Analysis and Prediction of Patient Mortality and Length of Stay

Danlin Chen, Yichen Ding, Wenxi Hu

1 Motivation and Background

1.1 Motivation

Two primary uncertainties hospitals are facing for resource management: who will request an ICU admission and how long he/she will stay once admitted. These uncertainties affect hospitals to make fast and accurate decisions for the patients to a large extent. Early estimation of a patient's length of stay could be immensely helpful for hospitals to allocate resources such as wards and caregivers properly, to optimize patients' treatment plans, and to lower the chance of getting a hospital-acquired condition such as staph infection. Early identification of mortality rate could also be significant for hospitals to prepare necessary measures and treatment procedures. Moreover, with the presumption of the mortality rate and length of stay, hospital facilities could efficiently arrange the most appropriate treatment solutions for the patient in intensive care units and could potentially increase the survival rate.

1.2 Background

There was an increasing interest in addressing hospital mortality prediction and length of stay in recent years since applying machine learning methodologies on the observational health datasets could potentially improve health care in many ways. For this purpose, the Medical Information Mart for Intensive Care (MIMIC-III) dataset (Johnson et al., 2016) was designed to help credentialed researchers utilize the real-world healthcare datasets. MIMIC-III is publicly available. However, the architecture complexity of the health data and the absence of background knowledge still made the data extracting process pretty hard. Recent work from Wang et al. (2019) provided an open-source pipeline for transforming raw electronic health records in the MIMIC-III database into a commonly available data frame, which significantly mitigated the tedious work on feature engineering and extraction. Another research (Purushotham et al., 2017) presented a benchmark for several clinical prediction tasks that supported us in identifying critical clinical features.

2 Problem Statement

We wanted to answer the following questions through our project:
1) Can we predict the length of stay of a patient at the time of admission to the hospital?
2) What clinical features affect a patient's length of stay (LOS)?
3) What clinical features affect a patient's mortality?

The main challenge we were facing was mainly on two aspects: lack of MIMIC-III data preprocessing framework (with limited medical knowledge), and model choosing for different prediction tasks.

The first challenge was the lack of standardized data processing frameworks for MIMIC-III. Although MIMIC-III provided a free and public usable database, the high complexity of the data structure, and encrypted patients' information (such as date of birth, waved admission timestamp) still made the data extracting process pretty hard. Additionally, we extracted vital signs as features. The values of vital signs had various issues that needed to be solved, such as inconsistent units, out-of-range values, semantically equivalent features aggregation, and missing values. The lack of background knowledge about the medical meaning of the vital signs made the aggregation of comparable laboratory testing data difficult. It increased the difficulty of identifying the real crucial variables that affected the prediction task results.

The second challenge was the model selection. For the mortality prediction, the prior work has conducted a bunch of measurements (Jo & Rose,2015; Grnarova et al., 2016), but there was no standard methodology that yielded high accuracy in the prediction task. For the LOS prediction, since the vital signs data preserved the time series nature, we integrated the data with conducting a progressive analysis of the patients' conditions. However, the time-series data for every subject in the table had different timestamps and contained missing values. It has been noted that missing values and their missing patterns are often correlated with the target labels, a.k.a., informative missingness. There was minimal work on exploiting the missing patterns for sufficient imputation and improving prediction performance. We needed to find a proper model that could interpolate sparse and irregularly sampled multivariate time series with missing values.

3 Data Science Pipeline

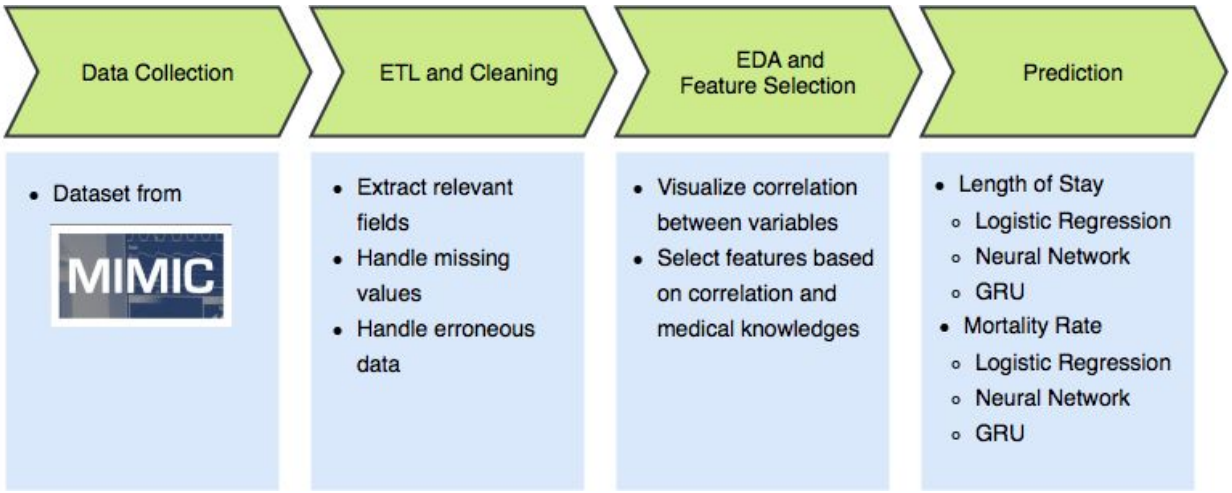

Fig 1. Pipeline

3.1 Data Collection

We chose the MIMIC-III database (Johnson et al., 2016), which contained the most comprehensive clinical data, to achieve our prediction goals.

MIMIC-III is a freely accessible database containing de-identified health data of patients who stayed in intensive care units at the Beth Israel Deaconess Medical Center from 2001 to 2012. MIMIC-III contains 50,000 distinct hospital admissions from 40,000 patients with associated demographics and physiological data collected during their hospital stay. We acquired the MIMIC-III database access following the official instructions, which involved the completion of an online course about healthcare.

3.2 Data Preprocessing

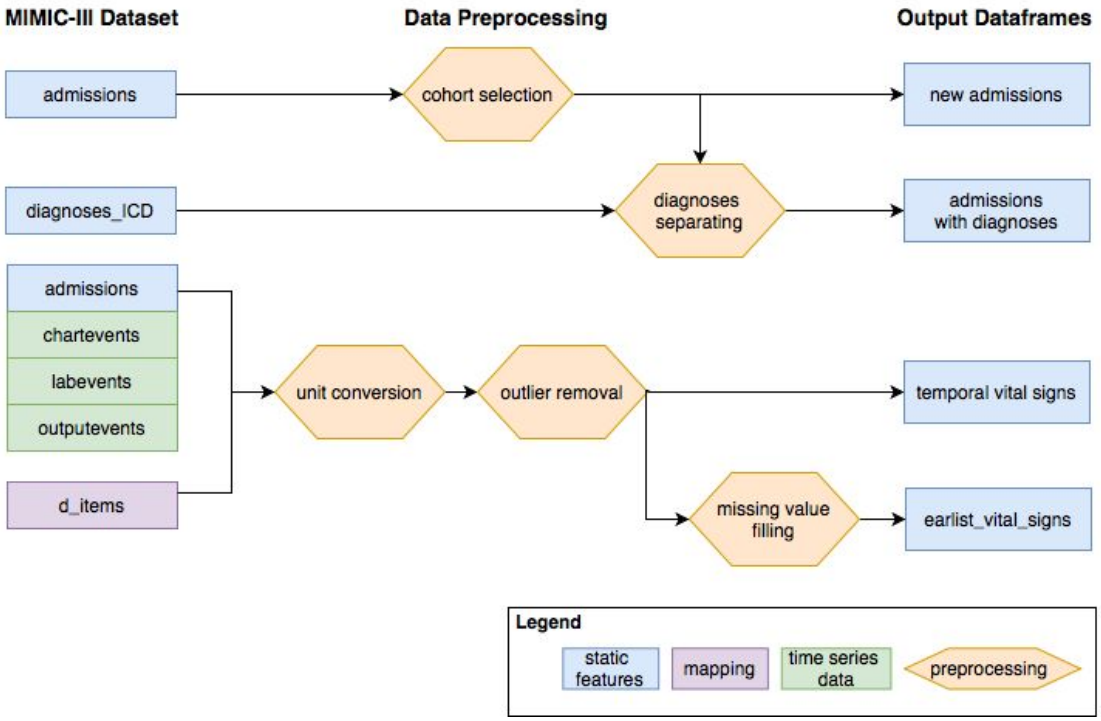

Fig 2. Data Preprocessing

3.2.1 Cohort Selection

Following previous studies (Purushotham et al., 2017), we excluded all ICU stays where the patient was younger than 15 due to the substantial differences between adult and pediatric physiology. The resulting root cohort had 38,579 unique records, with a total of 50,777 ICU admissions and over 250 million clinical events. The resulting cohort had a diverse demographic and admission coverage. This cohort selection was consistent with many papers using MIMIC-III (Ghassemi et al., 2014, 2015, 2016, 2017; Suresh et al., 2017; Raghu et al., 2017; McDermott et al.,2018).

3.2.2 Data Extraction

In MIMIC-III, there are 26 Tables of the total size about 50GB. Besides tables that record the demographic information of the patient and admission, charted events such as laboratory tests, fluids into and out of patients are stored in a series of 'events' tables. In order to obtain the most relevant clinical features for the length of stay and mortality prediction tasks, we extracted data for the cohort from the following tables:

Table	Description
Admissions	Gives information regarding a patient's admission to the hospital
Patients	Provides a single patient's static demographics and outcomes during the stay
Diagnoses_ICD	Contains ICD-9 diagnosis Code for patients
Chartevents	Contains all the charted data available for a patient (i.e. Routine vital signs, code status, mental status, etc)
Labevents	Contains information regarding laboratory based measurements
Outputevents	Output data for patients (i.e. Urine)
D_items	Dictionary of non-laboratory-related charted items.
D_labevents	Dictionary of laboratory-related items.

Table 1. Data Tables

We computed the age for each patient by subtracting their admission time and their DOB (modified date of birth) and computed the length of stay (in days) for each admission by subtracting the discharge time and the admission time. The HOSPITAL_EXPIRE_FLAG in the admission table was a binary value to identify if the patient died in the hospital, which we used as our target value in the mortality prediction task.

After researching the previous works, we noticed that SAPS II (Simplified Acute Physiology Score) was a generally acknowledged system that provided the ICU severity score, and many benchmark researchers used the variables in this system as features (Purushotham et al., 2017, Wang et al., 2019). Therefore, we extracted the corresponding physiological items based on the SAPS-II scoring system (Variable names listed in Appendix 1).

Besides, we also extracted vital signs that had the least missing values for further feature engineering (Appendix 2). Since we also conducted the time-series analysis in the prediction tasks, the temporal data of vital signs need to have relatively low-percentage missing values. The following extra vital signs have also been extracted: Peripheral Capillary Oxygen

Saturation, Respiratory Rate, Diastolic Blood Pressure, Trophoblast giant cells, Glucose, capillary refill rate, and Ph. All the above vital signs data are extracted from Chartevents, Labevents, and Outputevents. At this point, each vital sign had one data frame that contained all the timestamps results from all the patients, which could be used as raw input for the time-series analysis.

According to the official MIMIC-III description, there were duplicate ITEMID for the variables we wanted to extract. For example, both ITEMID 211 and ITEMID 220045 represented Heart Rate. Thus it was necessary to search for multiple ITEMID to capture a single physiological variable across the entire database. We found all ITEMIDs associated with the variables we need according to previous researches(Purushotham et al., 2017). We took the max value in case of multiple recordings present at the same time for each admission.

3.2.3 Data Cleaning

After we closely looked into the extracted data, we found that the data had lots of erroneous entries, such as missing values and outliers. We identified and handled the following issues:

3.2.3.1 Missing values

Missing values were frequently observed in data from events tables because a patient may not have the medical measurements that were not necessary for his/her medical condition. When combining all measurements for each admission, the amount of those admissions that had all selected features available might be minimal. Thus, we must handle those missing values instead of dropping them for the prediction task. However, as we showed in the data preprocessing Fig 2, we only filled the missing values for the data with the earliest timestamp. This preprocessing difference was caused by the input differences between the models that we chose. For continuous values, we filled the missing values by mean; For categorical values, we filled the missing values by the possible categories according to their proportions.

3.2.3.2 Outliers

For each clinical feature as well as demographic features, we plotted a histogram and checked the statistical descriptions of the feature to check potential outliers. We made use of a list of clinically reasonable variable ranges provided in the source code repository of Harutyunyan et al. (2019), which was developed in conversation with clinical experts, based on their knowledge of valid clinical measure ranges. Once we detected extreme values, we removed them from the table.

3.2.3.3 Inconsistent Units

We observed that for some items in the events tables, item values were recorded with inconsistent units. For example, item 113 (central venous pressure) had most of the rows recorded with unit ‘mmHG,’ and a small number of rows were recorded with ‘%.’ Because ‘mmHG’ was the primary unit that accounts for over 90% of the total number of records, we dropped the remaining records with different units.

3.2.3.4 Inconsistent data type

Some items had numeric values but were recorded with additional characters as a string(i.e., '< 5'). In this case, we examined the value distribution of such items and found that the majority (>90%) of the data were recorded generally as a number, and those exceptional cases were less than 10% of the total number of records. Thus we converted those particular values as None and then filled them by methods described above.

3.3 EDA

In order to gain a better understanding of the data we collected and find the underlying correlations between patient's length of stay, mortality, and each physiological feature, we performed exploratory data analysis on several features we extracted.

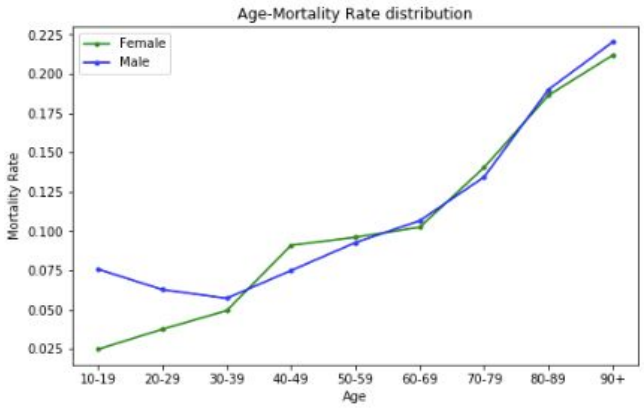

Fig.3 Gender v.s. Age v.s. mortality rate.

Fig.3 showed that as the age increased, the patients tended to have higher mortality rates. The difference between the gender difference for the mortality aspect was not quite obvious for ages over 50.

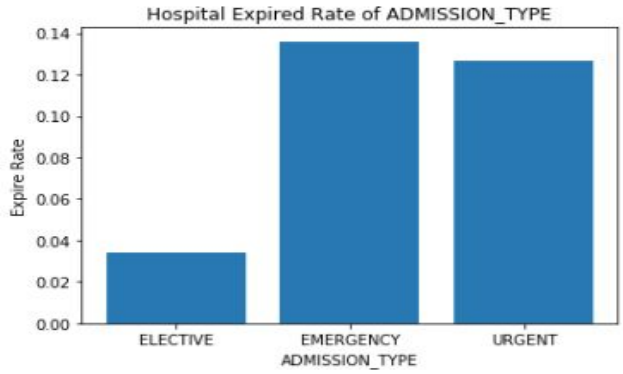

Fig 4. Admission Type v.s. Mortality Rate

Fig.4 showed the relationship between the admission types and hospital expired rate. The emergency type tended to have the highest mortality rate, and elective type tended to have the lowest mortality rate.

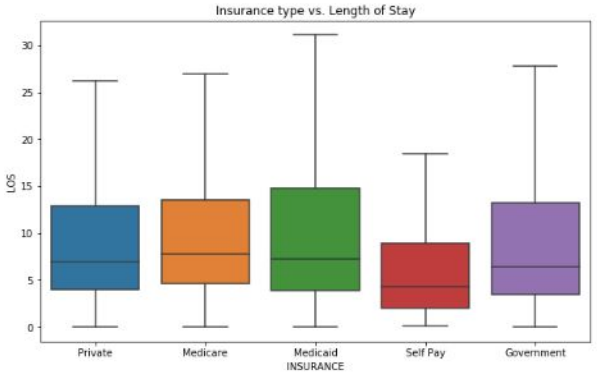

Fig 5. Insurance v.s. LOS

Fig. 5 showed the boxplots of insurance types and LOS. The type self-pay tended to have the lowest median LOS, this could be caused by the high expenses in the ICU stays.

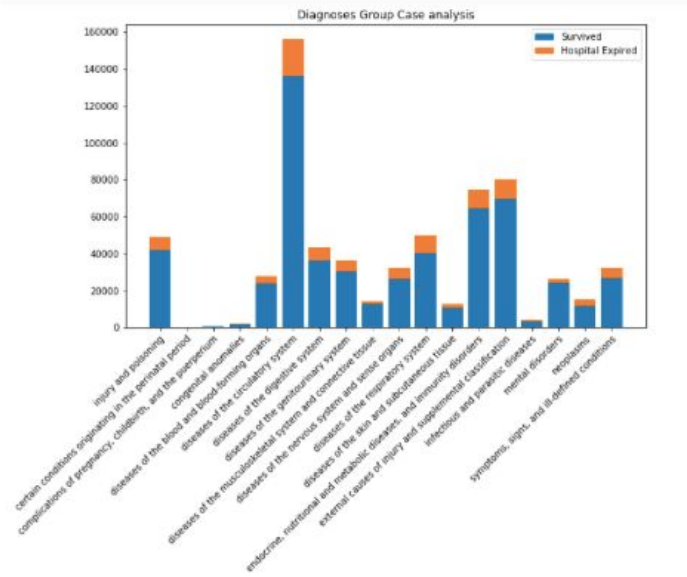

Fig 6. Diagnoses Mortality Distribution

Fig.6 showed the in-hospital mortality distribution for each diagnosis group diseases of circulatory system
The diagnosis that had the highest mortality rate was “certain conditions originating in the perinatal period”. The diagnosis that had the lowest rate was “complications of pregnancy,childbirth, and the puerperium”.

3.4 Feature Selection

We used two different approaches for our prediction tasks: one was using temporal features, and the other was using static features.

3.4.1 Static features

Because all the vital signs measurement data were temporal in the dataset, we converted them to static features by extracting the earliest measurement for each admission. Besides features from the SAPS-II scoring system(Appendix 1), we wanted to explore other potential features that might be helpful for our prediction tasks, and compare the performance on the 2 different feature sets.

We merged all the vital signs measurements we extracted from the database and essential admission and patient-level information we analyzed through EDA, and applied one-hot encoding for all categorical features. Then we used the RandomForest model to get the feature importance for both lengths of stay prediction and mortality prediction separately in order to get the most relevant features for each prediction task. Finally, we selected 3 additional features (43 in toal) for the length of stay prediction and 2 additional features (42 in total) for mortality prediction as our customized feature sets.

3.4.2 Temporal features

Due to the specificity of the model that we chose, the temporal input data of vital signs need to contain high-intensive timestamps and low-percentage missing values. After inspecting the

timestamps and value missingness for all the SAPS-II features, we only retained 4 features from the benchmark based on these criteria. For the extra vital signs data, we added 5 more features by using the RandomForest feature importance and the criteria. After merging the time-series vital signs data with the corresponding LOS and Hospital_Expire_Flag, all the records with LOS shorter than 24 hours were removed.

3.5 Modeling

We divided our prediction tasks into 3 binary classification problems:

- In-hospital mortality prediction: predict whether a patient dies during the hospital stay
- Short stay prediction: predict whether a patient will stay in the hospital for less than 3 days since admission
- Long stay days prediction: predict whether a patient will stay in the hospital for more than 7 days since admission

We noticed that the distribution of length of stay in terms of days was highly skewed, as shown in Fig 7. Additionally, the proportion of hospital expired admission cases was extremely unbalanced, with an overall in-hospital mortality rate of 12.14% from the cohort. To avoid the bias in each prediction task, we under-sampled our training dataset in order to balance the data and thus get a more reliable result.

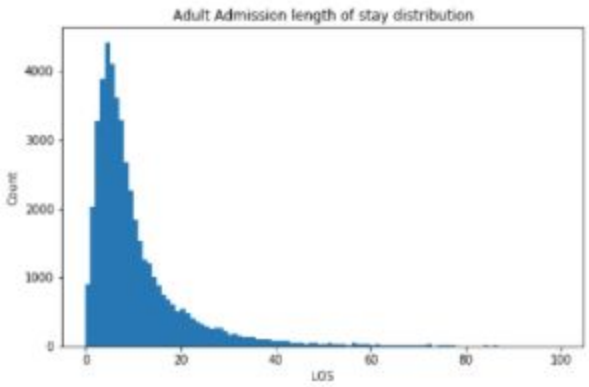

Fig 7. LOS Distribution

We chose Logistic Regression from sklearn as our baseline model for both length of stay and mortality prediction. We experimented with a multi-layer neural network model (MLNN) and a gated recurrent unit model (GRU) to compare the prediction performance.

4 Methodology

4.1 Multi-Layer Neural Network (MLNN)

We constructed a multi-layer neural network for our classification problems using static features. The reason we chose Neural Network was that Neural Networks had the ability to learn and model non-linear and complex relationships.

We used the following principal features for both length of stay and mortality prediction tasks:

- 16 earliest measurement for variables in SAPS-II scoring system
- Diagnosis group (Find the specific table in Appendix 3)
- Admission information: admission type, insurance type
- Patient information: age, gender

For both short term and long term length of stay prediction, we added the earliest measurement for SaO2, Respiratory Rate, CVP, Diastolic Arterial Blood Pressure, and Peripheral Capillary Oxygen Saturation; for in-hospital mortality prediction, we added the earliest measurement for Anion Gap, Creatinine, Respiratory Rate, and Peripheral Capillary Oxygen Saturation. The additional features were obtained from our feature selection section.

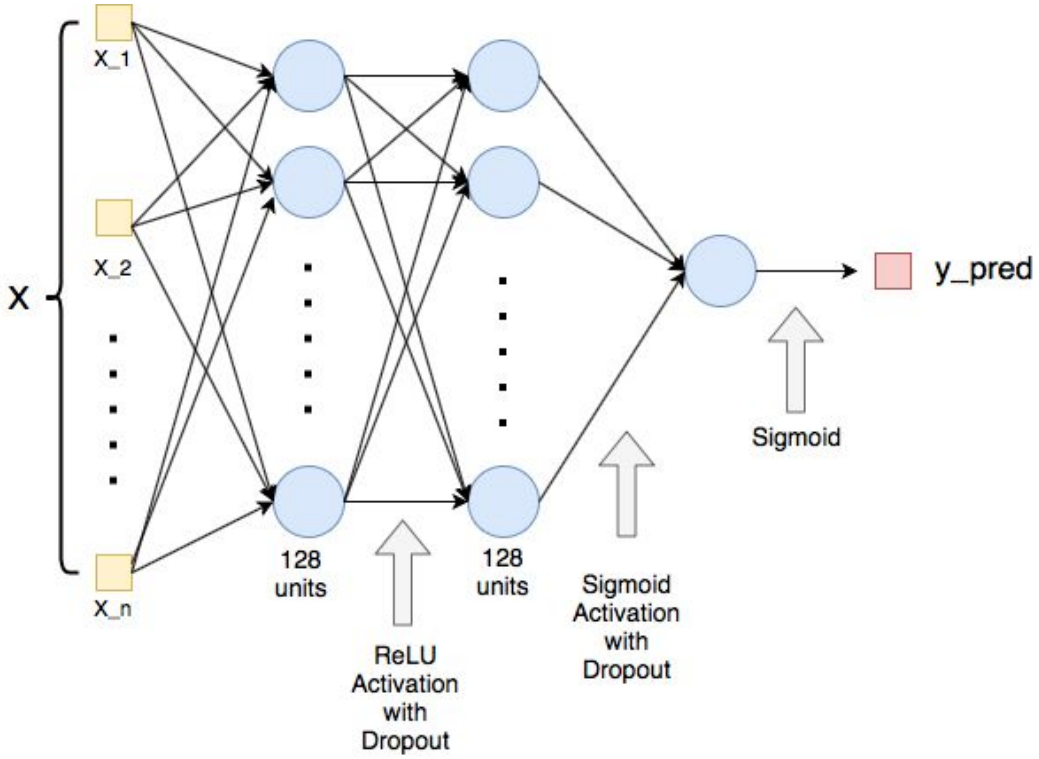

Fig 8: Multi-Layer Neural Network

The architecture of the neural network is shown in Fig 8.

4.2 Gated Recurrent Unit (GRU)

By the temporal nature of clinical data, we used the temporal data of vital signs to predict the mortality and length of stay. After inspecting all the time series data, we noticed that the timestamps were not uniformed and contained a certain amount of missing values. In the previous work (Che et al., 2019), the GRU model achieved impressive performance since it could deal with the problem of supervised learning with sparse and irregularly sampled multivariate time series data. In this project, we reimplemented the architecture of Shukla et al. (2019), which used a semi-parametric interpolation network followed by a GRU network as the prediction network to perform the prediction tasks.

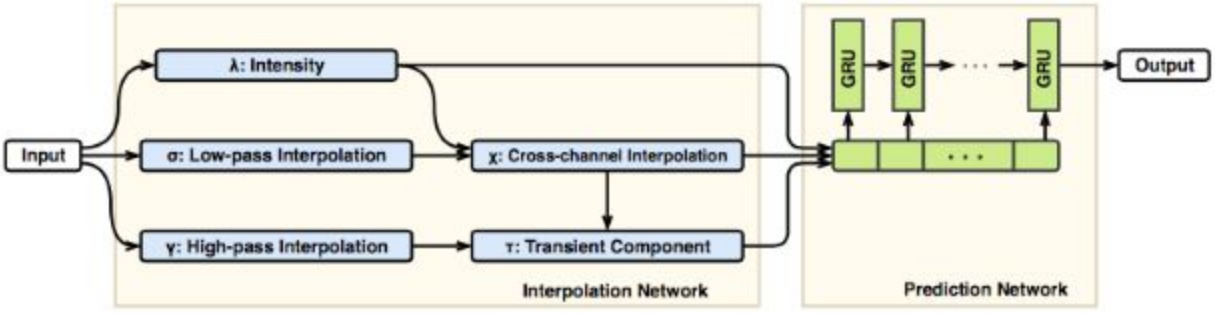

Fig 9. GRU Architre

Input Notation:

Let $D = \{(S_n, y_n) \mid n = 1, ..., N\}$ represents a data set containing N data cases. An individual data case consists of a target value y_n (categories for classification), and a D-dimensional, sparse and irregularly sampled multivariate time series S_n. Different features in the multivariate time series could have various observed timestamps and corresponding values. Let L_{dn} be the total number of observations. Thus, we represented time-series of feature d for data case n as $S_{dn} = (t_{dn}, x_{dn})$ where the length of t_{dn} and x_{dn} are L_{dn}. t_{dn} is the list of all observed timestamps, and x_{dn} is the corresponding list of observed values.

Architecture:

The model consisted of an interpolation network and a GRU prediction network. The interpolation network interpolated the multivariate, sparse, and irregularly sampled input time series against a set of reference time points $r = [r_1, ..., r_T]$. Every feature d in input D had a corresponding r_n, r_n contains all the observed timestamps for all the admission records. The time series were all defined within 24 hours after admission. The T reference time points r_t were chosen to be evenly spaced within that interval. The interpolation network contained two layers, and each layer performed a different type of interpolation. The first layer in the interpolation network separately performed 3 semi-parametric univariate transformations based on a radial basis function (RBF) network to accommodate the D continuous-time observations. The transformations were a low-pass (or smooth) interpolation σ_d, a high-pass (or non-smooth) interpolation γ_d, and an intensity function λ_d. The smooth interpolation σ_d used a squared exponential kernel with parameter α_d, while the non-smooth interpolation γ_d used a squared exponential kernel with parameter $\kappa\alpha_d$ for $\kappa > 1$. The equations 1,2, 3 and 4 showed the interpolation calculation process, where r was the reference timestamp, t was the observed

timestamp, d was the feature, and x was the observed feature value:

$$Z(r, \mathbf{t}, \alpha) = \sum_{t \in \mathbf{t}} w(r, t, \alpha)\,, \quad w(r, t, \alpha) = \exp(-\alpha(r-t)^2) \tag{1}$$

$$\lambda_{kd} = h_\theta^\lambda(r_k, \mathbf{t}_d, \mathbf{x}_d) = Z(r_k, \mathbf{t}_d, \alpha_d) \tag{2}$$

$$\sigma_{kd} = h_\theta^\sigma(r_k, \mathbf{t}_d, \mathbf{x}_d) = \frac{1}{Z(r_k, \mathbf{t}_d, \alpha_d)} \sum_{j=1}^{L_{dn}} w(r_k, t_{jd}, \alpha_d)\, x_{jd} \tag{3}$$

$$\gamma_{kd} = h_\theta^\gamma(r_k, \mathbf{t}_d, \mathbf{x}_d) = \frac{1}{Z(r_k, \mathbf{t}_d, \kappa\alpha_d)} \sum_{j=1}^{L_{dn}} w(r_k, t_{jd}, \kappa\alpha_d)\, x_{jd} \tag{4}$$

The second interpolation layer merged the time series of all the features at each reference timestamp by calculating a correlations $\rho_{dd'}$ across time series, and then calculated an across-dimension interpolation χ_d for each input dimension. A transient component Γ_d was calculated, which was the difference between the high-pass from the first layer and the smooth cross-dimension interpolation for each input dimension. The transient component Γ_d was calculated as shown in equation (5).

$$\chi_{kd} = h_\theta^\chi(r_k, \mathbf{s}) = \frac{\sum_{d'} \rho_{dd'}\, \lambda_{kd'}\, \sigma_{kd'}}{\sum_{d'} \lambda_{kd'}}\,, \qquad \tau_{kd} = h_\theta^\tau(r_k, \mathbf{s}) = \gamma_{kd} - \chi_{kd} \tag{5}$$

We defined $f_\theta(r,\ s_n)$ to be the function computing the output of the interpolation network. The output was $\hat{s_n} = f_\theta(r,\ s_n)$ a fixed-sized array with dimensions $(DC) \times T$ for all inputs. The prediction network took $\hat{s_n}$ as input, and output the estimated target value $\hat{y_n} = g_\phi(f_\theta(r,\ s_n))$.

Loss Function:

A composite objective function had a supervised and an unsupervised component was used to learn the model parameters. The supervised component was not insufficient to learn reasonable parameters given the amount of available training data. The unsupervised component used an autoencoder-like loss function. The model masked some observed data to force the semi-parametric RBF interpolation layers to learn correctly from the interpolated data. For each data point $(t_{jdn},\ x_{jdn})$, if $m_{jdn} = 1$, the model removed the data point from the input. We used $(1 - mn) \odot s_n$ to represent the values that were not masked out. Let l_p be the loss of the prediction network, and l_I be the loss of the interpolation network. L_2 regularizations were also added. The loss function was as equation 6:

$$\theta_*, \phi_* = \arg\min_{\theta,\phi} \sum_{n=1}^{N} \ell_P(y_n, g_\phi(f_\theta(\mathbf{s}_n)) + \delta_I \|\theta\|_2^2 + \delta_P \|\phi\|_2^2 \quad (6)$$
$$+ \delta_R \sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{j=1}^{L_{dn}} m_{jdn} \ell_I(x_{jdn}, h_\theta^\chi(t_{jdn}, (1 - \mathbf{m}_n) \odot \mathbf{s}_n))$$

Why interpolation layers with GRU?

It consisted of global interpolation layers. The proposed model used semi-parametric, deterministic, feed-forward interpolation layers as Shukla et al.(2019) described. Without encoding uncertainty, these layers allowed for very flexible interpolation both within and across layers. The interpolation layers produced regularly sampled interpolants that could serve as inputs for arbitrary, unmodified, deep classification networks. The model included the information about the timestamps at which corresponding observations occurred. Pre-discretizing the inputs, using the information in binary observation masks, and missing data indicator sets were generally used in the previous work (as Che et al. (2018a)) to deal with the sparse and irregularly sampled time series. This model used the sequence of observation events directly, and treated it as a point process in continuous time by fitting the semi-parametric intensity functions (Lasko, 2014).

5 Evaluation

In addition to Accuracy for evaluating the binary classification models' performance, we also used the Area under the ROC curve (AUROC), which indicated how much our model is capable of distinguishing between each class. AUROC measures how true positive rate (i.e., possibility of predicting 1 when the true label is 1) and false-positive rate (i.e., possibility of predicting 1 when the true label is 0) trade-off. Moreover, higher then AUROC, better the model is predicting 0s as 0s and 1s as 1s.

5.1 Length of Stay Prediction

		SAPS-II features only		Customized features	
		Test Accuracy	Test AUROC	Test Accuracy	Test AUROC
Short Stay Prediction	LR	0.6968	0.7599	0.6971	0.7614
	MLNN	0.7362	0.8091	0.7380	0.8140
	GRU	NaN		0.7532	0.8206

Long Stay Prediction	LR	0.6871	0.7615	0.6892	0.7626
	MLNN	0.7270	0.8075	0.7326	0.8097
	GRU	NaN		0.6715	0.7255

5.2 In-Hospital Mortality Prediction

	SAPS-II features only		Customized features	
	Test Accuracy	Test AUROC	Test Accuracy	Test AUROC
LR	0.7480	0.8260	0.7572	0.8383
MLNN	0.7814	0.8597	0.7821	0.8604
GRU	NaN		0.6635	0.7021

For all binary classification problems, we compared the performance of the 3 models over 2 different feature sets described in Section3.4. From the result, we achieved a slightly higher performance for both the baseline model and MLNN when we used our customized feature set. For short stay prediction, GRU had the best performance over the other 2 models with Accuracy 75.32% and AUROC 82.06%. Since the GRU model only used the time-series data and excluded the static features, the feature set of SAPS-II contained static features like age and weight. Hence, we could not apply the GRU models with the SAPS-II feature set in both tasks. For the LOS prediction and mortality prediction, MLNN had the best performance over the other 2 models with Accuracy 73.26% and AUROC 80.97%. For mortality prediction, MLNN had the best performance over the other 2 models with Accuracy 78.21% and AUROC 86.04%.

6 Data Product

Our data product is a set of IPython Notebooks providing detailed length of stay and mortality analysis, and solutions of predicting a patient's ICU stay outcome when he/she is admitted to the hospital. Due to time constraint, we could not build an interactive user interface for our solution.

7 Lesson Learned

Real-world datasets are usually unstructured and messy, thus data preprocessing plays a significantly important role in data science pipelines. We should always look closely at the data to eliminate outliers and missing values before using them. Successful detection of outliers and filling missing values with efficient methods could potentially improve the model performance.

Most importantly, when dealing with classification problems, we need to pay attention to the distribution of our target values. If the distribution is skewed, we need to apply proper up-sampling or down-sampling methods to balance our data for eliminating the skewness of the data.

8 Summary

In this project, we built a data science pipeline for analyzing and predicting the patient's length of stay and mortality. We have collected data from MIMIC-III, extracted, and cleaned clinical variables that were correlated with length of stay and mortality. We approached the prediction tasks in 2 different ways: using temporal vital signs measurements and applying to a GRU model, using static information combined with extracted earliest measurements of crucial vital signs and applying to an MLNN model. We selected features based on the SAPS-II system and RandomForest feature importance and then obtained 2 feature sets: SAPS-II features as a baseline feature set and our customized feature set, which we aimed to achieve a better performance than the baseline feature set. For the short term length of stay prediction, we got 75.32% accuracy and 82.06% AUROC with GRU model using our customized feature set, outperforming the baseline model. For the long term length of stay prediction, we got 73.26% accuracy and 80.97% AUROC with the MLNN model using our customized feature set, which performed better than the baseline model. For in-hospital mortality prediction, we got 78.21% accuracy and 86.04% AUROC with the MLNN model using our customized feature set, outperforming the baseline model as well.

Reference

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, N. Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In International Conference on Knowledge Discovery and Data Mining (KDD), pages 75–84. ACM, 2014.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate time series modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

Marzyeh Ghassemi, M. Wu, M. Feng, L.A. Celi, P. Szolovits, and F. Doshi-Velez. Understanding vasopressor intervention and weaning: Risk prediction in a public heterogeneous clinical time series database. Journal of the American Medical Informatics Association, page ocw138, 2016.

Marzyeh Ghassemi, Mike Wu, Michael Hughes, and Finale Doshi-Velez. Predicting intervention onset in the icu with switching state space models. In Proceedings of the AMIA Summit on Clinical Research Informatics (CRI), volume 2017. American Medical Infor- matics Association, 2017.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neu- ral networks. In Proceedings of the 2nd Machine Learning for Healthcare Conference, volume 68 of Proceedings of Machine Learning Research, pages 322–337, Boston, Mas- sachusetts, 18–19 Aug 2017. PMLR.

https://github.com/MIT-LCP/mimic-code/blob/master/concepts/pivot/pivoted-vital.sql
https://github.com/MIT-LCP/mimic-code/blob/master/concepts/pivot/pivoted-lab.sql

Jo, Yohan and Rose´, Carolyn Penstein. Time Series Analy- sis of Nursing Notes for Mortality Prediction via a State Transition Topic Model. Proceedings of the 24th ACM International Conference on Information and Knowl- edge Management, 2015.

T.Banerjee,M.Peterson,Q.Oliver,A.Froehle,L.Lawhorne,ValidatingaCommercialDeviceforContinuousActivityMeasurementinthe Older Adult Population for Dementia Management, Smart Health (2017)

Grnarova, Paulina, Schmidt, Florian, Hyland, Stephanie L, and Eickhoff, Carsten. Neural Document Embeddings for Intensive Care Patient Mortality Prediction. arXiv, cs.CL, 2016.

A. Raghu, M. Komorowski, L.A. Celi, P. Szolovits, and M. Ghassemi. Continuous state- space models for optimal sepsis treatment: a deep reinforcement learning approach. In Machine Learning for Healthcare Conference (MLHC), pages 147–163, 2017.

M.B.A. McDermott, T. Yan, T. Naumann, N. Hunt, H. Suresh, P. Szolovits, and M. Ghas- semi. Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs. In Association for the Advancement of Artificial Intelligence, New Orleans, LA, 2018.

Satya Narayan Shukla and Benjamin Marlin. Interpolation-prediction networks for irregularly sampled time series. In International Conference on Learning Representations, 2019.

Joseph Futoma, Sanjay Hariharan, and Katherine A. Heller. Learning to detect sepsis with a mul- titask gaussian process RNN classifier. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pp. 1174–1182, 2017. URL http://proceedings.mlr.press/v70/futoma17a.html.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. Scientific Reports, 8(1):6085, 2018a. URL https://doi.org/10.1038/s41598-018-24271-9.

Wang, Shirly & McDermott, Matthew & Chauhan, Geeticka & Hughes, Michael & Naumann, Tristan & Ghassemi, Marzyeh. (2019). MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III. arXiv:1907.08322.

Sanjay Purushotham, Chui zheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. Journal of Biomedical Informatics, 83, 2018. doi: https://doi.org/10.1016/j.jbi.2018.04.007. URL https://www.sciencedirect.com/science/article/pii/S1532046418300716.

Appendix 1 - SAPS-II Features

GCS Verbal Response	GCS Motor Response	GCS Eye Opening
Systolic Blood Pressure	Heart Rate	Body Temperature
Partial Pressure of Oxygen	Fraction Inspired Oxygen	Oxygen Saturation
Urine Output	Urea Nitrogen	White Blood Cells Count
Bicarbonate Level	Sodium Level	Potassium Level
Bilirubin Level	Age	Diagnoses Group
Admission Type		

Appendix 2 - Vital Signs with Least Missing Values

ITEMID	Variable Name	ITEMID	Variable Name
646,	SpO2	50868,	Anion Gap
212,	Heart Rhythm	161,	Ectopy Type
128,	Code Status	550,	Precautions
1125,	Service Type	159	Ectopy Frequency
1484,	Risk for Falls	8368,	Arterial BP [Diastolic]
5815,	HR Alarm [Low]	8549,	HR Alarm [High]
5820,	SpO2 Alarm [Low]	8554,	SpO2 Alarm [High]
5819,	Resp Alarm [Low]	8553	Resp Alarm [HIgh]
834,	SaO2	51248	MCH
581,	Previous WeightF	8441,	NBP [Diastolic]
456,	NBP Mean	31,	Activity
5817,	NBP Alarm [Low]	8551,	NBP Alarm [High]
113,	CVP	1703,	Health Care Proxy
467,	O2 Delivery Device	80,	Bowel Sounds
1337,	Riker-SAS Scale	674,	Temp. Site
432	Level of Conscious	5813,	ABP Alarm [Low]
8547,	ABP Alarm [High]	617,	Respiratory Pattern
210,	HOB	637,	Side Rails
198,	GCS Total	707	Urine Source
704,	Turn	479,	Orientation
54,	Assistance Device	32,	Activity Tolerance
547,	Position	154	Diet Type
51221,	Hematocrit	50912,	Creatinine
50902,	Chloride	51265,	Platelet Count
51222,	Hemoglobin	50931,	Glucose
51249,	MCHC	51279,	Red Blood Cells

Appendix 3 - Diagnosis Group

ICD9-CODE	Description
001-139	Infectious And Parasitic Diseases
140-239	Neoplasms
240-279	Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders
280-289	Diseases Of The Blood And Blood-Forming Organs
290-319	Mental Disorders
320-389	Diseases Of The Nervous System And Sense Organs
390-459	Diseases Of The Circulatory System
460-519	Diseases Of The Respiratory System
520-579	Diseases Of The Digestive System
580-629	Diseases Of The Genitourinary System
630-679	Complications Of Pregnancy, Childbirth, And The Puerperium
680-709	Diseases Of The Skin And Subcutaneous Tissue
710-739	Diseases Of The Musculoskeletal System And Connective Tissue
740-759	Congenital Anomalies
760-779	Certain Conditions Originating In The Perinatal Period
780-799	Symptoms, Signs, And Ill-Defined Conditions
800-999	Injury And Poisoning
E and V codes	External Causes Of Injury and Supplemental Classification