Multimodal Deep Learning for Early Length of Stay Prediction using Patient Similarity Embeddings

Arnon Vandenberghe[1†], Lyse Naomi Wamba Momo[1†][0000−0002−9019−0236], Vincent Scheltjens[1][0000−0002−6382−5750], and Bart De Moor[1][0000−0002−1154−5028]

KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{firstname.lastname}@kuleuven.be

Abstract. Background and objectives: Patient outcome prediction is a key challenge in view of logistics optimization and cost reduction in clinical contexts. Specifically in the Intensive Care Unit (ICU), where accurately estimating remaining Length of Stay (LoS) is highly valuable in regards to resource allocation and the identification of high-risk patients. In this study, a multimodal deep learning model is presented for early ICU-LoS prediction based on the first 24 hours of data after admission into ICU, extracted from the MIMIC-IV database.
Methodology: The multimodality is three-faceted, comprising a multivariate time-series, static demographic data and weighted graph-based similarity embeddings using the Node2Vec algorithm. Gated Recurrent Units (GRU) are used to establish baseline performance and are subsequently exploited with early and late fusion techniques.
Results and Conclusion: We show that augmentation with demographic data improves a temporal-only GRU baseline from a Mean Absolute Error (MAE) of 1.87 to 1.79 days, both with early and late fusion techniques. Furthermore, representing a patient's hospital transfers prior to 24 hours after ICU-admission as a weighted graph structure capturing patient similarities improves the prediction accuracy when fused on the time-series, to a MAE of 1.76 days. Though the latter representation does not always show significant superior performance over a vectorized transfer representation, it proves to be conceptually fitting and an intuitive way to represent and learn patient transfers. An Expected Gradients (EG) analysis for feature importance allows for additional insights into the model predictions.

Keywords: Length of Stay · Gated Recurrent Units · Graph Embeddings · Feature Explainability

† Equal contribution

1 Introduction

Predictive analytics for patient clinical outcomes like Length of Stay (LoS) in Intensive Care Units (ICU) has proven to be of high importance given the critical state of patients being admitted and the need to adequately distribute resources in proportion to the patients' needs. LoS, defined as the duration between admission and discharge of an inpatient in a single care episode [25], is an important tool used both for hospital quality evaluation and clinical optimization [16]. Modeling LoS has multiple benefits, including the early identification of patients at risk of extended stays (usually accompanied with high costs), reassuring patients and family members, improving patient satisfaction and optimally allocating resources (beds, staff, equipment, etc.) [11,12].

In clinical institutions, the digitization of data has led to the adoption of Electronic Health Records (EHRs). EHRs are an electronic version of a patient's medical history and include a myriad of information [6]. These sources of information appear in various formats, ranging from text (clinical reports), static data (demographics, admission, transfers), time-varying tabular data (vital signs, labs, medications, etc.), to images (scans, etc.) and more.

Given a single patient admission to ICU, termed a stay, multiple parameters are monitored at different time points. While demographics are charted only once at admission, labs and vital signs are recorded regularly (especially in ICU) and patient transfers are documented upon their occurrence. A means of combining all of these patient parameters with irregular charting frequencies for clinical analysis can be achieved by building a multimodal architecture that concatenates rich representations of the different data types [21].

In the literature, text data from clinical notes appears as one of the most exploited modalities for improving on baseline models with static demographic and time-varying data, resulting in an improved performance [2,15,30]. Closest to our work is the end-to-end hybrid GNN-LSTM model constructed by [23] for LoS and mortality predictions where time-varying information is processed by Long Short-Term Memory (LSTM) networks, static demographic data by a Fully Connected Network (FCN) and diagnoses shared among patients using a Graph Neural Network (GNN).

Though we do not use graphs in an end-to-end modelling strategy, the novelty in this work lies in the exploitation of patient transfer data, which has not yet been previously used, to learn the underlying similarities between patients. More precisely, patient transfers from one care unit to another are modeled as a graph to obtain a compact and rich representation of the underlying similarities. These representations are then used to augment static demographic and time-varying data for LoS predictions in terms of fractional days (as opposed to a more often performed classification in the literature). The use of graphs here enables us to encode not only the underlying connection between patients visiting the same care units, but also to implicitly avoid missing data imputation (as opposed to when transfers are stored in tabular form) for patients who did not have any transfer (56% of patients in the selected cohort).

The remainder of this work is organized as follows; Section 2 outlines related work and Section 3 gives a thorough explanation of the model architecture. Section 4, covers the data and preprocessing steps, after which the results are discussed in Section 5. Finally, Section 6 concludes the work and suggests future improvements and research directions.

2 Related work

2.1 Multimodal learning of EHRs for LoS prediction

Given its importance, LoS has gained a lot of attention over the years which has prompted many researchers to develop prediction algorithms both in unimodal and/or multimodal settings. In [30], binary classification of early LoS prediction was studied by applying both LSTM- and CNN-based algorithms on demographics, time-varying data and clinical notes. Embeddings from these data sources were fused in a late-fusion scheme to output predictions. Similarly, Convolutional Neural Network (CNN) embeddings extracted from irregularly charted clinical notes were fused with LSTM embeddings of hourly sampled time-varying information for continuous classification of remaining LoS in [15]. Extracted medical entities using the med7 model on clinical notes were combined in a late-fusion setting with Gated Recurrent Unit (GRU) embeddings from time-varying data for binary classification of early LoS in [2].

2.2 Graph based similarity embeddings of EHRs

Graph representation learning is still a young and emerging tool with an increasing number of use cases in research, including in the medical domain [24], to encode similarities that exist between entities. Some early applications of graph learning on EHRs involves knowledge representation in the form of a graph [5], message passing for missing data imputation [20] and the modeling of disease transition processes by building subgraphs for multiple patient visits and diseases [18]. In [23], an end-to-end hybrid LSTM-GNN model was used to learn patient similarities based on diagnoses and predict patient outcomes including LoS.

2.3 Model interpretability

Clinical validation of health-related models is essential and requires the comprehension of the model's reasoning by medical practitioners. Extracting what models examine and how this leads to a prediction is a difficult task with Deep Learning (DL) architectures and is one of the biggest limitations of such models [8]. In [23], attention weights assigned to edges (patient links) were used to assess an LSTM-GAT (LSTM-Graph Attention Network) model comprehension. Similarly, learned attention regions were visualized in [29] to show the time points that the model paid more or less attention to for decompensation prediction.

Feature attributions were computed using the integrated gradients method [26] in [22]. In [1], two-dimensional plots of the mortality risk space for the different ICU domains were obtained using t-SNE (t-distributed Stochastic Neighbor Embedding) [19].

3 Methods

The aim for the model is to accurately predict the expected LoS for a given patient shortly after ICU admission. To that end, the model should extract trends in the temporal input data. GRUs [1]are used to establish a baseline on time-series data, which is later enriched with demographic data and patient transfers in a multimodal setting.

Formally, the task at hand is to predict LoS based on the first 24 hours of patient data post ICU admission. As such, we use the time-series $\boldsymbol{x}^i_{1:24}$ [2]in combination with demographic data $\boldsymbol{d}^i$ and the graph-based similarity embeddings $\boldsymbol{e}^i$, where i denotes the i^{th} patient. From this, we intend to yield predictions $\hat{y}$ which accurately approximate the true LoS, y, calculated as the difference between T_D and T_A denoting respectively the time at discharge and at admission.

3.1 Model Architecture

Gated Recurrent Unit (GRU) At the core of the herein proposed architecture is a GRU, used for both the baseline and as one of the modalities in the multimodal approach. GRU is a subclass of the broader set of Recurrent Neural Networks (RNN) and offers reduced computational complexity as a result of the sole two gates it is comprised of. These are the reset and forget gates, denoted r_t and z_t respectively [4].

The equations governing the GRU cell are given in Eq. (1). Thus, for each discrete time step t in the input sequence, the following is computed:

$$\begin{aligned} r_t &= \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}) \\ z_t &= \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}) \\ n_t &= tanh(W_{in}x_t + b_{in} + r_t * (W_{hn}h_{(t-1)} + b_{hn})) \\ h_t &= (1 - z_t) * n_t + z_t * h_{(t-1)}, \end{aligned} \tag{1}$$

where $h_t, h_{(t-1)}$ are the hidden states at time $t, t-1$ and x_t the corresponding input at time t. Furthermore, n_t represents the new gates and $*$ corresponds to the Hadamard product. t corresponds to one hour in this work.

[1] Performance of LSTM networks is also evaluated and shown in Table 4. As LSTM does not strictly outperform GRU, subsequent analysis is performed with the less complex GRU architecture.

[2] The time-series input matrix $\boldsymbol{x}^i_{1:24}$ is augmented with a matrix of identical size, termed decay mask, that indicates for each feature in $\boldsymbol{x}^i_{1:24}$ the time since the last recording similar to [22,23,17]

Fully Connected Network (FCN) The output from the GRU network, i.e. the hidden state h_t, is passed through the FCN to yield a single LoS prediction. The FCN is made non-linear using the ReLU activation function and defined as:

$$\hat{y}_t = ReLU\left(W_{y_t} h_t + b_{y_t}\right), \tag{2}$$

where $\hat{y}_t$ is the prediction for LoS at time step t and $ReLU(x) = max(0, x)$. The latter enforces a strict constraint on the outcome to return only positive values.

Graph representation learning In order to learn representations that capture similarities among patients with a similar care trajectory prior to the first 24 hours post ICU admission, we model the patient transfers as a graph.

In the process, a bipartite graph is created for which the first set of nodes corresponds to the patients and the second to the different care units for which visits were observed within the considered time frame. An additional node was created for those patients that had no recorded transfers within that period to incorporate implicit missingness. The edges between the nodes denote the visit of a patient to a certain unit, and the weight assigned to an edge reflects the duration of the stay within that care unit. In Figure 1, a subgraph for a sample of 10 patients is shown. Patient with ID 37766114 is seen to have visited the care units PACU, Cardiology Surgery Intermediate and Cardiac Vascular ICU in-between admission to the hospital and 1 day post ICU admission, while patient with ID 31311981 was admitted to Cardiac Vascular ICU and had no further transfer prior to the 1 day cutoff. The different edge colors in the graph make a distinction between ongoing stays at the 24 hour cutoff and stays terminated prior to that.

Patient similarity embeddings Based on the constructed graph, representations are learned that capture similarities between patients in terms of care trajectory. For this purpose, the Node2Vec algorithm [10] is used, which in essence is a combination of random walks and skip-gram Word2Vec algorithm. By performing random walks through the graph, transition probabilities of moving from one node to the other are constructed and fed to the skip-gram algorithm for lower-dimensional embedding of the graph [10]. These embeddings are the third data modality considered in the overall architecture. In Figure 2, the t-SNE 2-D projection is shown where each point represents a patient. The color indicates the care unit at which a patient is residing at the 1 day post ICU admission time stamp. Figure 2 shows how the embeddings capture similarities between patients resulting in clearly defined clusters. Different clusters of the same color suggest the similarity between patients in terms of their trajectory prior to the cutoff time. As an example, one yellow cluster could signify an admission to the emergency department and subsequently a transfer to CVICU, while another yellow group could entail that those patients visited CCU prior to CVICU.

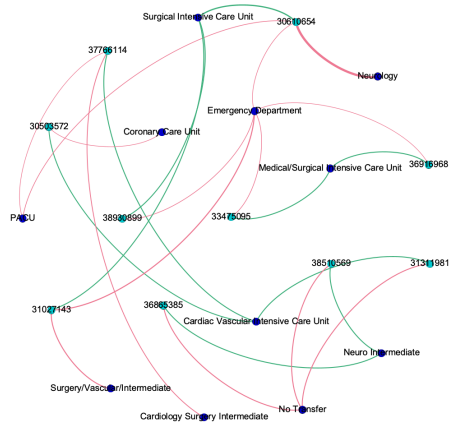

Fig. 1. Subgraph of a sample of 10 patients (green nodes) with corresponding transfers to different care units (blue nodes) prior to 24 hours post ICU admission. The color of the edge depicts a terminated stay (red) or ongoing stay (green) with the thickness representative of the overall duration of the stay at the given unit.

3.2 Models

Baseline The baseline model is established by only feeding the temporal data to the GRU and subsequently to the FCN components of the architecture.

Multimodal models In order to improve upon the baseline, the two additional data modalities are added, i.e., (*i*) the flat demographic data and (*ii*) the patient transfers data as learned similarity embeddings one after the other and then both.

As an intuitive example for including hospital tranfers, consider two patients that follow a similar care trajectory in terms of both the care units and time spent at each of the units. Our intuition here is that, when factoring in demographic similarities, the LoS for these two patients is more likely to be closer to one another than for patients with a dissimilar care trajectory and different demographic characteristics.

In the overall model architecture, different modalities are joined together through either early-fusion or late-fusion as shown in Figures 3 and 4.

Early-fusion In the early-fusion approach (Figure 3), the temporal, demographic and embedded or vectorized patient transfer data are joined together before passing through the main components of the model architecture. As the constructed graph can not be directly joined with the other sources of data, the lower dimensional vector representations are first learned and extracted. These are then fused with the temporal and/or demographic data.

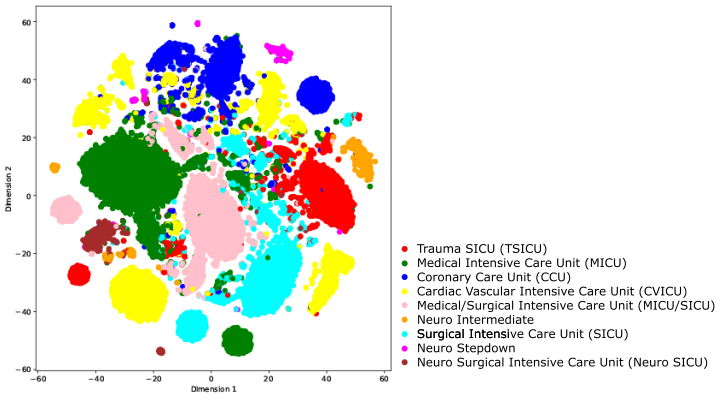

Fig. 2. t-SNE plot of the patient transfers graph embedding. Each dot represents one patient. Each color is associated with one of nine careunits for which there are ongoing stays registered at the 1 day timestamp with respect to ICU admission time.

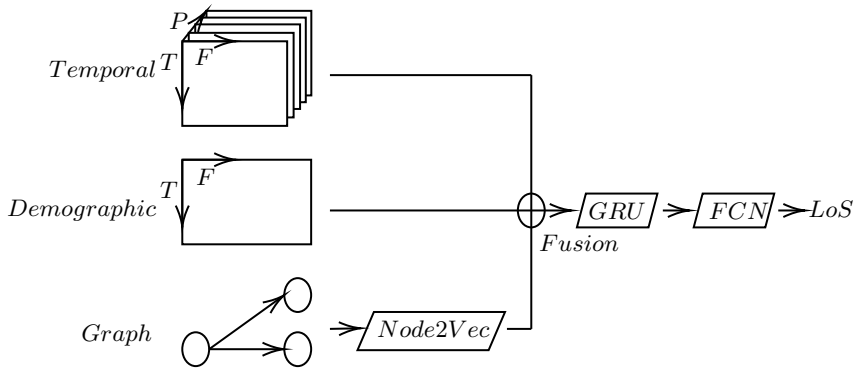

Fig. 3. Model architecture with early data fusion on the different modalities with indexing keys P for patients, T for time steps and F for features.

Late-fusion Here, each of the different data modalities is first processed by one of the model's sub-components, the representations of which are concatenated and passed through a FCN in order to obtain the final model outcome as seen in Figure 4.

Model optimization All models are optimized in two stages. First, a grid search is performed to determine the number of hidden layers L and the number of hidden units N in each layer. Three runs are considered while the learning rate η, the batch size m, the weight decay wd and the dropout rate r are randomly permuted in a confined interval in each of those runs (Table 2). Afterwards, population-based training (PBT) [13] is performed to fine-tune the choice of the

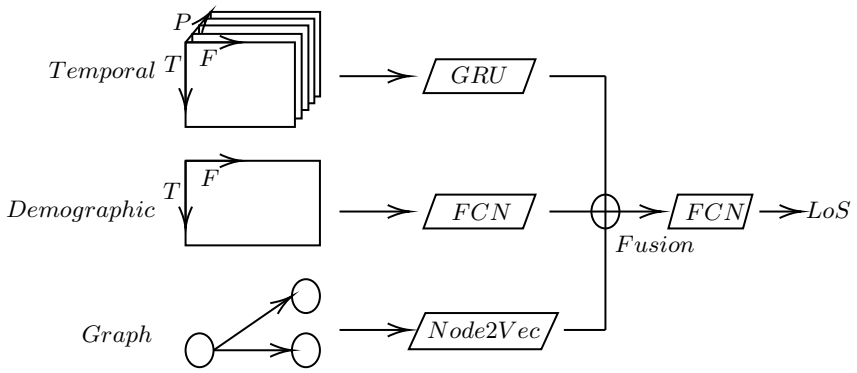

Fig. 4. Model architecture with late data fusion on the different modalities with indexing keys P for patients, T for time steps and F for features.

remaining hyperparameters, i.e., η, m, wd and r, within the intervals specified in Table 2. Best hyperparameters obtained are listed in Table 1.

Table 1. Optimal model hyperparameters obtained with grid search and a subsequent population-based training. L: Number of layers, N: number of hidden units, η: learning rate, m: batch size, wd: weight decay, r: dropout.

Model	L	N	η	m	wd	r
TS-GRU with decay mask	1	128	0.004	1024	5.20e-6	0.57
TS-GRU + demographic + transfers, early-fused	2	256	0.0004	2048	9.89e-6	0.10

Table 2. Hyperparameter search space where ";" signifies a choice and "," an interval.

Parameter	Grid search	Population-based training
L	[1; 2]	/
N	[32; 64; 128; 256]	/
η	[1e-5, 1e-2)]	[1e-5, 1e-2]
m	[32; 64; 128; 256; 512; 1024; 2048]	[32; 64; 128; 256; 512; 1024; 2048]
wd	[1e-8, 1e-2]	[1e-8, 1e-2]
r	[0, 0.8]	[0, 0.8]

4 Data

The data used in this study is extracted from the Medical Information Mart for Intensive Care IV (MIMIC-IV version 1.0) database [9,14]. All ICU stays of

adult patients (>18 years old) with LoS > 1 day with at least one documented recording during that stay are selected. This results in 60,494 distinct ICU stays from 44,246 patients, recorded in 55,535 separate hospital admissions described in Table 3. Multiple stays within the same hospital admission, i.e., ICU re-admissions, are kept. The target is clipped at 14 days (Figure 5) to reduce the effect of skewness as done in [27,28]. This cutoff excludes less than 5% of the stays from the initial cohort.

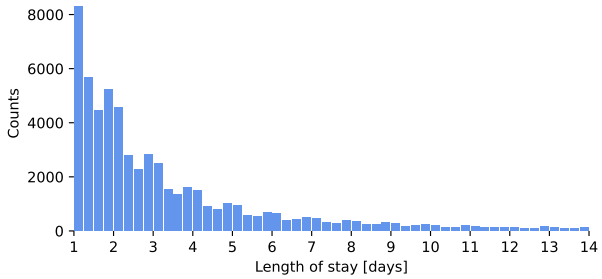

Fig. 5. Length of stay distribution of the cohort grouped in bins of 0.25 days. The right-ended tail of the distribution for stays longer than 14 days is omitted.

Table 3. Cohort overview

Number of stays	60,494
Train	42,345
Validation	9074
Test	9075
Gender (% female)	43.47
Age (mean)	65.10
Mean LoS	4.21
Median LoS	2.41
Number of features	96
Temporal	84
Demographic	12
Number of visited care units (mean per stay)	2.80
Number of patients having at least one transfer	19,616

For every ICU stay in the cohort, we extract all documented features from the *hosp.labevents* and *icu.chartevents* tables in a time window of -1 day to +1 day with respect to ICU admission time. This results in 1427 distinct features, which

is further reduced by removing those variables that are observed in less than 25% of stays and removing those that have an average measurement frequency of less than 2 and 3 for *hosp.labevents* and *icu.chartevents* respectively. 126 features remain, out of which 42 are removed as they are binary indicators or they are found to be duplicates. Forward-filling of missing data is performed up to the first observation post ICU admission time, similar to what is done in [22]. Inspired from the work in [22,3], a mask for every temporal feature is added as a decaying indicator of the time that has passed since the last recording. The decay curve corresponds to 0.75^j, where j is the number of hours since that last recording.

As demographic patient data, we extract 12 features from the *core.patients*, *core.admissions* and *icu.chartevents* tables. This is augmented with 30 features from the time-series data, considered to be not considerably varying over time. Furthermore, 3 features are extracted for the transfers data, which is in turn used for the corresponding graph construction and for building baseline vectorized transfer representations to compare the embeddings against. These features are the care unit, the duration in that care unit and a binary indicator for an ongoing stay at the 24 hour post ICU-admission time stamp. An overview of the temporal and demographic features is included in the GitHub repository [3].

The cohort is split into a train, validation and test set according to a 70/15/15 split. During model training, the MSE loss on the validation set is monitored and early stopping with a patience of 5 epochs is implemented to avoid overfitting on the training data. All metrics reported in the remainder of this work were evidently reported on the test data.

5 Results

5.1 Evaluation metrics

The metrics used for evaluating the model's performance are the coefficient of determination (R^2), the mean squared error (MSE) and the mean absolute error (MAE), defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the target LoS, $\hat{y}_i$ is the predicted LoS and $\bar{y}$ is the mean of the observed data.

5.2 Baselines

The performance of the baseline model is reported in Table 4. The inclusion of the decay masking is seen to significantly improve the performance, though it increases the model complexity by doubling the number of input features. As a comparison and validation of the baseline GRU model, two naive models are included that always predict either the mean or median from the training set.

[3] link to extracted features

Table 4. Performance of the baseline models. All metrics are reported as the average over 10 runs. TS-GRU/TS-LSTM: GRU/LSTM with time-series only. The error margin is the standard deviation among these iterations. TS-GRU and TS-LSTM always include decay masks.

Model	R^2	MSE	MAE
Mean	0.00	11.46	2.47
Median	-0.16	13.35	2.18
TS-LSTM without decay mask	0.29 ± 0.00	8.10 ± 0.05	1.95 ± 0.04
TS-LSTM	0.34 ± 0.01	7.54 ± 0.08	1.89 ± 0.02
TS-GRU without decay mask	0.30 ± 0.00	8.03 ± 0.05	1.94 ± 0.02
TS-GRU	0.34 ± 0.00	7.56 ± 0.05	1.87 ± 0.02

5.3 Multimodal models

Augmentations of the time-series GRU model with demographic data and/or transfer embeddings both in early and late fusion settings are evaluated.

From Table 5 and the t-test results from appendix A, all model extensions are seen to significantly outperform the TS-GRU baseline for all considered metrics on a 1% significance level. The inclusion of demographic data in both early and late fusion approaches yields an improvement in terms of MAE from 1.87 for the TS-GRU with decay masking to 1.79. The late fusion architecture is deemed slightly better than the early fusion one, reporting a lower MSE, while the MAE is not found to be significantly different.

The augmentation of the time-series with solely the transfer embeddings is less pronounced than with the demographic data, but nevertheless improves upon the baseline performance, showing its value towards early LoS prediction. Both fusion approaches for TS-GRU augmented with transfer embeddings are comparable in terms of performance.

To facilitate the interpretation of the increase in predictive power added by the inclusion of the transfer embeddings, we run the same models with an alternative vector representation for patient transfers termed v-transfers in Table 5. This vector representation $\boldsymbol{v}^i$ is a matrix of binary indicators capturing the presence of a patient in a given unit prior to the first 24 hours after ICU admission with the corresponding time spent in that unit. The results in Table 5 show that for the late fusion models, the vectorized representations do not outperform the proposed transfer embeddings as no statistically significant difference is observed. However, the early fusion models leveraging the transfer embeddings, both with and without the demographic data, significantly outperform the counterpart models leveraging the vectorized transfer representations at ($p < 0.01$) on MSE.

The overall best performance is obtained for a multimodal model that incorporates all data types, i.e., the time-series, the demographic data and the patient transfer embeddings, both with early and late fusion. The latter fusion type is favored over the former as it achieves a significantly lower MAE ($p < 0.05$).

Table 5. Performance of the TS-GRU model augmentations with early (EF) or late fused (LF) demographic data and/or transfer embeddings. The best performance for each metric is marked in bold. TS-GRU always include decay mask.

Model	Fusion	R^2	MSE	MAE
TS-GRU without decay mask	/	0.30 ± 0.00	8.03 ± 0.05	1.94 ± 0.02
TS-GRU	/	0.34 ± 0.00	7.56 ± 0.05	1.87 ± 0.02
TS-GRU+demographic	early	0.36 ± 0.00	7.29 ± 0.04	1.79 ± 0.04
	late	0.37 ± 0.01	7.20 ± 0.07	1.80 ± 0.02
TS-GRU+v-transfers	early	0.34 ± 0.04	7.51 ± 0.05	1.83 ± 0.03
	late	0.35 ± 0.00	7.43 ± 0.05	1.85 ± 0.03
TS-GRU+e-transfers	early	0.36 ± 0.00	7.39 ± 0.05	1.83 ± 0.03
	late	0.35 ± 0.01	7.45 ± 0.08	1.83 ± 0.03
TS-GRU+demographic+v-transfers	early	0.37 ± 0.00	7.26 ± 0.05	1.79 ± 0.02
	late	0.37 ± 0.01	7.15 ± 0.08	1.78 ± 0.01
TS-GRU+demographic+e-transfers	early	**0.38 ± 0.00**	**7.14 ± 0.03**	1.79 ± 0.02
	late	0.37 ± 0.01	7.18 ± 0.12	**1.76 ± 0.03**

5.4 Model interpretability

By means of an expected gradients (EG) analysis [7], the importance of each input feature to our model for early LoS prediction can be interpreted. This permits a validation of the results, as it allows for drawing parallels to clinical practice. In Figure 6, the feature importance of the 25 most decisive features out of a total of 137 are displayed as the mean of the absolute expected gradient values when passing a batch of test data through the model. After averaging over each sample, the values are normalized to show relative feature influence. The model used in this phase is the best performing three-faceted multimodal model.

The most influential feature is found to be *null_height*, which indicates the missingness of the patient's height as a binary indicator. Following this are the temporal features *GCS - Verbal Response* and *GCS - Eye Opening*, indicating a patient's consciousness level. These are all features that are documented by a clinician and not continuously through a machine, suggesting the importance of a personal interaction with clinical staff towards predicting early LoS, as the opinion of the clinician is implicitly captured in these features. Other striking features are the *Braden Score*, the binary *ventilated* indicator and the *O2 Flow*, out of which the first one is a risk scale used by clinical staff for prediction of pressure ulcer, while the second and third are important in terms of discharge as they are indicative of the patient's autonomous breathing ability. The top 25 features further include four transfer embedding dimensions, again proving the relevance of hospital transfers and the representation under graph similarity embeddings.

[5] https://github.com/slundberg/shap

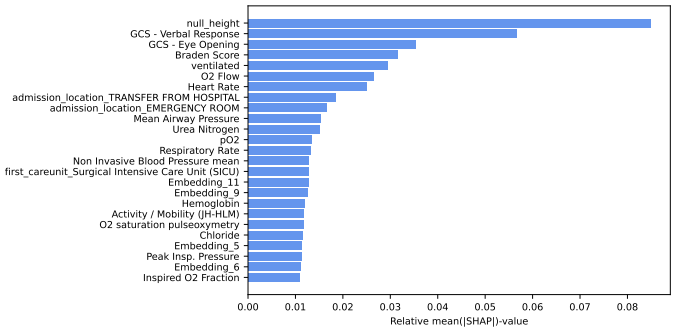

Fig. 6. Expected gradients feature importance using SHAP software package [5]. The values are normalized to show their relative weight.

6 Conclusion and future work

In this work, we showed the added value of including demographic data and patient similarity embeddings for early LoS prediction, on top of the time-series data as addressed in [12]. The three-faceted multimodal model is shown to outperform baseline models. Though the contribution of transfer patient information embedded into a graph was not always significantly superior to its vectorized binary representation, we explored a different means of encoding and learning patient similarities, incorporate additional knowledge (in terms of edge weights) and mitigate the effects of imputation.

This study contributes to existing works clinical analytics for early LoS prediction in ICU. Accurately predicting early LoS is highly valuable regarding resource allocation and identification of high-risk patients, targeting logistics optimization and cost reductions in clinical contexts. In future work, expanding the graph with additional sources of data, including for example medications is envisaged. Regarding the transfer embeddings, taking into account the order of patient transfers as a time-dependent, directed graph is another interesting avenue as is benchmarking the proposed method on other datasets.

Acknowledgements

This work was supported by KU Leuven: Research Fund (projects C16/15/059, C3/19/053, C24/18/022, C3/20/117, C3I-21-00316), Industrial Research Fund (Fellowships 13-0260, IOFm/16/004, IOFm/20/002) and several Leuven Research and Development bilateral industrial projects; Flemish Government Agencies: FWO: EOS Project no G0F6718N (SeLMA), SBO project S005319N, Infrastructure project I013218N, TBM Project T001919N; PhD Grants (SB/1SA1319N, SB/1S93918, SB/1S1319N), EWI: the Flanders AI Research Program VLAIO: CSBO (HBC.2021.0076) Baekeland PhD (HBC.20192204) and Innovation mandate (HBC.2019.2209) European Commission: European Research Council under the European Union's Horizon 2020 research and

innovation programme (ERC Adv. Grant grant agreement No 885682); Other funding: Foundation 'Kom op tegen Kanker', CM (Christelijke Mutualiteit)

References

1. Alves, T., Laender, A., Veloso, A., Ziviani, N.: Dynamic prediction of icu mortality risk using domain adaptation. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 1328–1336. IEEE (2018)
2. Bardak, B., Tan, M.: Improving clinical outcome predictions using convolution over medical entities with multimodal learning. Artificial Intelligence in Medicine **117**, 102112 (2021)
3. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Scientific Reports **8**(1), 6085 (2018). https://doi.org/10.1038/s41598-018-24271-9
4. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. CoRR **abs/1406.1078** (2014), http://arxiv.org/abs/1406.1078
5. Choi, E., Xu, Z., Li, Y., Dusenberry, M.W., Flores, G., Xue, Y., Dai, A.M.: Graph convolutional transformer: Learning the graphical structure of electronic health records. arXiv preprint arXiv:1906.04716 (2019)
6. Critical Data, M.: Secondary analysis of electronic health records. Springer Nature (2016)
7. Erion, G., Janizek, J.D., Sturmfels, P., Lundberg, S.M., Lee, S.I.: Improving performance of deep learning models with axiomatic attribution priors and expected gradients. Nature machine intelligence **3**(7), 620–631 (2021)
8. Gao, J., Xiao, C., Glass, L.M., Sun, J.: Dr. agent: Clinical predictive model via mimicked second opinions. Journal of the American Medical Informatics Association **27**(7), 1084–1091 (2020)
9. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000), circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215
10. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. CoRR **abs/1607.00653** (2016), http://arxiv.org/abs/1607.00653
11. Gruenberg, D.A., Shelton, W., Rose, S.L., Rutter, A.E., Socaris, S., McGee, G.: Factors Influencing Length of Stay in the Intensive Care Unit. American Journal of Critical Care **15**(5), 502–509 (09 2006). https://doi.org/10.4037/ajcc2006.15.5.502
12. Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. Scientific Data **6**(1), 96 (2019). https://doi.org/10.1038/s41597-019-0103-9
13. Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W.M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., Fernando, C., Kavukcuoglu, K.: Population based training of neural networks. CoRR **abs/1711.09846** (2017), http://arxiv.org/abs/1711.09846
14. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L.A., Mark, R.: Mimic-iv (version 1.0) (2021). https://doi.org/10.13026/s6n6-xd98
15. Khadanga, S., Aggarwal, K., Joty, S., Srivastava, J.: Using clinical notes with time series data for icu management. arXiv preprint arXiv:1909.09702 (2019)

16. Lingsma, H.F., Bottle, A., Middleton, S., Kievit, J., Steyerberg, E.W., Marang-van de Mheen, P.J.: Evaluation of hospital outcomes: the relation between length-of-stay, readmission, and mortality in a large international administrative database. BMC Health Services Research **18**(1), 116 (2018). https://doi.org/10.1186/s12913-018-2916-1
17. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R.: Learning to diagnose with lstm recurrent neural networks (2015). https://doi.org/10.48550/ARXIV.1511.03677
18. Lu, C., Han, T., Ning, Y.: Context-aware health event prediction via transition functions on dynamic disease graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 4567–4574 (2022)
19. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
20. Malone, B., Garcia-Duran, A., Niepert, M.: Learning representations of missing data for predicting patient outcomes. arXiv preprint arXiv:1811.04752 (2018)
21. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
22. Rocheteau, E., Liò, P., Hyland, S.: Temporal pointwise convolutional networks for length of stay prediction in the intensive care unit. In: Proceedings of the Conference on Health, Inference, and Learning. pp. 58–68. CHIL '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3450439.3451860
23. Rocheteau, E., Tong, C., Veličković, P., Lane, N., Liò, P.: Predicting patient outcomes with graph representation learning. arXiv preprint arXiv:2101.03940 (2021)
24. Schrodt, J., Dudchenko, A., Knaup-Gregori, P., Ganzinger, M.: Graph-representation of patient data: a systematic literature review. Journal of medical systems **44**(4), 1–7 (2020)
25. Segen's Medical Dictionary: length of stay (2011), https://medical-dictionary.thefreedictionary.com/length+of+stay
26. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning. pp. 3319–3328. PMLR (2017)
27. Suresh, H., Hunt, N., Johnson, A., Celi, L.A., Szolovits, P., Ghassemi, M.: Clinical intervention prediction and understanding using deep networks. arXiv preprint arXiv:1705.08498 (2017)
28. Wang, S., McDermott, M.B., Chauhan, G., Ghassemi, M., Hughes, M.C., Naumann, T.: Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 222–235 (2020)
29. Xu, Y., Biswal, S., Deshpande, S.R., Maher, K.O., Sun, J.: Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In: Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining. pp. 2565–2573 (2018)
30. Zhang, D., Yin, C., Zeng, J., Yuan, X., Zhang, P.: Combining structured and unstructured data for predictive models: a deep learning approach. BMC medical informatics and decision making **20**(1), 1–11 (2020)

Appendices

A Model significance

Table A.1. Identifiers for the different models, used in Table A.2.

TS0	Time-series gated recurrent unit (TS-GRU)
TS1	TS-GRU with decay mask
DG, EF	TS-GRU + demographic data, early fusion
DG, LF	TS-GRU + demographic data, late fusion
TV, EF	TS-GRU + vectorized transfers, early fusion
TV, LF	TS-GRU + vectorized transfers, late fusion
TF, EF	TS-GRU + transfer embeddings, early fusion
TF, LF	TS-GRU + transfer embeddings, late fusion
DV, EF	TS-GRU + demographic data + vectorized transfers, early fusion
DV, LF	TS-GRU + demographic data + vectorized transfers, late fusion
DT, EF	TS-GRU + demographic data + transfer embeddings, early fusion
DT, LF	TS-GRU + demographic data + transfer embeddings, late fusion

Table A.2. t-tests on the MSE (upper) and MAE (lower) for the models in Table 5. The model identifiers are explained in A.1. *: significant on 5% significance level. **: significant on 1% significance level.

MSE

	TS0	TS1	DG, EF	DG, LF	TF, EF	TF, LF	DT, EF	DT, LF	TV, EF	TV, LF	DV, EF	DV, LF
TS0	/											
TS1	1.67e-14**	/										
DG, EF	4.42e-19**	3.43e-11**	/									
DG, LF	1.61e-17**	4.36e-11**	1.66e-3**	/								
TF, EF	1.49e-16**	5.51e-7**	1.06e-4**	1.40e-6**	/							
TF, LF	1.48e-13**	1.98e-3**	1.66e-5**	4.96e-7**	5.32e-2	/						
DT, EF	3.51e-21**	5.96e-15**	9.66e-9**	2.30e-2*	1.53e-10**	1.31e-9**	/					
DT, LF	4.63e-19**	4.55e-13**	1.39e-6**	5.08e-2	5.71e-9**	8.89e-9**	9.96e-1	/				
TV, EF	2.53e-15**	3.75e-02*	7.20e-10**	4.27e-10**	3.50e-05**	6.06e-02	4.21e-14**	7.47e-07**	/			
TV, LF	1.92e-16**	1.30e-05**	5.05e-07**	4.06e-08**	6.70e-02	5.20e-01	2.55e-12**	2.61e-05**	1.61e-03**	/		
DV, EF	7.81e-18**	1.31e-10**	1.64e-01	4.79e-02*	3.81e-05**	6.74e-06**	1.58e-05**	1.23e-01	1.97e-09**	5.13e-07**	/	
DV, LF	1.04e-16**	7.70e-11**	1.88e-04**	1.95e-01	5.43e-07**	1.97e-07**	6.83e-01	5.62e-01	5.80e-10**	2.91e-08**	4.18e-03**	/

MAE

	TS0	TS1	DG, EF	DG, LF	TF, EF	TF, LF	DT, EF	DT, LF	TV, EF	TV, LF	DV, EF	DV, LF
TS0	/											
TS1	3.58e-7**	/										
DG, EF	6.05e-9**	8.55e-6**	/									
DG, LF	4.68e-12**	1.62e-8**	5.34e-1	/								
TF, EF	3.69e-8**	1.14e-3**	1.47e-2*	7.22e-3**	/							
TF, LF	4.45e-8**	1.44e-3**	1.28e-2*	6.00e-3**	9.39e-1	/						
DT, EF	6.63e-13**	1.56e-9**	6.42e-1	7.50e-1	2.97e-3**	2.44e-3**	/					
DT, LF	5.75e-14**	3.10e-11**	2.94e-1	6.73e-3**	3.69e-5**	3.10e-5**	6.36e-3**	/				
TV, EF	6.26e-09**	2.24e-04**	1.89e-02*	8.55e-03**	7.85e-01	7.24E-01	3.15e-03**	9.38e-05**	/			
TV, LF	1.55e-06**	5.23e-02	1.98e-03**	5.49e-04**	2.46e-01	2.75e-01	2.22e-04**	1.37e-05**	1.45e-01	/		
DV, EF	1.79e-12**	3.41e-09**	9.48e-01	3.60e-01	1.49e-03**	1.24e-03**	4.86e-01	5.18e-02	1.54e-03**	1.29e-04**	/	
DV, LF	5.10e-14**	3.93e-11**	6.93e-01	6.35e-02	1.76e-04**	1.45e-04**	7.68e-02	1.17e-01	1.35e-04**	1.76e-05**	3.95e-01	/