Discriminative Features Generation for Mortality Prediction in ICU

Suresh Pokharel[1(✉)], Zhenkun Shi[2(✉)], Guido Zuccon[1(✉)], and Yu Li[1(✉)]

[1] The University of Queensland, Brisbane, Australia
{s.pokharel,g.zuccon}@uq.edu.au, yuli@itee.uq.edu.au
[2] Jilin University, Changchun, China
shizk14@mails.jlu.edu.cn

Abstract. Effective methods for mortality prediction for Intensive Care Unit (ICU) patients assist health professionals by producing alerts ahead of time regarding the critical changing degeneration of a patient's health. This can guide health professionals to take immediate interventions to rescue the lives of patients. However, existing methods only use raw clinical features and ignore the compound information exhibited by Electronic Health Records (EHRs) data, i.e., the co-occurrence of different clinical events at the same point of time (or within a short time interval). In this paper we use a recently proposed method, called Temporal Tree, to capture the compound information and use them as discriminative features. In addition, to test the impact of preserving temporal information, we capture compound information in terms of patient situations (i.e., the patient's medical condition at particular point of time), and represent a patient as a sequence of patient situations. This is contrasted with the baseline approach of representing a patient by the associated sequence of clinical events (bag-of-words like). These representations are further processed to obtain low dimensional embeddings used to represent patients and their situations.

The effectiveness of the proposed methods is evaluated using a real ICU dataset with nine different baselines and state-of-the-art classifiers. The empirical results show the Temporal Tree method is able to generate discriminative patient representations. Classifiers that exploited the compounded information significantly improved the quality of ICU mortality predictions, in all cases and across both bag-of-words (up to 7.814% improvements) and patient situations representations (up to 2.720% improvements).

Keywords: Compound information · Mortality prediction · Temporal tree

1 Introduction

Intensive care units (ICUs) services are one of the largest ticket items in an hospital operational expenditure budget: according to Coopersmith et al. [7]

X. Yang et al. (Eds.): ADMA 2020, LNAI 12447, pp. 324–338, 2020.
https://doi.org/10.1007/978-3-030-65390-3_25

in fact, these costs amount to between 17.4% and 39.0% of the total hospital costs. ICUs provide highly specialised and costly treatments, such as mechanical ventilation, to acutely ill or injured patients. Among the ICU settings, the prediction of patient likelihood of mortality (i.e. whether the patient dies in ICU or is discharged alive at the end of the ICU stay), is a crucial task for assessing the critically of the patient's illness and of the possible interventions and the treatments parameter settings. However, only between 10% to 15% of ICUs in the US use mortality prediction scoring systems [4], with the major obstacle to adoption being accuracy, cost and applicability to the patient population.

Popular mortality prediction scoring methods such as Sequential Organ Failure Assessment (SOFA) [25], Simplified Acute Physiology Score 3 (SAPS 3) [19], and Acute Physiology and Chronic Health Disease Classification System III (APACHE III) [13] use clinical and laboratory variables for informing the prediction. They do so however based on simple rules or heuristics, and on a limited number of variables. For example, the SOFA score is calculated using data from only six organ systems (respiratory, cardiovascular, hepatic, coagulation, renal and neurological). But the data routinely captured in ICUs is much richer: for example the MIMIC III dataset [12] contains more than 4,000 indicators that influence a patient's conditions. It is not surprising then that research has recently been directed towards using such rich data to further improve the accuracy of ICU mortality prediction: this has taken the form of using machine learning to address this task.

Machine learning research for ICU mortality prediction has seen the application of both traditional classification methods such as Logistic Regression [15,16,26] and Support Vector Machines [9], and of recent deep learning methods such as Long Short-Term Memory (LSTM) [1,24] and Recurrent Neural Networks [5] (although this last line of work attempted to predict the illness severity at any point of time, rather then predict mortality). A common problem all these methods face is producing or learning an effective representation of the data: in either case, previous work has mainly exploited simple temporal characteristics and features of the ICU Electronic Health Records datasets such as MIMIC III (see more detail in Sect. 3.3).

However, EHRs are characterised by

1. *Complex Characteristics*: EHRs are sparse, irregular, temporal, heterogeneous and multivariate
2. *Compound Information*: along with temporal information (events occurring at subsequent points in time), EHRs contain compound information defined as co-occurrence of different clinical events at the same point of time (or within a short time interval).

Existing work ignores these types of inherent relationships between clinical events – but modelling these relationships may provide deeper representational power and insights. These relationships are due to the temporal and multivariate characteristics of EHRs. Figure 1a shows an example of the occurrence of clinical events at different time points for a *chronic renal failure* ICU patient. In the

example, laboratory tests performed on 2153-07-6 at 20:11 report *high magnesium* and *high glucose*; we shall treat this as a single compound information. In EHRs, there are a large amount of these kinds of compound information which, if adequately modelled, may add insights to the EHR representation, and thus likely improve the downstream machine learning tasks, e.g., mortality prediction. Thus, it is desirable to have a way to handle the complex data characteristics as well as to capture the compound information. To achieve this, in this paper we adapt a novel temporal structural network representation called *Temporal Tree* [21].

Temporal Tree is a structural network representation which preserves sequence information related to clinical temporal events and captures compound information. Compound information is generated through a re-labelling approach and is represented in a hierarchical form. A Temporal Tree is constructed for each patient, and the tree preserves the sequence of medical situations for that patient. A medical patient situation, simply called a *patient situation*, is a medical condition observed for a patient, e.g., body temperature, blood pressure, consciousness level, at a particular point of time (or within a short time interval).

Once EHR data is represented by means of a Temporal Tree, patients and their medical situations are represented into a lower dimensional vector space through the use of embeddings. A well known document-level embedding technique [14] (doc2vec) is applied so as to learn lower vector representations. This is done by using an optimisation function tailored to patients as well as their situations. The motivations for adapting the embedding method are that (i) all patients and their situations are represented within the same dense, lower dimensional embedding space, so that processing is easier and faster; and, (ii) similar patients, and similar patients situations, have similar embeddings: this can potentially translate into improvement in the downstream classification tasks.

Patients embeddings are generated in two ways: (1) by treating all clinical events of a patient as a Bag of Word (BOW), so that the corresponding patient vector is generated. In this method, the patient's temporal information is lost. And, (2) by modelling a patient with a sequence of the patient's medical situations, so that then a patient vector is calculated from the patient situation vectors. This preserves the patient's temporal information. In summary, in this paper we provide the following contributions:

1. We adapt the Temporal Tree representation to incorporate compound EHR information, which captures the complex relationships between clinical events, thus providing a richer representation of the EHR data.
2. We adapt an embedding technique to be able to produce embeddings from Temporal Tree to represent a patient. In addition, we show that considering temporal information improves the performance of the downstream task of ICU mortality prediction.
3. We demonstrate the effectiveness of the proposed representation technique using real ICU data, across an extensive set of baselines and state-of-art methods.

2 Related Work

ICU Mortality Prediction. The mortality prediction task for ICU patients has received considerable attention. Existing methods use either structured data, e.g., data from observations such as blood pressure, temperature, laboratory test values [1,11,20,23,24], or a combination of structured and unstructured data, e.g., include clinical notes [9,15]. Our work only focuses on representing and using structured data, leaving the interpretation and representation of clinical notes to future work. Thus, next we consider some of the most representative works that have used structured data.

Suresh et al. [24] have proposed a two-step pipeline to (i) learn patient subgroups using an LSTM autoencoder, and (ii) predict patients mortality for separate subgroup of patients within a multi-task framework. Harutyunyan et al. [11] have proposed linear regression models and neural baselines (a standard LSTM, a channel-wise LSTM, and LSTMs with deep supervision and multitask training) for four prediction tasks: mortality, forecasting the length of stay, detecting physiologic decline, and phenotype classification. Nori et al. [20] have proposed a method which integrates domain knowledge, showing this is more effective than a logistic regression baseline, both with and without multitask learning. Shi et al. [23] have proposed the Deep Interpretable Mortality Model (DIMM), which employs Multi-Source Embedding, Gated Recurrent Units (GRU), Attention mechanism and Focal Loss techniques for mortality prediction. Aczon et al. [1] have proposed a deep learning architecture composed of three LSTMs. Luo et al. [16] have proposed a Subgraph Augmented Non-negative Matrix Factorization (SANMF) where time series data is represented within a graph which is then used to automatically extract the temporal trends by applying frequent subgraph mining. Trends are then grouped using matrix factorization and logistic regression is applied using features from trend groups. Lehman et al. [15] have proposed combining the learned "topic" structure from nursing notes using Hierarchical Dirichlet Processes (HDP) with physiologic data (from the Simplified Acute Physiology Score I (SAPS I)). They then use multivariate logistic regression for hospital mortality prediction. Similarly, Ghassemi et al. [9] used Support Vector Machine for mortality prediction, where they generate aggregated features which are the combination of structured information (age, sex, admitting SAPS-II score as well as derived features such as maximum/minimum SAPS-II score) and features obtained from free-text clinical notes using Latent Dirichlet Allocation. However, all these methods do not consider the compound information.

Patient Embeddings. Patient embedding techniques generate a fixed, low dimensional vectors that are used to represent patients, such that patients that are "similar" are represented by similar embeddings. Intuitively, effective patient embedding techniques may be useful because they may generate representations able to discriminate e.g. between survival and non-survival patients. A number of patient embedding techniques have been recently proposed. Patient2Vec [27] learns an interpretable deep representation for a patient

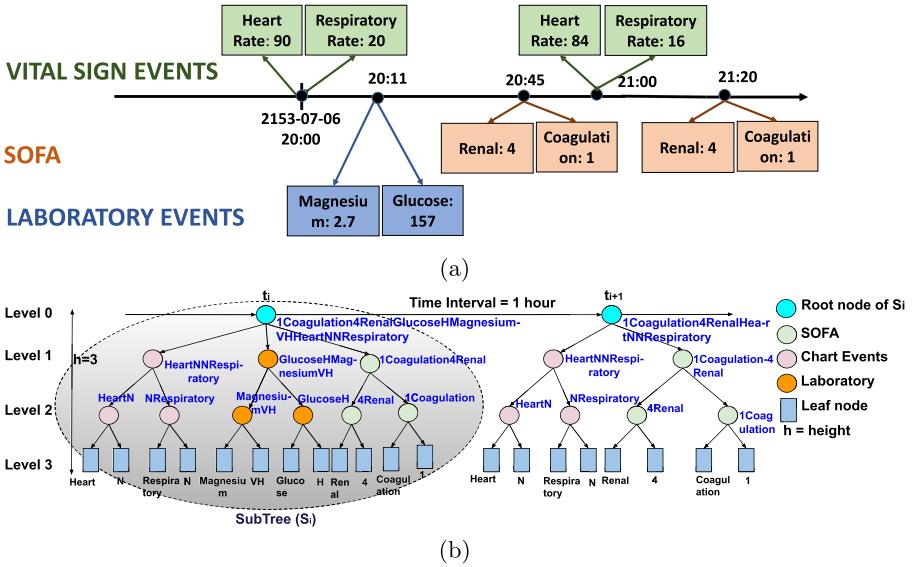

Fig. 1. (a) An example of EHR data for a patient treated in ICU. (b) Temporal Tree representation, where vital signs, sofa and laboratory are shown as examples of events.

by relying on *word2vec* [18] to embed clinical events into vectors – in a similar manner to how we rely on *doc2vec* (a *word2vec* variant) for the same goal. Bajor et al. [2] used the document-level embeddings [14] to represent patients and data elements from clinical codes and laboratory tests, thus generating a sequence of data elements from temporal data. Unlike our work, however, they do not consider the values associated to clinical events and the compound information. Choi et al. [6] proposed an embedding method for learning lower dimensional medical concept representations. Their method considers co-occurrence information along with visit sequence information present in EHRs data, while, Glicksberg et al. [10] used *word2vec* to create medical concept embeddings of the phenotype space and ranked patients based on the distance from the corresponding query embedding. Unlike this prior work, we generate temporal medical data sequences by considering clinical events along with their compound events. In addition, we also consider the patient's temporal situations while generating the patient embeddings so as to retain the temporal information.

3 Methodology

The framework we propose to represent ICU patients data and perform mortality prediction is shown in Fig. 2. The Temporal Tree is constructed for each patient and captures temporal discriminative data. Then, the embedding technique is applied to obtain a low dimensional vector representation of the patient as well as their medical situations. These embeddings are used as input to train

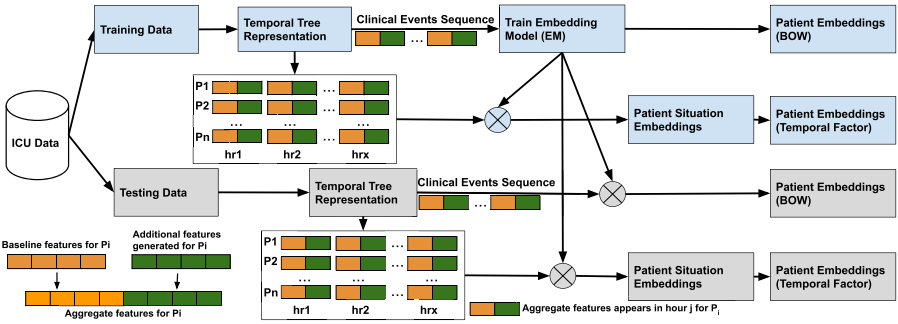

Fig. 2. Framework to represent ICU patients and perform mortality prediction.

a classification model (several considered and evaluated in this work). Details about the components of this framework are provided in the next sub-sections.

3.1 Temporal Tree

The Temporal Tree technique [21] is used to capture and model the compound information present in EHR data. A Temporal Tree is a temporal hierarchical structural network which is constructed based on the temporal co-occurrence of clinical events. An example of Temporal Tree is shown in Fig. 1b. In Temporal Tree, compound information is generated based on the local neighbourhood relationships between clinical events and is represented in a hierarchical form. Here, a leaf node represents the actual clinical events that occur at particular time, while the upper levels of the tree represent compound information. An approach to model event information across multiple levels of the Temporal Tree is to generate compound terms by re-labelling the Temporal Tree using the Weisfeiler-Lehman graph kernels re-labelling method [22]. The Temporal Tree technique offers the following advantages:

1. It models the temporal aspects of the data, along with other complex properties typical of EHR data. Each subtree, S_i, is constructed based on temporal information. A temporal event can encompass any number and any type of attributes (thus modelling multivariate data). Each subtree only considers the available data and there is no need of ad-hoc policies to deal with missing data (thus tackling issues related to irregular and sparse data). The representation also models sparsity in the data due to the frequency and time interval in which events occur, e.g., laboratory test data is often reported at a larger time interval than chart event data. Finally, the heterogeneous nature of EHR data is tackled by the use of abstract values (the more detail is described in Sect. 3.2).
2. No domain knowledge is required to construct the temporal tree.

3.2 Abstract Values

EHR data often contains both numerical and categorical values. To deal with this heterogeneous data, for all the clinical events, we convert each numerical value to one of five categories: Very Low (VL), Low (L), Normal(N), High(H), Very High(VH), using value abstraction [3]. We refer to these as *abstract values* or simply *values*. But in case of SOFA events, we put sofa score as categorical value (refer to Sect. 4.2 for more details).

3.3 Baseline and Aggregate Features

Most existing methods for mortality prediction only use basic temporal features, e.g., sequences of clinical events with their values – we shall refer to those as *baseline features*. However, EHRs contain many compound information which express very important temporal relationships between clinical events; these can be used as additional discriminative features for prediction tasks. While existing methods do not consider these kinds of temporal relationships, they can be captured by constructing a Temporal Tree. Then, the baseline and the additional features from the Temporal Tree can be combined to form a set of *Aggregate Features*. Note that baseline features are also directly extractable from the Temporal Tree: these are the features captured at Level 2 of the Temporal Tree hierarchy (see Fig. 1b). Each Temporal Tree (a tree is constructed for each patient) is used to assemble all baseline features and the Term Frequency - Inverse Document Frequency (TF-IDF) weighting scheme is applied to extract the top additional features in such a way that the amount of additional features is equivalent to the amount of baseline features. The *Aggregate Features* are constructed by appending the additional features to the baseline features.

3.4 Embeddings

Once the Temporal Tree for a patients is constructed, the patient and their patient situations are embedded into fix length lower dimensional vector. For this, we use Paragraph Vector [14], a state-of-art embedding technique. To prepare data ready for training the embedding model, we consider each node label as a word and each Temporal Tree as a document. We then traverse the Temporal Tree using Breath First Search (BFS) to generate the sequence of clinical events which is then the input used for training the embedding model (EM). The EM generates both patient and clinical event embeddings. To generate the patient situation embeddings, we consider a patient situation as being compose of many clinical events (including compound information) at a specific time (or within a short time interval) and thus feed these events into the EM to generate an embedding for each patient situation. Each patient is considered as a sequence of patient situations and a patient vector can thus be generated by taking the average of its patient situation vectors. Note that patient vectors are generated thus in two ways: (1) by representing a patient with a bag-of-words (BOW) of the clinical events relevant to the patient, or, alternatively, (2) by representing

a patient with the sequence of relevant patient situations. The advantage of this second approach is that the embeddings constructed in that way preserve the temporal information about the patient and the sequence of clinical events, while the BOW approach does not preserve such important information.

4 Evaluation Methodology

In this work we are aim to answer the following research questions:

1. Do the aggregate features generated using the Temporal Tree model improve ICU mortality prediction?
2. Does the modelling of patients temporal information ICU mortality prediction, compared to modelling clinical events as bag-of-words?

To this aim, we set up an empirical evaluation that considers classifiers used for the task ICU mortality prediction and explore the impact different representation settings have on the effectiveness of the classifiers. The reminder of this section describes the experiment settings, while empirical results are reported in Sect. 5.

4.1 Dataset and Patient Cohort Selection

To evaluate the proposed approach we use MIMIC III[1] [12], a dataset of publicly available de-identified ICU encounters. It contains structured (e.g., real time sensor data, laboratory tests, prescriptions) as well as unstructured data (e.g., free-text clinical notes) for more than 60,000 ICU admissions between 2001 and 2012 from a US hospital.

The dataset is processed so as to select a cohort of patients that provide a meaningful mean of evaluation mortality prediction methods. Patients were selected according to the following criteria: (i) adults (patients aged 16 years or more), (ii) length of stay in ICU greater than 24 h, and 48 h respectively[2], (iii) have at least one vital signs entry recorded in the dataset (see Table 2 for more details), (iv) have been admitted to ICU for the first time. The reason for excluding re-admitted patients is that it is likely a patient is re-admitted for the same condition, and thus the data would show a high correlation; in addition there are only a small number of re-admissions in the dataset that satisfy the other criteria.

Following the criteria above, we generate a dataset where each ICU admission is regarded as a unique patient. This data is then randomly split into training (this is further divided for folds for cross-validation) and testing subsets using a 80:20 ratio. Details of the patient cohort used in our evaluation are provided in Table 1. Note the dataset presents a bias towards survival (not died) patients.

[1] https://mimic.physionet.org/.

[2] We consider two settings, one based on a minimum length of stay of 24 h, and one of 48 h.

Table 1. Details regarding the patient cohort used for evaluation – patient data is extracted from the MIMIC III dataset.

Time interval	Subset	Number of patients	Not dead (N)	Dead (n)
24 h	Training	25,889	24,130	1,759
	Testing	6,473	6041	432
	Total	32,362	30,171	2,191
48 h	Training	15,947	14,759	1,188
	Testing	3,988	3,687	301
	Total	19,935	18,446	1,489

4.2 Feature Selection

For the experiment, we use the structured data present in MIMIC III for the patient cohort detailed in Sect. 4.1; this includes vital signs, laboratory tests, and static information (demographic information and type of admission). Table 2 reports the information types extracted for each patient. The vital signs are highly sampled while laboratory tests are irregular. Along with this information, we also add the SOFA score [25]: this measures the severity of organ disfunction (score between 0 and 4) for six organ systems: respiration, coagulation, liver, cardiovascular, neurological, and renal. The higher the value, the higher the severity of the dysfunction for that patient. The overall severity is calculated by adding all the individual scores of each organ system. At each time point, or within a short time interval[3], we capture the following information: static information, vital signs, laboratory tests and individual and overall SOFA score. As there are many missing observations which may impact the calculation of the SOFA severity score, the SOFA score at any point in time is calculated based on the latest six hours observations.

4.3 Experiment Setting

In our experiments, patient mortality is predicted for two time intervals: (a) after 36 h from admission at ICU, based on the first 24 h of data, and (b) after 72 h from admission at ICU, based on the first 48 h of data. We intentionally consider a time gap between the observation data and the outcome being predicted. This is so that the medical practitioners would have enough time to consider the output of the system and administer a treatment plan. Both the considered time interval, and the gap between observed data and predicted outcome are commonly used in the relevant prior works is counted which is common in literature because it allows enough time for the medical practitioners to take proper decision such as intervene or other treatment parameter settings: specifically, Legman et al. [15]

[3] We consider one hour as per interval because the clinical events that occur in close temporal proximity often have a stronger relationship than those far away, at least in the ICU context.

Table 2. Information about a patient captured in the Temporal Tree.

Static information	Admission type, gender, age
Vital signs	SpO2, Arterial PaCO2, Arterial pH, Heart Rate, Arterial Blood Pressure Systolic, Arterial Blood Pressure Diastolic, Respiratory Rate, GCS - Eye Opening, Temperature Celsius, Inspired O2 Fraction, GCS - Verbal Response, GCS - Motor Response, Anion Gap, Prothrombin Time
Laboratory tests	Bicarbonate, Bilirubin - Total, Calcium - Total, Chloride, Creatinine, Glucose, Potassium, Sodium, Urea Nitrogen, Hematocrit, White Blood Cells, Hemoglobin, Magnesium, INR(PT), Phosphate, pH, Lactate, Platelet Count
SOFA score	respiration, coagulation, liver, cardiovascular, cns, renal

have considered the 24 h setting, Harutyunyan et al. [11] have considered the 48 h setting, while Suresh et al. [24] have considered both settings.

4.4 Evaluation Measure

We use Area Under the Curve (AUC) for measuring the effectiveness of the studied methods for ICU mortality prediction. AUC has been consistently used as the target evaluation measure in prior work that addressed this prediction task [1,11,16,17,20,23,24]. Classifiers are trained and validated using 5-fold cross validation on the training data, and then evaluated on the withhold testing data for the purpose of model comparison and effectiveness. This process is repeated 10 times and effectiveness is averaged to weed out bias due to the random partition of the training data for validation purposes into 5-folds.

The MIMIC-III dataset contains mortality information for each patient including date and time of death, if death occurred. This information is used as ground truth to evaluate the classifiers.

4.5 Baselines and State-of-art Methods

The problem of ICU mortality prediction for a given patient is cast into the problem of assigning a binary label (died, not died) to a patient, given the patient's EHR data as input. To investigate the effectiveness of the representation method proposed in this paper, the following baseline as well as state-of-art classification methods are implemented.

- *Logistic Regression (LR)* is a popular method used in numerous previous work that has considered the problem of ICU mortality prediction [11,15,16].
- *Support Vector Machines (SVM)* have also been often applied to the ICU mortality prediction problem, e.g., [9].
- *Random Forest (RF)* has been shown to provide the highest mortality prediction accuracy among a number of other alternative methods when evaluated

for predicting the six-month mortality in a population of elderly Medicare beneficiaries [17].
- *Gradient-Boosted Trees (GBDT)* produce an ensemble of weak prediction models, typically decision trees. Darabi et al. [8] have shown that GBDT provide high effectiveness for the task of predicting the 30-day mortality risk after admission to ICU. This however was demonstrated on a small dataset only. We further note that their method, like ours, relied on the use of medical embeddings, but in their case this was so as to reduce the dimensionality of the data.
- *Gaussian Naive Base (GaussianNB)* is a simple statistical classifier technique based on the Bayes Theorem, and thus forms a naive baseline.
- *Extreme Gradient Boosting (xgbGBT)* is a specialisation of the Gradient Boosting method which uses a more regularised model formalization for controlling over-fitting, often delivering better classification performance.
- *K-Nearest Neighbourhood (KNN)* exploits proximity among the representation of data items for classification. In the situation at hand, it assumes that similar patients are likely represented by embeddings that are close to each other. It is clear then that the effectiveness of the KNN method depends on the fidelity of the representation in terms of maintaining the similarity between patients. In our implementation, cosine similarity is used to determine the similarity between patients embeddings. The value of k is determined from training data. Specifically, 10-fold cross validation on the training data is used to tune the value of k with respect to AUC; k is varied in the range $[0, 100]$ with step 2.
- *Aczon2017:* Aczon et al. [1] consider the task of predicting the mortality risk for paediatric critical care patient. To that aim, they construct a model architecture comprised of three LSTMs. We replicate this approach by considering the patient situation vectors as the input for the LSTM architecture.
- *Suresh2018:* Suresh et al. [24] propose a method that comprises two steps: (1) Learn relevant patient subgroups in an unsupervised manner. They use a LSTM autoencoder to produce dense representations and then use a Gaussian Mixture Model (GMM) to identify the patient group. (2) Use multitask learning for each separate subgroup. In our experiments, we adapt the same GMM method to cluster the patient into three groups. To this aim, patients embeddings generated by the methods described in Sect. 3.4 are used as input for clustering. Then the same single task and multi-task model is used for each group. Effectiveness is measured across subgroups and averaged into a single value.

In the case of *LR*, *SVM*, *RF*, *GBDT*, *GaussianNB*, *xgbGBT* and *KNN*, patient embeddings are considered as the input for training and testing the model. Conversely, for *Aczon2017* and *Suresh2018*, patient situation embeddings are considered as the input for training and testing the model.

5 Empirical Results

5.1 Effectiveness of Aggregate Features

We consider whether the use of aggregate features as extracted from Temporal Tree improve the effectiveness of classifiers on the task of ICU mortality prediction. To this aim, we consider the results obtained (1) When considering the patient as a sequence of clinical events, i.e. a BOW approach. The results obtained for this setup are reported in Table 3). (2) When preserving patients temporal information in terms of patients situations, and thus considering patients as sequences of patients situations. The results obtained for this setup are reported in Table 4).

Aggregate features improve ICU mortality prediction effectiveness (AUC) when using a BOW-based approach of on average 7.814% (for 24 h) and 4.889% (for 48 h), compared to when aggregate features are not considered. This occurs across all considered classifiers, and the highest AUC is obtained when using the *KNN* approach, for both time intervals.

When analysing the effect preserving temporal information has on the results, we find that aggregate features provide between 3.130% (for 24 h) and 2.562% (for 48 h) average improvements in terms of AUC. Moreover, all classifiers obtain higher prediction effectiveness when aggregate features are used, except when considering KNN for predicting mortality in the 48 h setting. In both time settings, the best effectiveness is provided by the SVM classifier when considering aggregate features.

Among all classifiers, KNN and SVM, when used on aggregate features, provide the highest ICU patient mortality predictions, independently of the time interval considered. We stress the fact that the effectiveness of the KNN method is highly dependent on the representation chosen for the patients. Thus, the fact that KNN is among the best classifiers, if not the best, when using the proposed representation techniques, speaks in favour of the Temporal Tree approach coupled with the embedding method, and specifically of their ability to effectively model patient similarity.

Table 3. ICU mortality prediction effectiveness (AUC) across methods and for the bag-of-word approach. Results are reported distinguished between the 24hr and 48hr setup. Standard deviation is provided in brackets and represents the variation obtained across different rounds of tuning of the learnt classifier.

Classifers	24hr_baseline	24hr_aggregate	% Improvement	48hr_baseline	48hr_aggregate	% Improvement
LR	0.836 ± 0.002	0.862 ± 0.001	3.110	0.818 ± 0.002	0.838 ± 0.002	2.445
SVM	0.820 ± 0.004	0.851 ± 0.003	3.780	0.814 ± 0.004	0.816 ± 0.004	0.246
RF	0.724 ± 0.015	0.773 ± 0.009	6.768	0.716 ± 0.01	0.727 ± 0.009	1.536
GBDT	0.788 ± 0.008	0.849 ± 0.003	7.741	0.775 ± 0.006	0.818 ± 0.005	5.548
GaussianNB	0.474 ± 0.006	0.613 ± 0.003	29.325	0.513 ± 0.002	0.622 ± 0.004	21.248
XgbGBT	0.849 ± 0.003	0.867 ± 0.002	2.120	0.810 ± 0.004	0.833 ± 0.002	2.840
KNN*	0.864 ± 0.000	**0.880 ± 0.000**	1.852	0.832 ± 0.000	**0.835 ± 0.000**	0.361
Average			**7.814**	Average		**4.889**

Table 4. ICU mortality prediction effectiveness (AUC) across methods and for the approach that preserves temporal information. Results are reported distinguished between the 24hr and 48hr setup. Standard deviation is provided in brackets and represents the variation obtained across different rounds of tuning of the learnt classifier.

Classifers	24hr_baseline	24hr_aggregate	% Improvement	48hr_baseline	48hr_aggregate	% Improvement
LR	0.832 ± 0.000	0.870 ± 0.001	4.567	0.795 ± 0.001	0.831 ± 0.001	4.528
SVM	0.847 ± 0.001	**0.875 ± 0.001**	3.306	0.821 ± 0.001	**0.844 ± 0.001**	2.801
RF	0.760 ± 0.005	0.781 ± 0.009	2.763	0.743 ± 0.008	0.745 ± 0.012	0.269
GBDT	0.842 ± 0.002	0.867 ± 0.002	2.969	0.823 ± 0.004	0.829 ± 0.003	0.729
GaussianNB	0.558 ± 0.002	0.582 ± 0.002	4.301	0.563 ± 0.003	0.592 ± 0.002	5.151
XgbGBT	0.853 ± 0.002	0.874 ± 0.002	2.462	0.831 ± 0.002	0.838 ± 0.002	0.842
KNN*	0.852 ± 0.000	0.862 ± 0.000	1.174	0.820 ± 0.000	0.810 ± 0.000	−1.220
Aczon2017	0.829 ± 0.024	0.858 ± 0.014	3.498	0.757 ± 0.06	0.813 ± 0.019	7.398
Suresh2018-Single	0.831 ± 0.000	0.839 ± 0.000	1.003	0.769 ± 0.00	0.784 ± 0.000	1.950
Suresh2018-Multitask	0.832 ± 0.000	0.842 ± 0.000	1.161	0.781 ± 0.000	0.771 ± 0.000	−1.323
Average			**2.720**	Average		**2.113**

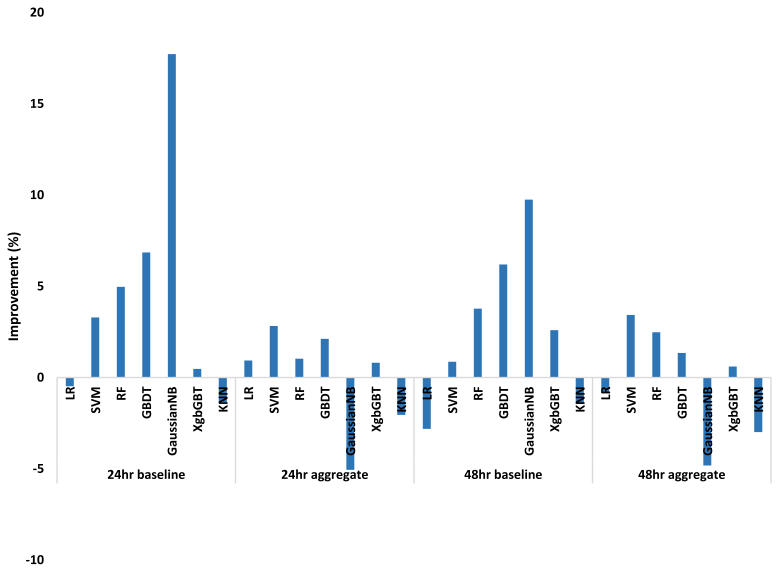

Fig. 3. Performance Improvement of Temporal Information over Bag-Of-Word Approach.

Finally, the empirical results also clearly suggest that ICU mortality prediction is easier when performed on the first 24 h of data obtained after admission, compared to the 48 h setting. We note a that similar result was found in previous studies, e.g., [24].

5.2 Impact of Modelling Temporal Information

We now consider whether modelling temporal information provide an edge over the more simplistic bag-of-word approach. Figure 3 shows the percentage improvement while considering patient's temporal information (i.e. patient situations) over bag-of-word indicated by y-axis value. SVM, RF, GBDT, and

XgbGBT always provide better performance when preserving temporal information; the results however exhibits the opposite trend when KNN is considered, while LR and GaussianNB provide mixed results. Overall, we found that, in most cases, preserving temporal information leads to better mortality prediction accuracy.

6 Conclusion and Future Work

In this paper we have investigated the problem of ICU mortality prediction from EHR data: this is a challenging but important prediction task as improvements in prediction accuracy translate into better clinical decision support and thus likely better healthcare delivery. To provide an effective mean for accurately predicting patient mortality, in this paper we propose to represent patients EHR data using the Temporal Tree technique, a recently introduced method for representing compounded EHR data. This method is used to generate patients and patients situation embeddings, which are then used as the input to a suite of common and state-of-the-art classifiers for the ICU mortality prediction task. Our extensive empirical results on a dataset of real ICU EHR data demonstrate that compound information generated by Temporal Tree is useful for producing discriminative representations, which in turn improve the mortality prediction accuracy of the considered classification methods. In addition, the results also demonstrate that preserving temporal information leads to further gains in effectiveness in most cases and settings.

References

1. Aczon, M., et al.: Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. Stat 1050, 23 (2017)
2. Bajor, J.M., Mesa, D.A., Osterman, T.J., Lasko, T.A.: Embedding complexity in the data representation instead of in the model: a case study using heterogeneous medical data. arXiv preprint arXiv:1802.04233 (2018)
3. Batal, I., Valizadegan, H., Cooper, G.F., Hauskrecht, M.: A temporal pattern mining approach for classifying electronic health record data. ACM Trans. Intell. Syst. Technol. (TIST) **4**(4), 63 (2013)
4. Breslow, M.J., Badawi, O.: Severity scoring in the critically ill: part 1–interpretation and accuracy of outcome prediction scoring systems. Chest **141**(1), 245–252 (2012)
5. Chen, W., Long, G., Yao, L., Sheng, Q.Z.: AMRNN: attended multi-task recurrent neural networks for dynamic illness severity prediction. World Wide Web **23**(5), 2753–2770 (2019). https://doi.org/10.1007/s11280-019-00720-x
6. Choi, E., et al.: Multi-layer representation learning for medical concepts. In: Proceedings of the 22nd ACM SIGKDD, pp. 1495–1504. ACM (2016)
7. Coopersmith, C.M., et al.: A comparison of critical care research funding and the financial burden of critical illness in the united states. Crit. Care Med. **40**(4), 1072–1079 (2012)
8. Darabi, H.R., Tsinis, D., Zecchini, K., Whitcomb, W.F., Liss, A.: Forecasting mortality risk for patients admitted to intensive care units using machine learning. Procedia Comput. Sci. **140**, 306–313 (2018)

9. Ghassemi, M., et al.: Unfolding physiological state: mortality modelling in intensive care units. In: Proceedings of the 20th ACM SIGKDD, pp. 75–84. ACM (2014)
10. Glicksberg, B.S., et al.: Automated disease cohort selection using word embeddings from electronic health records. In: PSB, pp. 145–156. World Scientific (2018)
11. Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. Sci. Data **6**(1), 96 (2019)
12. Johnson, A.E., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016). https://doi.org/10.1038/sdata.2016.35
13. Knaus, W.A., et al.: The apache iii prognostic system: risk prediction of hospital mortality for critically iii hospitalized adults. Chest **100**(6), 1619–1636 (1991)
14. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
15. Lehman, L.W., Saeed, M., Long, W., Lee, J., Mark, R.: Risk stratification of ICU patients using topic models inferred from unstructured progress notes. In: AMIA Annual Symposium Proceedings, vol. 2012, p. 505. American Medical Informatics Association (2012)
16. Luo, Y., Xin, Y., Joshi, R., Celi, L., Szolovits, P.: Predicting ICU mortality risk by grouping temporal trends from a multivariate panel of physiologic measurements. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
17. Makar, M., Ghassemi, M., Cutler, D.M., Obermeyer, Z.: Short-term mortality prediction for elderly patients using medicare claims data. Int. J. Mach. Learn. Comput. **5**(3), 192 (2015)
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
19. Moreno, R.P., et al.: Saps 3-from evaluation of the patient to evaluation of the intensive care unit. part 2: development of a prognostic model for hospital mortality at ICU admission. Intensive Care Med. **31**(10), 1345–1355 (2005)
20. Nori, N., Kashima, H., Yamashita, K., Ikai, H., Imanaka, Y.: Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care. In: Proceedings of the 21th ACM SIGKDD, pp. 855–864. ACM (2015)
21. Pokharel, S., Zuccon, G., Li, X., Utomo, C.P., Li, Y.: Temporal tree representation for similarity computation between medical patients. Artif. Intell. Med. **108**, 101900 (2020)
22. Shervashidze, N., Schweitzer, P., Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. J. Mach. Learn. Res. **12**(Sep), 2539–2561 (2011)
23. Shi, Z., Chen, W., Liang, S., Zuo, W., Yue, L., Wang, S.: Deep interpretable mortality model for intensive care unit risk prediction. In: Li, J., Wang, S., Qin, S., Li, X., Wang, S. (eds.) ADMA 2019. LNCS (LNAI), vol. 11888, pp. 617–631. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35231-8_45
24. Suresh, H., Gong, J.J., Guttag, J.V.: Learning tasks for multitask learning: heterogenous patient populations in the ICU. In: Proceedings of the 24th ACM SIGKDD, pp. 802–810. ACM (2018)
25. Vincent, J.L., et al.: The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive Care Med. **22**(7), 707–710 (1996)
26. Xu, Y., Zhang, Z., Lu, G., Yang, J.: Approximately symmetrical face images for image preprocessing in face recognition and sparse representation based classification. Pattern Recogn. **54**, 68–82 (2016)
27. Zhang, J., Kowsari, K., Harrison, J.H., Lobo, J.M., Barnes, L.E.: Patient2vec: a personalized interpretable deep representation of the longitudinal electronic health record. IEEE Access **6**, 65333–65346 (2018)