI uncovers the key temporal interactions utilized for the said tensor factorization, by attractively capturing both linear relationships and non-linear dependencies contributing to the prediction of disease onset.

3.As per author's knowledge no contemporary study has obtained such results with this limited count of raw features.

The rest of the paper is organized as follows. Related work is discussed in Section 2. Section 3 deals with various materials and methods employed in this proposed study. Section 4 details the experimental results and comparison with baseline-studies. The various strength and limitations of proposed study and its comparative analysis with other sate-of-art are discussed in Section 5. Section 6 concludes the research.

2. Related work

The recent modest literature on the early prediction of sepsis appears from the PhysioNet 2019 Challenge [37] that aimed at the development of various effective signal processing, machine learning, and deep learning schemes for the early prediction of sepsis six hours before clinical recognition of sepsis onset from EHR data. True (correct) sepsis predictions up to 12 hours before sepsis onset were rewarded, whilst false (incorrect) predictions were lightly penalized, and predictions that failed to detect sepsis near disease onset were strongly penalized. The choice of 6 hours window (optimal) comes from the fact that the observed median time to antimicrobial therapy is found to be 6 hours [38]. Furthermore, each hour in the delay of treatment is associated with an average decrease in survival of 7.6% [38].

Early prediction of such chronic diseases assists clinicians with effective and improved treatment. Effective treatment of sepsis in the intensive care unit (ICU) requires early antibiotic administration and injecting intravenous (IV) fluids immediately after diagnosis. The time to initiation of effective antimicrobial therapy following sepsis-induced hypotension (i.e., septic shock) is very crucial and vital in predicting the outcome of success and effectiveness of any form of treatment. Furthermore, early treatment reduces sepsis mortality. Most of these state-of-art submissions [39, 40, 41, 42, 43, 44] rely on extracting the hand-crafted features like statistical, dynamic (derivative), etc., from multiple domains since it is challenging to extract the required feature representation for disease onset with EHR data. In this regard, our recent work [45] aimed at deploying soft computing and machine learning scheme for prediction of sepsis onset, wherein, various clinically potential ratios and powers were computed as descriptors by applying a genetic algorithm on given raw covariates for both feature transformation and optimal selection.

Regardless of faithful results obtained by these predictive analytics based risk assessment models for sepsis onset prediction, the existing literature is naïve towards the noteworthy application of higher-order multi-modal techniques such as tensor analysis on such sparse multivariate time-series clinical data for sepsis onset prediction [46, 47, 37].

3. Materials and Methods

3.1. Study Population and Data Sources

The data considered for the study is the PhysioNet/Computing in Cardiology Challenge 2019 [37] data retrieved from ICU patients of three separate hospital systems A, B, and C.

EHR data of 40336 ICU patients (37,404 non-sepsis and 2,932 sepsis) from A and B are publically available for training and model development, while sequestered data of 24,819 from the remaining content of A and B, along with independent hospital C forms the test set. Each patient's record data consists of multivariate clinical time series of 40 covariates (26 laboratory, 8 vital signs, and 6 demographic variables), recorded for every hour of admission to ICU. The custom defined utility function is used for scoring the predictions. The Sepsis-3 clinical criterion is used for the annotation of considered EHRs. Consent of approval from the appropriate institutional review boards is obtained by challenge organizers to collect data [37]. The publically available EHR data has a high density of missing information. Table.1 presents the given 40 clinical variables along with their percentage of missing information. For this study clinical covariates having missing information less than 95% are only considered which corresponds to 24 clinical covariates.

3.2. Data Pre-processing

The proposed approach employs data imputation as a pre-processing step to fill in the missing values. In this work, the *'forward fill'* imputation method, which is most commonly used imputation technique on the given sequential EHR data is employed [48, 49, 50]. In a real-time setting, the current missing value encountered needs to be filled with available previous measurements to adhere to causality. Hence, only the previous values of EHR data are fetched in advance for imputation of every current observation in hand.

In this study, data imputation is performed in two rounds: first locally, for individual records, and then globally for the combined records. During local imputation, the trailing missing values for a particular clinical covariate (or feature vector) in a row are forward filled with the previous neighboring non-missing value locally in that row for the given record. Ipso facto, if the record encounters 'NaN' values, in the beginning, i.e., for the first alone measurement at time instant *t=0*, they are retained as it is initially and then later replaced with *'global mean'* for that particular covariate.

As an illustration, let us consider EHR data as a matrix A with clinical values of covariates arranged across rows of the matrix. The missing values are first filled locally across the columns (time- axis) to obtain matrix F' and later globally with their respective mean values to obtain matrix F.

$$A = \begin{bmatrix} NaN & NaN & 5 & 3 & NaN & 5 & 7 & NaN & 9 & NaN \\ 8 & 9 & NaN & 1 & 4 & 5 & NaN & 5 & NaN & 5 \\ NaN & 4 & 9 & 8 & 7 & 2 & 4 & 1 & 1 & NaN \end{bmatrix}$$

$$F' = \begin{bmatrix} NaN & NaN & 5 & 3 & 3 & 5 & 7 & 7 & 9 & 9 \\ 8 & 9 & 9 & 1 & 4 & 5 & 5 & 5 & 5 & 5 \\ NaN & 4 & 9 & 8 & 7 & 2 & 4 & 1 & 1 & 1 \end{bmatrix}$$

$$F = \begin{bmatrix} 5.8 & 5.8 & 5 & 3 & 3 & 5 & 7 & 7 & 9 & 9 \\ 8 & 9 & 9 & 1 & 4 & 5 & 5 & 5 & 5 & 5 \\ 4.1 & 4 & 9 & 8 & 7 & 2 & 4 & 1 & 1 & 1 \end{bmatrix}$$

Table 1: Details of the given clinical covariates with missing values information in percentage

Sr. No.	**Covariates**	**Units**	**Missing values(%)**
1	Heart rate	beats/min	9.8
2	O_2Sat	%	13
3	Temperature	^oC	66
4	SBP	mm Hg	14.5
5	MAP	mm Hg	12.45
6	DBP	mm Hg	31.34
7	Resp	breaths/min	15.35
8*	$EtCO_2$	mm Hg	96.28
9*	Excess bicarbonate	mmol/L	95.57
10	Bicarbonate	mmol/L	94.81
11	FiO_2	%	91.66
12	pH	-	93.06
13	$PaCO_2$	mm Hg	94.44
14*	SaO_2	%	96.54
15*	Asparatate transaminase	IU/L	98.37
16	BUN	mg/dL	93.13
17*	Alkaline phosphatase	IU/L	98.39
18	Calcium	mg/dL	94.11
19	Chloride	mmol/L	94.46
20	Creatinine	mg/dL	93.90
21*	Direct bilirubin	mg/dL	99.8
22	Gl	mg/dL	82.89
23*	Lactic acid	mg/dL	97.32
24	Magnesium	mmol/dL	93.68
25	Phosphate	mg/dL	93.98
26	Potassium	mmol/L	90.68
27*	Total bilirubin	mg/dL	98.50
28*	Troponin I	ng/mL	99.04
29	Hematocrit	%	91.14
30	Hemoglobin	g/dL	92.61
31	PTT	s	94.05
32	WBC	count/L	93.59
33*	Fibrinogen concentration	mg/dL	99.34
34	Platelet count	count/mL	94.05
35*	Age	yr	0
36*	Gender	Female (0) or male (1)	0
37*	Unit 1	false (0) or true (1)	39.42
38*	Unit 2	false (0) or true (1)	39.42
39*	HospAdmTime	hours	0
40*	LOS	hours	0
41	SepsisLabel	Is 1 for septic patients, and 0 for nonseptic patients.	0

*Excluded Covariates.

Note: O_2Sat: Pulse oximetry, SBP: Systolic Blood Pressure, MAP: Mean arterial pressure, DBP: Diastolic Blood Pressure, Resp: Respiration rate, $EtCO_2$: End tidal carbon dioxide, FiO_2: Fraction of inspired oxygen, $PaCO_2$: Partial pressure of carbon dioxide from arterial blood, SaO_2: Oxygen saturation from arterial blood, BUN: Blood urea nitrogen, Gl: Serum glucose, PTT: Partial thromboplastin time, WBC: Leukocyte count, Unit 1: Administrative identifier for ICU unit (medical ICU), Unit 2: Administrative identifier for ICU unit (surgical ICU), HospAdmTime: Time between hospital and ICU admission, LOS: ICU length of stay.

3.3. Pointwise Mutual Information

Mutual information measures the amount of information that a given random variable contains about another random variable. It relates to the reduction in the uncertainty of one random variable due to the presence of the other. Pointwise mutual information (PMI) measures the association looking at particular instances of the two random variables X and Y. More specifically, PMI measures the difference between the probability of their co-occurrence given their joint distribution and the probability of their co-occurrence given the marginal distributions of X and Y (thus assuming the two random variables are independent).

PMI between hour-covariate pair, for the given sepsis onset prediction window (W) of 6 hours duration is defined as in (1)

$$\mathrm{PMI}(f) = \log \frac{p(f_{i,j})}{p(f_i)p(f_j)} \tag{1}$$

where,

$$p(f_{i,j}) = \frac{f_{i,j}}{\sum_{i=1}^{6}\sum_{j=1}^{24} f_{i,j}}, \quad p(f_i) = \frac{\sum_{j=1}^{24} f_{i,j}}{\sum_{i=1}^{6}\sum_{j=1}^{24} f_{i,j}}, \quad p(f_j) = \frac{\sum_{i=1}^{6} f_{i,j}}{\sum_{i=1}^{6}\sum_{j=1}^{24} f_{i,j}} \tag{2}$$

Here, $f_{i,j}$ is the EHR data of 24 covariates for the window W with 6 hours duration, $i \subseteq \{1,2,\ldots,6\}$ and $j \subseteq \{1,2,\ldots,24\}$.

$P(f_{i,j})$ is the joint probability distribution that measures the pairwise association between the covariates for the particular time instant under the given window W, $P(f_i)$ is the marginal probability distribution that measures the association of all the 24 covariates for a given hour, and $P(f_j)$ is the marginal probability distribution that measures the association of given covariate for all 6 hours.

PMI measures how far the relativeness of clinical observations obtained on a given pair of covariates for a particular time instant is from what would be expected if the covariates were independent of each other.

Such PMI matrix entails relationships between the clinical values of the different covariates within the micro-environment of sepsis onset window, wherein differences may be compared from hour-to-hour, covariate-to-covariate, and finally subject-to-subject. Overall the heterogeneity in EHR data is quantified by the PMI matrix [51], wherein the entries reveal the measure of how frequently a temporal interaction between two covariates for the given hour (referenced by the row and column number) occurs within the EHR data as compared to the interactions predicted by the individual (marginal) distributions. Each bin of the PMI matrix represents the joint dependencies among clinical values of given covariates as a pair of different or itself, relative to the marginal distribution of the individual covariates over the entire dataset [51]. A significant temporal interaction would be considered when the PMI value is either very high or very low. High PMI values (more than 1) signify that this particular temporal interaction among covariates occurs more frequently than observed in the marginal distribution alone. Low PMI values (less than 1) signify that this particular

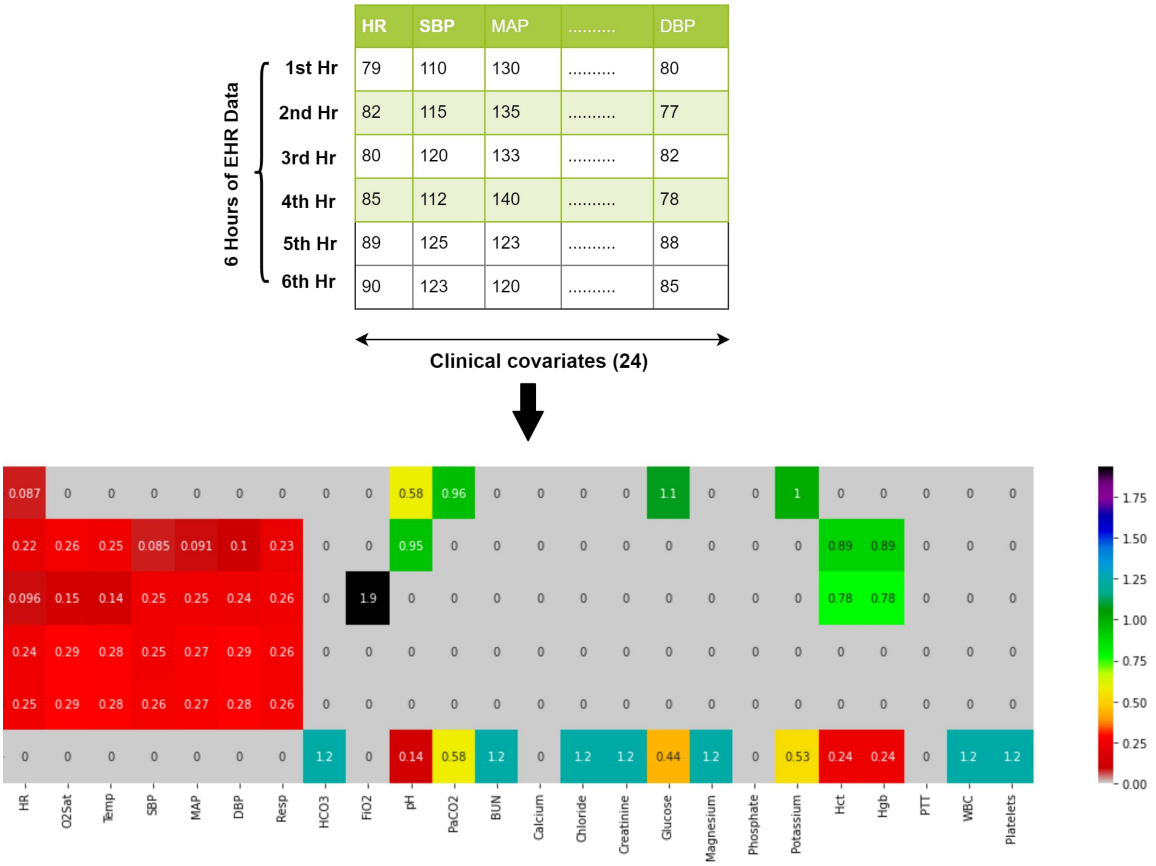

Figure 1: Typical PMI Matrix of EHR data for 6 hours window duration with 24 covariates

temporal interaction among covariates was observed less frequently than expected from the marginal distribution. PMI values (close to 0) signify interactions that may adequately be described by the marginal distribution.

This PMI matrix $f_{i,j}$ can be treated as a vertical slice of the full tensor computed using every 6 hours of EHR data with selected 24 covariates as shown in the Figure 1.

3.4. Tensor-Learning of PMI matrices for Prediction of Sepsis Onset

The proposed Tensor-factorization method uses PMI matrices that present pairwise statistics of clinical associations from selected clinical covariates as input for the given length of stay (L). Three-way tensors are formulated to envision the high-dimensional modalities among covariates of the patients for every hour after admission to ICU.

The algorithmic approach for the formulation of tensor followed by its decomposition is explained as follows: It is to be noted that,there is no unique way or there are multiple ways to build a higher-order tensor for a given EHR data of a record. However, we preferably resort to choose a 3-way tensor that can present ternary interactions among every hour, given covariates, and the given length of stay. The xy plane is chosen to represent the

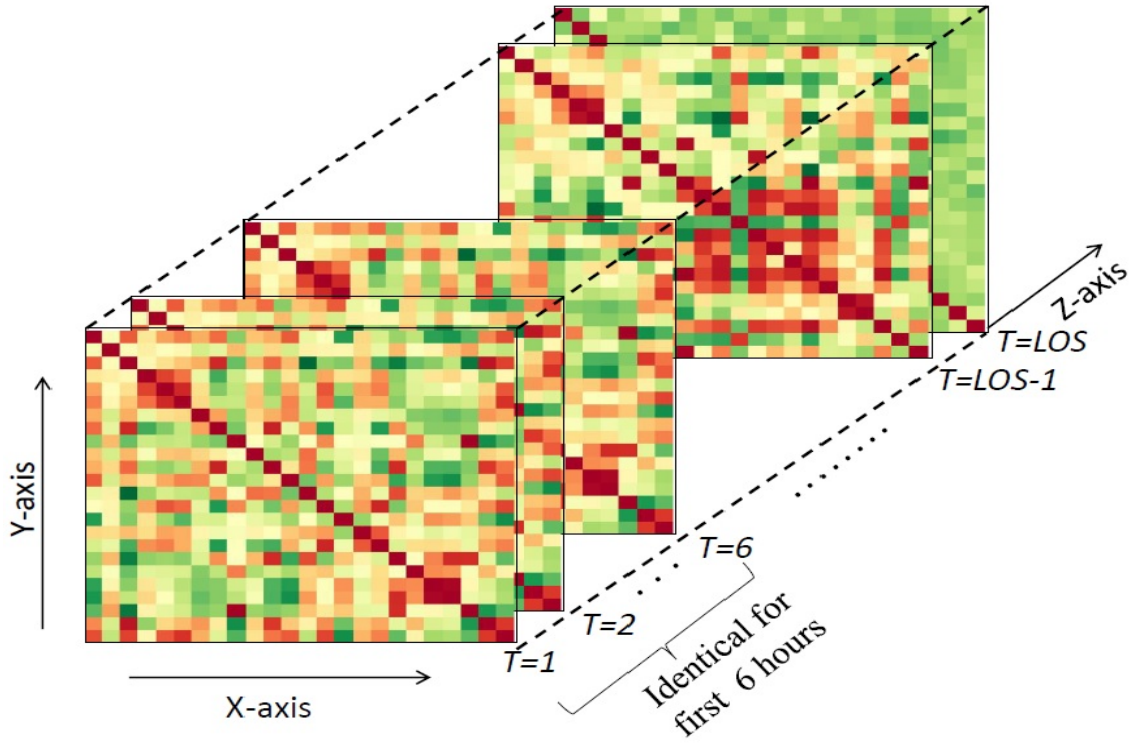

Figure 2: Input data representation for proposed tensor factorization.Each slice along z-axis is a PMI matrix corresponding to every 6 hours of EHR data, arranged with a stride of 1 hour.

interactions between the selected covariates for considered hours, with rows representing every hour data and columns representing covariates. The task under consideration is to detect sepsis onset before 6 hours. So, accordingly, the x-axis represents every 6 hours of data, the y-axis represents 24 covariates, and the z-axis represents the count of such xy time-frames for the given length of stay computed with the stride of 1 hour for every 6 hours.

We collected all the PMI matrices equal to the length of stay from an individual record and built a third-order observed tensor $\theta_{x,y,z}$. Entries at $(x; y; z)$ of the tensor $\theta_{x,y,z}$ is the PMI between any 2 covariates considered at the given hour, for a window of 6 hours, along the x and y-axis, and the z-axis represents the count of such matrices for every 6 hours of EHR data, computed with a stride of 1 hour, equal to L resulting in dimension $\theta \in \mathbb{R}^{6\times 24\times L}$.

In the process of constructing the tensor $\theta_{x,y,z}$ for every record, the first 6 PMI matrices are kept identical which represents the pairwise associations between the selected covariates for the first 6 hours of data, i.e., PMI matrices from $T = 1$ to $T = 6$ are identical. Next, the trailing matrices from $T = 7$ (representing pairwise associations between covariates for next 6 hours of data, i.e., from 2^{nd} hour to 7^{th} hour), $T = 8$ (representing pairwise associations between covariates for next 6 hours of data, i.e., from 3^{rd} hour to 8^{th} hour), and so on to $T = L$ are unique and are fused to form a three-way tensor $\theta_{x,y,z}$ as shown in Figure 2.

Such a fused tensor $\theta_{x,y,z}$ for every record undergoes tucker decomposition [52, 53] with

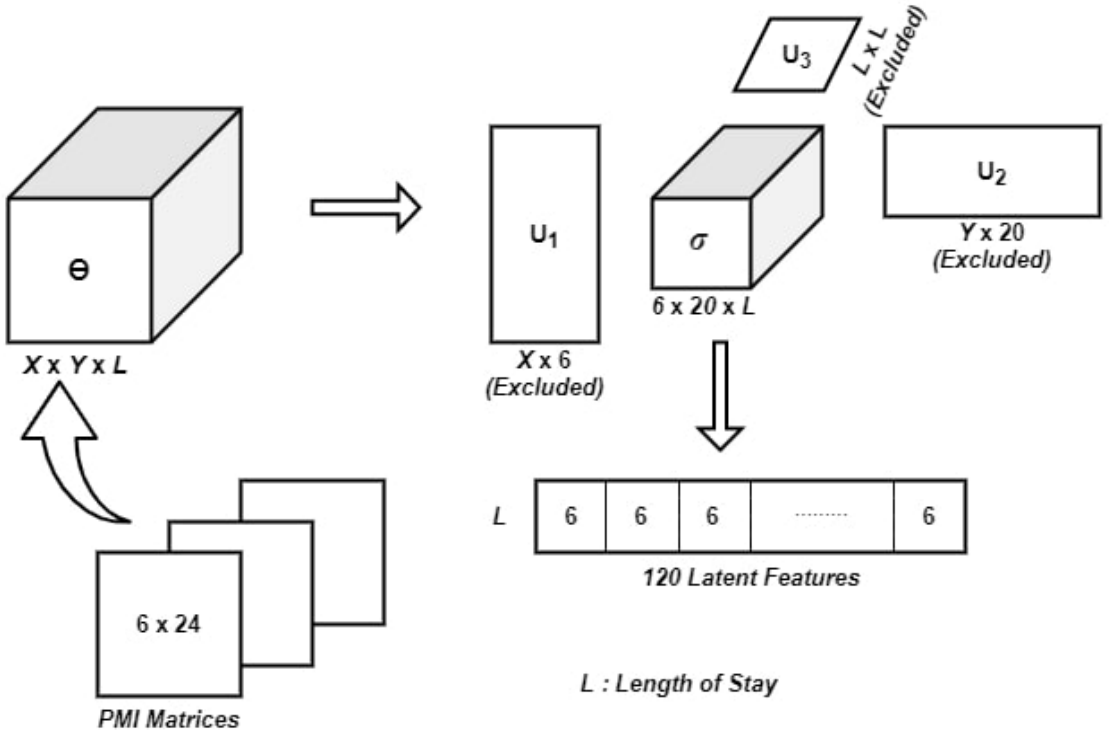

Figure 3: Proposed tensor-factorization for prediction of sepsis onset

rank $\mathbb{R} \in 6 \times 20 \times L$. Decomposition yields core-tensor $\sigma_{x,y,z}$ and factor matrices (U_1, U_2, U_3) as $latent factors$ with dimensions $\sigma \in \mathbb{R}^{6\times20\times L}$, $U_1 \in \mathbb{R}^{6\times6}$, $U_2 \in \mathbb{R}^{24\times20}$, and $U_3 \in \mathbb{R}^{L\times L}$ respectively as shown in Figure 3. Among the latent factors, only the $\sigma_{x,y,z}$ is retained and all other factor matrices are excluded. The reason behind this is that U_1 and U_2 have information limited to only hours and covariates within duration (W), and nothing to say about length of stay, whilst U_3 has information limited only to length of stay and no information about hours and covariates interactions within (W). So the ubiquitous choice is the core-tensor, that not only includes the temporal interactions of hours and covariates within (W), but also across the entire length of stay for a record. This assist the design to form the feature set in terms of predictions for given length of stay as input to machine learning framework.

So, now the extracted $\sigma_{x,y,z}$ is unfolded to give 120 latent features as a feature set. In the context of this sepsis onset prediction task, we explored matrix/tensor factorization to extract these latent features from the core of every patient-tensor representing ternary temporal interactions among covariates. Then, the machine learning models are trained with these latent features for the said tensor-factorization prediction method. Please refer to this recent review [3, 54] for preliminaries and further details on the matrix and tensor factorizations.

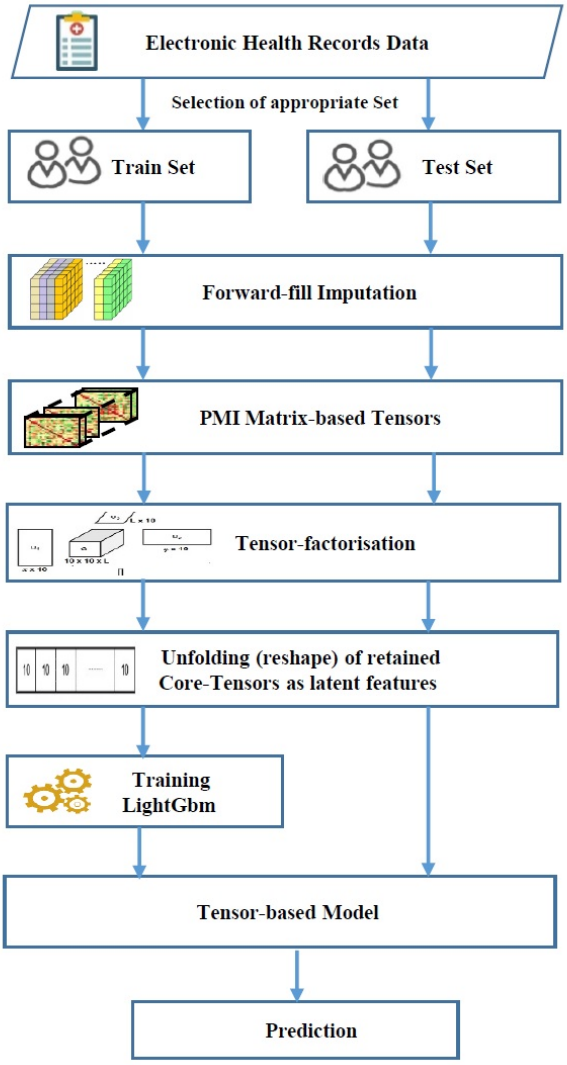

Figure 4: Workflow of sepsis onset prediction system

3.5. Sepsis onset prediction system

In general practice, Sepsis-3 guidelines are recommended by the involved task force which is all about a two-point change in the patient's Sequential Organ Failure Assessment (SOFA) score and clinical suspicion of infection (as defined by the ordering of blood cultures or IV antibiotics)[55]. These guidelines involve a rule-based procedure associated with manual interpretations. For the given clinical signs, the advantage of such traditional techniques is that they are relatively easy to describe and evaluate [29]. However, the medical assessment of the patient involves measuring sophisticated parameters like Glasgow Coma Score.

On the other hand, machine learning models differ from such traditional screening tools for sepsis which relies on strong conceptual models [29]. The outcome of such machine learning-based automated systems using few easily recorded clinical variables can be po-

Table 2: Hyper-parameters values for the optimized models.

Hyper-parameters	**Values**
lambda_l1	0.544
lambda_l2	9.989
learning_rate	0.0501
max_depth	20
min_data_in_leaf	70
num_leaves	51
reg_alpha	8.163
reg_lambda	8.157
scale_pos_weight	23.50

tentially higher as compared to abundant and cumbersome data. Even recent studies have demonstrated that machine learning algorithms outperform such traditional alternatives in contexts wherein data inputs are abundant and where there is a high potential for complex variable interactions [56, 57, 58]. Because of these attractive features, pattern cum adaptive rule-based machine learning models are supplanting traditional rule-based models. Given the complexity and acknowledged gaps in our understanding of sepsis, primary screening for sepsis seems like an ideal use case for machine learning. Some studies have already shown that machine learning models offer improved sepsis prognostication compared with the Systemic Inflammatory Response Syndrome (SIRS), quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA), and the Modified Early Warning Score (MEWS) among ICU patients[59, 60, 61, 62].

Figure 4 presents the work-flow of the proposed method for the sepsis onset prediction. At the onset, an appropriate set (selected 24 covariates) of EHR data is imputed using forward-fill as described earlier in *subsection* 2.2. The 3-way tensors based on the PMI matrices are constructed and decomposed using tucker decomposition for every record. The extracted cores are unfolded and combined to form the feature set. These features are used to build latent feature set-based gradient boosting machine learning models in a five-fold cross-validation scheme wherein the model in each fold is trained with, 80% of stratified EHR data, and tested with, hold out 20% as a test set for the prediction.

4. Experiments and Results

In this section, the predictive analytics of the proposed framework from the given EHRs is demonstrated to determine the development of sepsis onset risk before 6 hours. The experimental results are evaluated and verified using the custom-defined utility score as given by challenge organizers. The given utility score rewards or penalizes predictions of models within twelve hours before and three hours after the onset of sepsis time and is normalized as described in [37]. The experimental study is performed using a five-fold cross-validation scheme wherein five LightGBM [63] models are developed using patient-wise stratified data in each fold containing a unique 20% of the entire training set.

Table 3: Summary of results for the proposed method.

Fold	**AUROC**	**F1 Score**	**Utility Score Proposed(Previous)**
1	0.8593	0.1405	0.4109(0.3906)
2	0.8605	0.1767	0.4634(0.3915)
3	0.8674	0.1741	0.4859(0.3982)
4	0.8640	0.1893	0.4838(0.4429)
5	0.8597	0.1373	0.4155(0.3927)
Average	0.8621	0.1636	0.4519 (0.4013)
Training on cohort A and Testing on cohort B			0.3157
Training on cohort B and Testing on cohort A			0.4103

Table 4: Summary of the baseline results using selected 24 clinical covariates only under a five-fold cross-validation scheme

Fold	**AUROC**	**F1 Score**	**Utility Score**
1	0.7675	0.1121	0.1744
2	0.7053	0.0920	0.1672
3	0.7514	0.1007	0.2333
4	0.7102	0.0957	0.2220
5	0.7410	0.0979	0.1550
Average	0.7351	0.0968	0.1904

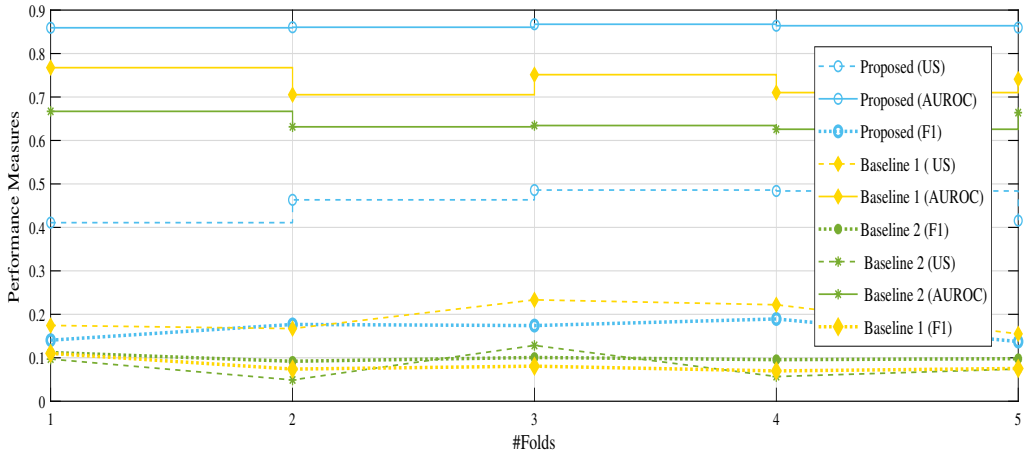

Figure 5: Comparison of results by proposed method with the two base-line studies in a five-fold cross-validation scheme. US:Utility Score, F1: F1 Score, AUROC: Area Under Receiver Operating Characteristic Curve

The hyper-parameter optimization of the model hyperparameters is achieved using automatic utility *'bayesopt'* in Python [64, 65]. The formulation of the underlying objective function is intended to maximize the utility score. At the outset, the values of optimal hyper-parameters that minimize the cross-validation loss using Bayesian optimization are enlisted in Table 2 and the experimental study results are summarised in Table 3. Results also include inter-cohort performances. This is done to probe the reason that models trained are unbiased towards the cohorts and learned solely on the underlying pathology. So to claim the robustness of the unbiased scheme, models trained with the data of cohort A were scored on cohort B data and vice versa. This certainly avoids the doubt of the over-fitting as claimed by [66] in their study. Inter-cohort scores of 0.3157 and 0.4103 are obtained for A and B respectively with significant improvement as compared to 0.3031 and 0.3183, respectively, in [45].

4.1. Comparison with Baselines

Further, to verify and emphasize the advantages and clinical relevance of PMI matrix-based tensor factorization, a subjective analysis of the derived latent feature set is done against the ablation experiments that exclude computation of both PMI matrix with tensor-analysis and secondly, experiments without tensor-factorization alone. Figure 5 depicts the graphical view of the obtained results and its comparison with baseline-studies.

For the aforementioned baseline-studies, we performed two well-tuned simpler methods: In the first method, the imputed version of EHR data with selected 24 covariates data is fed directly to LightGBM models without computing the PMI matrix and without tensor-factorization under a five-fold cross-validation scheme. In the second method, the PMI matrices fused tensor with dimensions $\theta \in \mathbb{R}^{6\times24\times L}$, is reshaped to $\theta \in \mathbb{R}^{(L\times144)}$, and fed to LightGBM models without tensor factorization in a five-fold cross-validation scheme. The

Table 5: Summary of the baseline results using PMI matrices of covariates without tensor-factorization under a five-fold cross-validation scheme

Fold	**AUROC**	**F1 Score**	**Utility Score**
1	0.6670	0.1101	0.0970
2	0.6314	0.0740	0.0490
3	0.6344	0.0807	0.1284
4	0.6258	0.0698	0.0568
5	0.6640	0.0754	0.0741
Average	0.6445	0.0820	0.0811

results of these baseline studies are presented in Tables 4 and 5 accordingly. Hyperparameters are always consistent across evaluations as in Table 2. As we can see, the tensor factorization based methods outperform the other non-tensor based baselines for all the formulations of this task.

Note that exclusion of factor matrices is a sacrifice of rotational invariant components and it can be argued that the retained Tucker core thus exhibits non-uniqueness. However, the core tensor accounts for all the possible complex interactions among the considered tensor components (modes), leading to improved classification results. Infact, injecting more data during the training phase by adding multiple representations/rotations in the data may lead to improved performance [67].

The obtained results reflect the extent to which a tensor-factorization captures third-order relationships between the association of covariates. The quantitative results of our tensor methods validate our approach and justify the fact that third-order information sheds light on second-order information in a way that is not invertible [68]. To illustrate this hypothesis, consider three covariates $z; x; y$. If z has often temporal interactions with both x and y, it is clear that z also often co-occurs with x, but the two separate analogies, that z co-occurs with x and z co-occur with y does not imply anything about the 3-way temporal interaction of $(z; x; y)$. Further, the pairwise association of x and y can be extracted from the 3-way fused and factorized core-tensor data: $\#(x; y) = \sum \#(x; y; z)$. It is thus believable that utilizing third-order factorization can lead to improved insights on tasks that rely on pairwise information.

Determining the optimal value of the rank for tensor-factorization is a difficult task [69]. However, in this study, we empirically experimented with different values of the rank and eventually resort to choose the value of first mode rank (r_1) in tucker decomposition to be 6 (equal to window duration), a second mode (r_2) to 20, and third mode (r_3) equal to the length of stay (L). The reason behind the choice of (r_2) value equal to 20 was just to keep the latent feature set count almost equal to the previous study [45] for a fair comparison. We implemented the proposed method by using the 'tensorly-toolbox' in python [70].

5. Discussion

This study explores the deployment of the PMI matrix to capture the pairwise statistics of clinical associations from selected clinical covariates as input to the tensor factorization. The obtained enhanced results assist us to hypothesize the claim that PMI uncovers the key temporal interactions utilized for the said tensor factorization, by attractively capturing both linear relationships and non-linear dependencies contributing to the prediction of disease onset [71, 51, 72]. At the outset of factorization, the resultant latent-space in the form of the core-tensor-based latent features further enhances the prediction mechanism [7].

As an inspiration from the various proposed techniques in natural language processing for the word embeddings [73, 68], we aim to generalize those approaches in this study by creating pairwise covariates embeddings given by latent-space of core-tensors containing higher-order temporal-interactions of EHR data.Another perspective is that the consideration of third-order moment further contextualizes the PMI associations [72], with additional information that is missing by only considering matrix factorization[68, 7].

The primary contribution of this study is fourfold. First, a limited number of clinical covariates (24 only) from a total of 40 variables are used and obtained enhanced results. Next, higher-order multi-modalities among the pairwise clinical associations are captured for binary classification. Further, this study excludes clinical covariates such as length of stay and other demographics that may bias the algorithms, and finally, achieve an enhanced utility score compared to previous work that used an almost equal number of features and fewer latent feature set in comparison to the contemporary state of arts.

Previous study [45] dealt with static interrelations among the covariates, instead of considering them directly and the enhanced results justified the clinical potential of the extracted feature set for automated analysis of sepsis onset. In literature, such inter-relations like shock-index, PaO2/FiO2, etc. among the covariates expressed as ratio and power have clinical significance in anomaly detection systems [29, 59]. As an extensive analysis, the proposed study explored the potential of higher-order modalities through multidimensional latent information among the pairwise clinical interactions. The enhanced predictive power proves the clinical potential of the proposed approach against our previous study that employs the same amount of features. The comparative results of the proposed work as evaluated against the previous study [45] are shown in Table 3. The proposed model outperforms [45] consistently across each fold and inter-cohort sets. It is to be noted that the same consistent patient-wise stratified folds were used for both studies. A substantial improvement of $\sim 5\%$ in terms of overall utility score as compared to previous work shows that the PMI-based tensor-factorization scheme is quite efficient in the early identification of patients who will develop sepsis in near future.

We have also compared our method with the training results of various submission approaches, which reported the best prediction results in the PhysioNet Challenge 2019 [37]. Most of the submissions employed a 5 or 10 fold cross-validation scheme and obtained utility scores in the range of 0.37 to 0.44 as listed in Table 6. The top-entrant team by [39] derived a signature transform-based regression model in conjunction with a gradient boosting scheme. Yang et al. [41, 74] employed a multi-feature fusion-based XGBoost model for clas-

Table 6: Summary of the results by proposed solutions to PhysioNet Challenge 2019 using 5/10-fold cross-validation on the entire training data.

Reference	**Methodology**	**Utility Score**	**AUROC**
Morrill et al. [39]	A signature-based transform model	0.4340	-
Zabihi et al. [40]	An Ensemble of XGBoost models	0.4280	0.8333
Yang et al. [41]	Fusion-based XGBoost learning	0.4300	0.8400
Chang et al. [42]	Temporal Convolutional Networks (TCN)	0.4170	-
Lee et al. [75]	Graph Convolutional Networks (GCN)	0.3820	0.8170
Du et al. [43]	Gradient Boosted Decision Trees	0.4090	0.8630
Lyra et al. [76]	Using Random forest classification	0.3760	0.8100
Li et al. [44]	A Time-phased model	0.4300	-
Nesaragi and Patidar [45]	Ratio and Power-based Rusboost model	0.4013	0.8432
Proposed Work	PMI-based Tensor factorization	**0.4519**	**0.8621**

sification. Zabihi et al. [40] presented an ensemble scheme of five XGboost models. While Chang et al. [42] employed temporal convolutional networks (TCN),Lee et al. [75] employed graph convolutional networks (GCN). Du et al. [43] introduced a gradient boosting decision tree (GBDT)based classifier to estimate the sepsis onset. Lyra et al. [76] obtained a utility score of 0.376 under a 10-fold cross-validation scheme using random forest, while Li et al. [44] used a time-phased model under a 5-fold cross-validation scheme.Apart from these challenge submissions, recently Rafiei et al. [77] performed a unique predictive analysis study with LSTM-CNN model for different sepsis onset window durations using this challenge data and reported only AUROC, but not the utility score. Another study by Kok et al. [78] used TCN and obtained a utility score of 0.43 under ten-fold cross-validation.

As seen in table 6 the proposed method achieved the best performance on public data especially in terms of overall utility score with $\sim 1.5-2\%$ improvement as compared to top entrant Morrill et al. [39]. It is to be noted that most of the reference methods by Morrill et al. [39], Yang et al. [41, 74], Zabihi et al. [40], and Du et al. [43]'s approaches heavily relied on hand-crafted features for the feature extraction and selection. Further, unlike,Yang et al. [41, 74], Zabihi et al. [40],Du et al. [43], even this proposed study employed the LightGBM based GBDT scheme, which might have limited capability in handling high-dimensional features. But, the tensor-factorization of the pairwise associations among the covariates developed in this study was able to not only capture the high-dimensional features in time series observations but also utilize the pairwise associations to further improve prediction performance. There are a few important limitations to be highlighted and dealt with in this study. First, the predictive analysis is limited to the retrospective setting of publicly available EHR data from 2 cohorts. So the generalization and stability of the proposed model need to be evaluated systematically in perspective settings. Second, the investigation of risk prediction performance is according to custom-defined task-specific utility score that reward predictions between twelve hours before and three hours after the onset of sepsis (Sepsis-3 guidelines) with an optimum window definition of 6 hours before onset. The performance evaluation with different T hours (where T = 6, 12, 18, or 24) is to be explored with different

optimum and prediction window definitions between before and sepsis onset.

It can be argued that besides PMI, many other non-parametric and robust correlation measures such as Pearson correlation, distance correlation, and maximal information coefficient could have been the choice to depict the clinical associations. However, such correlation analysis takes into account only the linear relationship between multiple variables whereas PMI can attractively capture both linear relationships and non-linear dependence but again limited to pairwise associations [71]. The proposed method does consider the non-linear correlation between multivariate as an added advantage through PMI-based tensor analysis [72]. Thus, a multivariate-interaction non-linear analysis method based on mutual information tensor analysis is proposed.

Deep learning schemes with neural networks could have been applied as an alternative to learn such multiplicative interactions and might have yielded satisfactory results. However, this comes at the penalty of high computational cost [79]. It should be noted that tree-based methods can be best thought of as scaled-down versions of neural networks [80, 81], thereby approaching feature classification, information flow, optimization, etc. in simpler terms. For such a structured (tabular) EHR data deterministic modeling approach is more suitable and consistent. The tree-based methods rely on such a deterministic approach as compared to a probabilistic approach by neural networks. Furthermore, most of the challenge submissions [35, 40, 76, 82] reported boosting schemes to have emerged as better candidates, since they efficiently handle the class-imbalance problems associated with this task, when compared to the neural networks. Therefore, we refrained from a baseline study with deep learning schemes in the current study, and resort to GBDT for solving the class-imbalance problem.

In this study, the enhanced results of the proposed method as compared to baselines validates that, factorizing the third moment provably recovers the parameters of the implanted latent temporal associations for the sepsis onset, in a way that only capturing pair-wise interactions, i.e. the second-order moment alone uniquely cannot [68]. Tensor factorization turns out to be a highly effective and applicable tool for learning such embeddings, with increased potential. Leveraging higher-order data assists in encoding new types of latent information and semantic relationships compared to models that are trained using only pairwise data [7].

6. Conclusion

Uncovering temporal interactions among the covariates of patient records are very important for the diagnosis of sepsis and its prevention onset and diagnosis of diseases. Most existing state of art relies on extracting the handcrafted features like statistical, derivate, etc from multiple domains since it is challenging to extract the features for disease with EHR data. This research demonstrates that tensor factorization is an appropriate and reliable modeling technique to encapsulate the inherent quiescent high-dimensional interrelations among given clinical covariates of the patients and sequential patterns for the sepsis onset prediction. The proposed tensor-based machine learning predictive models can predict the sepsis onset at the individual level with enhanced utility scores as compared to benchmark machine learning techniques.

The proposed approach achieves significant improvements compared with existing state-of-the-art algorithms when applying the learned latent representations for hour-hour, covariate-covariate, and patient-patient association prediction. To the best of the authors' knowledge, our approach is the first of its kind in the literature that has employed a tensor factorization technique for exploiting the temporal interactions using PMI to study the sepsis onset prediction.

In most patients, admitted with septic cases, the condition remains undiagnosed or incorrectly diagnosed because of delayed prognosis, thus sepsis has significant mortality among patients admitted to ICUs.Therefore, using algorithms with significant clinical potential can assist in early clinical intervention, and thereby mitigate the economic burden, morbidity and mortality rate of sepsis, and the ICU length of stay. The encouraging results of this study may aid clinicians in predicting sepsis onset in ICUs after it has been trialed in a prospective clinical setting.

Our future research work in the pipeline involves the employment of graph signal processing techniques and kernel-based entropy methods and testing by acquiring data from more cohorts with different optimum onset definitions.

7. Conflict of interest statement

Authors have no competing interests to declare.

References

[1] Y. Luo, F. Wang, P. Szolovits, Tensor factorization toward precision medicine, Briefings in bioinformatics 18 (2017) 511–514.

[2] Z. Rahimi, M. M. Homayounpour, Tens-embedding: A tensor-based document embedding method, Expert Systems with Applications 162 (2020) 113770.

[3] E. E. Papalexakis, C. Faloutsos, N. D. Sidiropoulos, Tensors for data mining and data fusion: Models, applications, and scalable algorithms, ACM Transactions on Intelligent Systems and Technology (TIST) 8 (2016) 1–44.

[4] M. Hong, J. J. Jung, Multi-criteria tensor model for tourism recommender systems, Expert Systems with Applications (2020) 114537.

[5] H. Wang, Q. Zhang, Y. F. Chen, E. Leung, E. Wong, E.-K. Yeoh, Tensor factorization-based prediction with an application to estimating the risk of chronic diseases, BioRxiv (2019) 810556.

[6] W. Wang, Y. Sun, B. Eriksson, W. Wang, V. Aggarwal, Wide compression: Tensor ring nets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9329–9338.

[7] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, M. Telgarsky, Tensor decompositions for learning latent variable models (a survey for alt), in: International Conference on Algorithmic Learning Theory, Springer, pp. 19–38.

[8] S. Khan, G. Xu, R. Chan, H. Yan, An online spatio-temporal tensor learning model for visual tracking and its applications to facial expression recognition, Expert Systems with Applications 90 (2017) 427–438.

[9] J. Chambua, Z. Niu, A. Yousif, J. Mbelwa, Tensor factorization method based on review text semantic similarity for rating prediction, Expert Systems with Applications 114 (2018) 629–638.

[10] Y. Luo, F. S. Ahmad, S. J. Shah, Tensor factorization for precision medicine in heart failure with preserved ejection fraction, Journal of cardiovascular translational research 10 (2017) 305–312.

[11] S. Rabanser, O. Shchur, S. Günnemann, Introduction to tensor decompositions and their applications in machine learning, arXiv preprint arXiv:1711.10781 (2017).

[12] E. C. Chi, T. G. Kolda, On tensors, sparsity, and nonnegative factorizations, SIAM Journal on Matrix Analysis and Applications 33 (2012) 1272–1299.
[13] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (2009) 455–500.
[14] H. Wang, Q. Zhang, H.-Y. So, A. Kwok, Z. S.-Y. Wong, Temporal prediction of in-hospital falls using tensor factorisation, BMJ innovations 4 (2018).
[15] D. Goldfarb, Z. Qin, Robust low-rank tensor recovery: Models and algorithms, SIAM Journal on Matrix Analysis and Applications 35 (2014) 225–253.
[16] D. Yach, C. Hawkes, C. L. Gould, K. J. Hofman, The global burden of chronic diseases: overcoming impediments to prevention and control, Jama 291 (2004) 2616–2622.
[17] R. Kohli, S. S.-L. Tan, Electronic health records: how can is researchers contribute to transforming healthcare?, Mis Quarterly 40 (2016) 553–573.
[18] G. Shmueli, O. R. Koppius, Predictive analytics in information systems research, MIS quarterly (2011) 553–572.
[19] K. He, S. Huang, X. Qian, Early detection and risk assessment for chronic disease with irregular longitudinal data analysis, Journal of biomedical informatics 96 (2019) 103231.
[20] E. A. Ashley, The precision medicine initiative: a new national effort, Jama 313 (2015) 2119–2120.
[21] F. S. Collins, H. Varmus, A new initiative on precision medicine, New England journal of medicine 372 (2015) 793–795.
[22] R. Mirnezami, J. Nicholson, A. Darzi, Preparing for precision medicine, N Engl J Med 366 (2012) 489–491.
[23] G. Wardi, M. Carlile, A. Holder, S. Shashikumar, S. Hayden, S. Nemati, Predicting progression to septic shock in the emergency department using an externally generalizable machine learning algorithm, medRxiv (2020).
[24] H. J. Kam, H. Y. Kim, Learning representations for the early detection of sepsis with deep neural networks, Computers in biology and medicine 89 (2017) 248–255.
[25] G. Wardi, S. Shashikumar, M. Carlile, M. Krak, S. Hayden, A. Holder, S. Nemati, 301 use of transfer learning to improve external validity of a machine-learning algorithm to predict septic shock in the emergency department, Annals of Emergency Medicine 76 (2020) S116.
[26] A. Gupta, M. T. Lash, S. K. Nachimuthu, Optimal sepsis patient treatment using human-in-the-loop artificial intelligence, Expert Systems with Applications (2020) 114476.
[27] F. Cismondi, A. L. Horn, A. S. Fialho, S. M. Vieira, S. R. Reti, J. M. Sousa, S. Finkelstein, Multi-stage modeling using fuzzy multi-criteria feature selection to improve survival prediction of icu septic shock patients, Expert Systems with Applications 39 (2012) 12332–12339.
[28] D. W. Shimabukuro, C. W. Barton, M. D. Feldman, S. J. Mataraso, R. Das, Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial, BMJ open respiratory research 4 (2017).
[29] R. J. Delahanty, J. Alvarez, L. M. Flynn, R. L. Sherwin, S. S. Jones, Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis, Annals of emergency medicine 73 (2019) 334–344.
[30] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, T. G. Buchman, An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU:, Critical Care Medicine 46 (2018) 547–553.
[31] C. Barton, U. Chettipally, Y. Zhou, Z. Jiang, A. Lynn-Palevsky, S. Le, J. Calvert, R. Das, Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs, Computers in biology and medicine 109 (2019) 79–84.
[32] D. B. da Silva, D. Schmidt, C. A. da Costa, R. da Rosa Righi, B. Eskofier, Deepsigns: A predictive model based on deep learning for the early detection of patient health deterioration, Expert Systems with Applications 165 (2020) 113905.
[33] M. Saqib, Y. Sha, M. D. Wang, Early Prediction of Sepsis in EMR Records Using Traditional ML Techniques and Deep Learning LSTM Networks, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, Honolulu, HI, 2018, pp. 4038–4041.

[34] C. Lin, Y. Zhangy, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, M. Chi, Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-lstm, in: 2018 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, pp. 219–228.
[35] Y. Wang, B. Xiao, X. Bi, W. Li, J. Zhang, X. Ma, Prediction of sepsis from clinical data using long short-term memory and extreme gradient boosting, in: 2019 Computing in Cardiology (CinC), IEEE, pp. Page–1.
[36] M. Scherpf, F. Gräßer, H. Malberg, S. Zaunseder, Predicting sepsis with a recurrent neural network using the mimic iii database, Computers in biology and medicine 113 (2019) 103395.
[37] M. A. Reyna, C. S. Josef, R. Jeter, S. P. Shashikumar, M. B. Westover, S. Nemati, G. D. Clifford, A. Sharma, Early Prediction of Sepsis From Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019, Critical care medicine 48 (2020) 210–217. Publisher: Lippincott Williams & Wilkins.
[38] A. Kumar, D. Roberts, K. E. Wood, B. Light, J. E. Parrillo, S. Sharma, R. Suppes, D. Feinstein, S. Zanotti, L. Taiberg, et al., Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock, Critical care medicine 34 (2006) 1589–1596.
[39] J. H. Morrill, A. Kormilitzin, A. J. Nevado-Holgado, S. Swaminathan, S. D. Howison, T. J. Lyons, Utilization of the signature method to identify the early onset of sepsis from multivariate physiological time series in critical care monitoring, Critical Care Medicine 48 (2020) e976–e981.
[40] M. Zabihi, S. Kiranyaz, M. Gabbouj, Sepsis Prediction in Intensive Care Unit Using Ensemble of XGboost Models.
[41] M. Yang, X. Wang, H. Gao, Y. Li, X. Liu, J. Li, C. Liu, Early Prediction of Sepsis Using Multi-Feature Fusion Based XGBoost Learning and Bayesian Optimization.
[42] Y. Chang, J. Rubin, G. Boverman, S. Vij, A. Rahman, A. Natarajan, S. Parvaneh, A multi-task imputation and classification neural architecture for early prediction of sepsis from multivariate clinical time series, in: 2019 Computing in Cardiology (CinC), IEEE, pp. Page 1–Page 4.
[43] J. A. Du, N. Sadr, P. d. Chazal, Automated prediction of sepsis onset using gradient boosted decision trees, in: 2019 Computing in Cardiology (CinC), pp. Page 1–Page 4.
[44] X. Li, X. Xu, F. Xie, X. Xu, Y. Sun, X. Liu, X. Jia, Y. Kang, L. Xie, F. Wang, et al., A time-phased machine learning model for real-time prediction of sepsis in critical care, Critical Care Medicine 48 (2020) e884–e888.
[45] N. Nesaragi, S. Patidar, Early prediction of sepsis from clinical data using ratio and power-based features, Critical Care Medicine 48 (2020) e1343–e1349.
[46] B. A. Goldstein, A. M. Navar, M. J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, Journal of the American Medical Informatics Association 24 (2017) 198–208.
[47] B. A. Goldstein, M. J. Pencina, Developing implementable risk prediction models with electronic health records data, Wiley StatsRef: Statistics Reference Online (2014) 1–8.
[48] Z. C. Lipton, D. C. Kale, R. Wetzel, Modeling missing data in clinical time series with rnns, arXiv preprint arXiv:1606.04130 (2016).
[49] R. J. Little, D. B. Rubin, Statistical analysis with missing data, volume 793, John Wiley & Sons, 2019.
[50] D. B. Rubin, Inference and missing data, Biometrika 63 (1976) 581–592.
[51] D. M. Spagnolo, R. Gyanchandani, Y. Al-Kofahi, A. M. Stern, T. R. Lezon, A. Gough, D. E. Meyer, F. Ginty, B. Sarachan, J. Fine, et al., Pointwise mutual information quantifies intratumor heterogeneity in tissue sections labeled with multiple fluorescent biomarkers, Journal of pathology informatics 7 (2016).
[52] A. H. Phan, A. Cichocki, Tensor decompositions for feature extraction and classification of high dimensional datasets, Nonlinear theory and its applications, IEICE 1 (2010) 37–68.
[53] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, H. A. Phan, Tensor decompositions for signal processing applications: From two-way to multiway component analysis, IEEE signal processing magazine 32 (2015) 145–163.
[54] R. Orús, A practical introduction to tensor networks: Matrix product states and projected entangled

pair states, Annals of Physics 349 (2014) 117–158.
[55] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J.-L. Vincent, D. C. Angus, The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), JAMA 315 (2016) 801.
[56] A. L. Beam, I. S. Kohane, Big data and machine learning in health care, Jama 319 (2018) 1317–1318.
[57] L. Breiman, et al., Statistical modeling: The two cultures (with comments and a rejoinder by the author), Statistical science 16 (2001) 199–231.
[58] R. J. Delahanty, D. Kaufman, S. S. Jones, Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients, Critical care medicine 46 (2018) e481–e488.
[59] K. E. Henry, D. N. Hager, P. J. Pronovost, S. Saria, A targeted real-time early warning score (trewscore) for septic shock, Science translational medicine 7 (2015) 299ra122–299ra122.
[60] J. S. Calvert, D. A. Price, U. K. Chettipally, C. W. Barton, M. D. Feldman, J. L. Hoffman, M. Jay, R. Das, A computational approach to early sepsis detection, Computers in Biology and Medicine 74 (2016) 69–73.
[61] J. Calvert, T. Desautels, U. Chettipally, C. Barton, J. Hoffman, M. Jay, Q. Mao, H. Mohamadlou, R. Das, High-performance detection and early prediction of septic shock for alcohol-use disorder patients, Annals of Medicine and Surgery 8 (2016) 50–55.
[62] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, R. Das, Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach, JMIR Medical Informatics 4 (2016) e28.
[63] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in: Advances in neural information processing systems, pp. 3146–3154.
[64] E. Brochu, V. M. Cora, N. De Freitas, A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, arXiv preprint arXiv:1012.2599 (2010).
[65] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, Advances in neural information processing systems 25 (2012) 2951–2959.
[66] P. Biglarbeigi, D. McLaughlin, K. Rjoob, A. Abdullah, N. McCallan, A. Jasinska-Piadlo, R. Bond, D. Finlay, K. Ng, A. Kennedy, J. McLaughlin, Early prediction of sepsis considering early warning scoring systems, in: 2019 Computing in Cardiology (CinC), pp. Page 1–Page 4.
[67] S. Patidar, Diagnosis of sepsis using ratio based features, in: 2019 Computing in Cardiology (CinC), IEEE, pp. Page–1.
[68] E. Bailey, S. Aeron, Word embeddings via tensor factorization, arXiv preprint arXiv:1704.02686 (2017).
[69] C. J. Hillar, L.-H. Lim, Most tensor problems are np-hard, Journal of the ACM (JACM) 60 (2013) 1–39.
[70] J. Kossaifi, Y. Panagakis, A. Anandkumar, M. Pantic, Tensorly: Tensor learning in python, The Journal of Machine Learning Research 20 (2019) 925–930.
[71] W. Lin, J. Ji, Y. Zhu, M. Li, J. Zhao, F. Xue, Z. Yuan, Pminr: pointwise mutual information-based network regression–with application to studies of lung cancer and alzheimer's disease, Frontiers in genetics 11 (2020) 1043.
[72] L. Lu, X. Ren, C. Cui, Y. Luo, M. Huang, Tensor mutual information and its applications, Concurrency and Computation: Practice and Experience (2020) e5686.
[73] O. Levy, Y. Goldberg, Neural word embedding as implicit matrix factorization, Advances in neural information processing systems 27 (2014) 2177–2185.
[74] M. Yang, C. Liu, X. Wang, Y. Li, H. Gao, X. Liu, J. Li, An explainable artificial intelligence predictor for early detection of sepsis, Critical Care Medicine 48 (2020) e1091–e1096.
[75] B. T. Lee, O.-Y. Kwon, H. Park, K.-J. Cho, J.-M. Kwon, Y. Lee, et al., Graph convolutional networks-based noisy data imputation in electronic health record, Critical Care Medicine 48 (2020) e1106–e1111.

[76] S. Lyra, S. Leonhardt, C. Hoog Antink, Early Prediction of Sepsis Using Random Forest Classification for Imbalanced Clinical Data.

[77] A. Rafiei, A. Rezaee, F. Hajati, S. Gheisari, M. Golzan, Ssp: Early prediction of sepsis using fully connected lstm-cnn model, Computers in biology and medicine 128 (2021) 104110.

[78] C. Kok, V. Jahmunah, S. L. Oh, X. Zhou, R. Gururajan, X. Tao, K. H. Cheong, R. Gururajan, F. Molinari, U. R. Acharya, Automated prediction of sepsis using temporal convolutional network, Computers in Biology and Medicine 127 (2020) 103957.

[79] H. Zhu, X. Li, P. Zhang, G. Li, J. He, H. Li, K. Gai, Learning tree-based deep model for recommender systems, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (2018).

[80] R. Krishnan, G. Sivakumar, P. Bhattacharya, Extracting decision trees from trained neural networks, Pattern recognition 32 (1999).

[81] O. Boz, Extracting decision trees from trained neural networks, in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 456–461.

[82] B. Sweely, A. Park, L. Winter, L. Liu, X. Zhao, Time-padded random forest ensemble to capture changes in physiology leading to sepsis development, in: 2019 Computing in Cardiology (CinC), IEEE, pp. 1–4.