Building Deep Learning Models to Predict Mortality in ICU Patients https://youtu.be/5RzyXMPvLK0

Yuanfei Bi
College of Computing, George Institute of Technology, Atlanta, GA, USA

1 Abstract

Mortality prediction in the Intensive Care Units (ICUs) is considered as one of critical steps for the efficient treatment of patients in serious condition. As a result, various prediction models have been developed to address this problem based on modern electronic healthcare records (EHR). However, it becomes increasingly challenging to model such task as time-series variables because some of the laboratory test results such as heart rate and blood pressure are sampled with inconsistent time frequencies. In this paper, we propose several deep learning models using the same features as the SAPS-II score[1]. To derive insight into the performance of the proposed models, several experiments have been conducted based on the well-known clinical dataset Medical Information Mart for Intensive Care III (MIMIC-III, v1.4)[2]. The prediction results demonstrates the capability of the proposed models in terms of precision, recall, F1-score and area under the receiver operation characteristic curve (AUC).

2 Introduction

Mortality prediction in Intensive Care Units (ICUs) is wards in hospital where specially trained physicians provide support to the most severely ill patients. However, it is a common challenge that physicians do not have intelligent tools to process massive amount of modern electronic healthcare records (EHR). The accurate and reliable mortality prediction for ICU patients is crucial for physicians to assess severity of illness, determine appropriate levels of cares and provide radical life saving treatment. Patients are monitored closely within ICUs to ensure any deterioration is detected and corrected before it becomes fatal. As a result, there is an increasingly large amount of ICUs data in EHR. Today, deep learning models (aka Deep Neural Networks) have revolutionized many fields such as natural language processing (NLP), voice recognition and computer vision, and are being increasingly adopted in clinical healthcare fields.

The goal of this paper is to develop a deep learning model which can identify patients hospitalized in the ICUs at high risk for death during the ICU stay based on the EMR dataset accumulated by the first 48 hours of the first ICU admission. We propose a method to extract both sequential and non-sequential features from the MIMIC-III (v1.4) database[2] and build several recurrent neural network (RNN) models to predict hospital mortality, i.e., death inside the hospital.

The rest of the paper is organized as follows: in Section 3, we present a literature review on the related studies; in Section 4, we provide a basic statistics of MIMIC-III (v1.4) dataset; in Section 5, we describe the pre-processing step we employed to obtain the features and the proposed RNN models; the experimental results are presented and discussed in Section *Results* and *Discussion*, respectively; and we conclude with summary in Section *Conclusion*.

3 Related Work

Recent advances and success of machine learning and deep learning have facilitated the adoption of these models into mortality prediction tasks for ICU patients. Early work[3–5] showed that machine learning models obtain good results on mortality prediction in ICUs. Recently, an ensemble technique called Super Learner (SL) is proposed to offer improved performance of mortality prediction in ICU patients[6]. Among a given set of candidate algorithms, the SL technique builds an aggregate algorithm as the optimal weighted combination of the candidate algorithms. Their work has demonstrated that machine learning models outperform the prognostic scores.

With freely-available datasets such as MIMIC-III, the development of novel models for mortality prediction is gaining increased attention. Lee et al.[7] demonstrated a personalized 30-day mortality prediction model by analyzing similar past patients. Johnson et al.[8] compared multiple published mortality prediction works against gradient boosting and logistic regression model using a simple set of features extracted from MIMIC-III dataset. Recently, researchers have attempted to applied deep learning based methods to EHR to utilize its ability to learn complex patterns from data. Dabek et al. showed that a neural network model can improve the prediction of several psychological conditions such as anxiety, depression and behavioral disorders[9]. Che et al.[10] developed a novel recurrent neural network (RNN) model based on Gated Recurrent Unit (GRU) which demonstrates promising performance for ICU mortality prediction. Some RNN models with LSTM units are also proposed and compared with baseline models to show better ICU mortality prediction accuracy[11–14].

4 Data

MIMIC-III (v1.4)[2] is a publicly available critical care database maintained by the Massachusetts Institute of Technology (MIT). This database integrates clinical data of over 40,000 patients admitted to ICUs of the Beth Israel Deaconess Medical Center during 2001 to 2012. MIMIC-III consists of 26 relational tables, where 16 of them contain timestamped event information. Table 1 shows the statistics of MIMIC-III (v1.4) dataset. In this project, we will focus on the ICU-related data of adult patients.

Table 1: Summary statistics of MIMIC-III (v1.4) dataset.

# of patients	46520
# of adult patients [a]	38597
Median age of adult patients	65.8 years
In-hospital mortality of adult patients	11.5%
# of admissions	58976
# of ICU stays	61532
# of ICU stays of adult patients	53423
# of long ICU stays [b] of adult patients	53133
# of the first long ICU stay of adult patients	38418
Avg. length of long ICU stays of adult patients	4.17 days
Avg. length of ICU stays of adult patients	4.14 days
Avg. length of the first long ICU stays of adult patients	4.07 days

[a] Adults: $\geq$ 16 years old.
[b] Long ICU stays: $\geq$ 4 hours.

5 Methodology

5.1 Problem Definition

The model we proposed to identify patients hospitalized in the ICU is based on the EMR data accumulated by the first 48 hours into the first ICU stay, as illustrated in Figure 1. Here for each patient, we exclude readmissions of ICU stays, which can prevent possible information leakage in subsequent analysis. And we choose the prediction time point as the first 48 hours into the first ICU stay because empirical assessment shows that it is impossible to predict ICU mortality accurately without enough data accumulated.

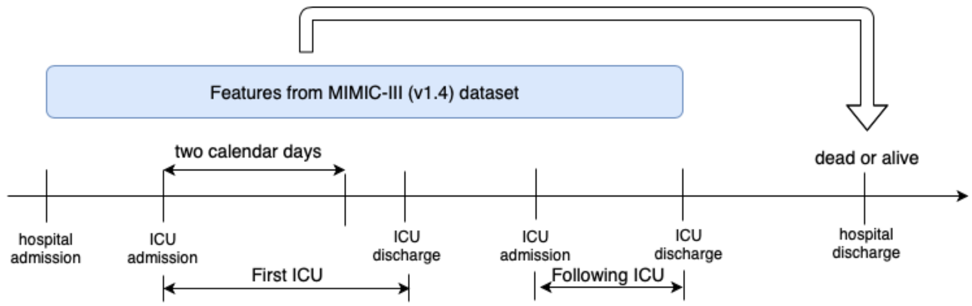

Figure 1: ICU mortality prediction problem

5.2 Cohort Selection

We use two sets of inclusion criterion to select the ICU stays. First, as mentioned in Section 5.1, we exclude readmissions of ICU stays. Second, we choose ICU stays which meet the following criteria: age of patient $\geq$ 16 years at the time of ICU admission and the ICU stay is longer than 48 hours.

5.3 Data Cleaning

Due to noice, missing values, outliers or incorrect records, the data extracted from MIMIC-III database has lots of erroneous entries. Therefore we need to identify and handle these inconsistent or erroneous records. First, we observed that there is inconsistency in the measure units of some variables. For example, the body temperature are measured in either Fahrenheit or Celsius units. Second, some numerical values are missing or recorded as error texts. Third, some variables have multiple values recorded at the same time. We addressed these issues by following procedures.

- To handle inconsistent units in body temperature records, we represent all data in Fahrenheit unit.
- For missing records, there are two circumstances. First, if the record is only missing occasionally with 48 hours, we do forward imputation and backward imputation. Second, if there is no such data for a period of 48 hours, we take the average value of that variable of all patients.
- For multiple records of the same variable in an hour, we randomly pick one value as there are vey small changes.

5.4 Feature Selection and Extraction

We extract data from following tables: *admissions*, *services*, *outputevents*, *chartevents*, *icustays*, *labevents* and *diagnoses_icd*, etc, because they provide the most relevant clinical features of ICU stays. To enable an exhaustive feature map which can measure the severity of disease for patients admitted to ICUs efficiently, we select the same set of features which is used in the calculation of the SAPS-II score which consists of the 17 features. Table 2 lists all the 17 processed features and their corresponding entries in the MIMIC-III database tables.

Table 2: 17 features used in SAPS-II scoring system

Features	ItemID	Item Name	Table
glasgow coma scale	723	GCSVerbal	chartevents
	454	GCSMotor	chartevents
	184	GCSEyes	chartevents
	223900	Verbal Response	chartevents
	223901	Motor Response	chartevents
	220739	Eye Opening	chartevents
systolic blood pressure	51	Arterial BP [Systolic]	chartevents
	442	Manual BP [Systolic]	chartevents
	455	NBP [Systolic]	chartevents
	6701	Arterial BP # 2 [Systolic]	chartevents
	220179	Non Invasive Blood Pressure systolic	chartevents
	220050	Arterial Blood Pressure systolic	chartevents
heart reate	211	Heart Rate	chartevents
	220045	Heart Rate	chartevents
body temperature	678	Temperature F	chartevents
	223761	Temperature Fahrenheit	chartevents
	676	Temperature C	chartevents
	223762	Temperature Celsius	chartevents
pao2 / fio2	50821	PO2	labevents
	50816	Oxygen	labevents
	223835	Inspired O2 Fraction (FiO2)	chartevents
	3420	FiO2	chartevents
	3422	FiO2 (Meas)	chartevents
	190	FiO2 set	chartevents
urine output	40055	Urine Out Foley	outputevents
	43175	Urine	outputevents
	40069	Urine Out Void	outputevents
	40094	Urine Out Condom Cath	outputevents
	40715	Urine Out Suprapubic	outputevents
	40473	Urine Out IleoConduit	outputevents
	40085	Urine Out Incontinent	outputevents
	40057	Urine Out Rt Neophrostomy	outputevents
	40056	Urine Out Lt Neophrostomy	outputevents
	40405	Urine Out Other	outputevents
	40428	Orine Out Straight Cath	outputevents
	40086	Urine Out Ureteral Incontinent	outputevents
	40096	Urine Out Ureteral Stent # 1	outputevents
	40651	Urine Out Ureteral Stent # 2	outputevents
	226559	Foley	outputevents
	226560	Void	outputevents
	226561	Condom Cath	outputevents

Table 2: 17 features used in SAPS-II scoring system

Features	ItemID	Item Name	Table
	226584	Ileoconduit	outputevents
	226563	Suprapubic	outputevents
	226564	R Nephrostomy	outputevents
	226565	L Neophrostomy	outputevents
	226567	Straight Cath	outputevents
	226557	R Ureteral Stent	outputevents
	226558	L Ureteral Stent	outputevents
	227488	GU Irrigant Volume In	outputevents
	227489	GU Irrigant/Urine Volume Out	outputevents
serum urea nitrogen level	51006	Urea Nitrogen	labevents
white blood cells count	51300	WBC Count	labevents
	51301	White Blood Cells	labevents
serum bicarbonate level	50882	BICARBONATE	labevents
sodium level	950824	Sodium White Blood	labevents
	50983	Sodium	labevents
potassium level	50822	Potassium, whole blood	chartevents
	50971	Potassium	chartevents
bilirubin level	50885	Bilirubin Total	labevents
age	-	intime	icustays
	-	dob	patients
immunodeficiency syndrome	-	icd9_code	diagnoses_icd
hematologic malignancy	-	icd9_code	diagnoses_icd
metastatic cancer	-	icd9_code	diagnoses_icd
admission type	-	curr_service	services
		ADMISSION_TYPE	admissions

The 17 features in Table 2 can be divided into two categories: non-sequential features such as chronic diseases, admission types and age, and sequential features that represent time-series patient characteristic such as blood pressure, heart rate and body temperature, etc. For each patient admitted into ICU, each time-series feature is sampled every 1 hour so that the time-series information for each patient is represented by a 48×13 matrix.

5.5 Deep Learning Models

Recently, deep learning models have demonstrated promising performance in mortality prediction of ICU patients. Deep learning models consist of a layered, hierarchical architectures of neurons for learning and representing data. One of the main advantages of the deep learning models is its ability to automatically learn good feature from raw data and thus significantly reduces the effort of handcrafted feature engineering.

Some recent works have demonstrated that deep learning models achieve state-of-the-art performance in health-related fields, such as ICU mortality prediction[8], phenotype discovery[15] and disease prediction[16]. In this work, we applied RNN model which is appropriate for modeling sequence and time-series data.

5.5.1 Implementation Details

Here we implemented a basic 3-layer LSTM model in PyTorch[17]. The model is trained with Adam optimizer with learning rate of 0.001. The batch size is 32 and max epoch number is 10. Early stopping with best weight is applied during training. We randomly sample 20% of the patients for the test set, and 20% for the validation set. The remaining 60% of the patients are used during training.

5.5.2 Evaluation Metrics

As the ICU mortality is a binary classification problem, we choose *Precision*, *Recall*, *F1* and *AUC* to evaluate our models.

6 Prediction Results

In this paper, we compared the RNN-LSTM-based model with a logistic regression model with L2 regularization. The input feature value of the logistic regression model are those measured at the last hour of the 48 hours window. The metrics results of the basic LSTM model and the comparison logistic regression model are reported in Table 3 and the receiver operating characteristic (ROC) curve of the RNN-LSTM model is in Figure 2. From Table 3, RNN-LSTM model consistently outperforms the baseline logistic regression model. On the test dataset, the AUC of the RNN-LSTM model is higher than logistic regression by 4%. Figure 2 shows that a basic LSTM model can achieve good performance in mortality prediction, which implies a promising future of deep learning models in health-related projects.

Table 3: Metrics evaluation of different models.

Model	Precision	Recall	F1	AUC
RNN-LSTM model	0.620	0.711	0.662	0.600
Logistic Regression	0.610	0.650	0.620	0.560

7 Discussion

In this study of nearly 40,000 ICU stays, we found that a RNN-LSTM based model is able to take advantage of sequential nature of time-series features to achieve a higher accuracy in identification of patients at high risk of death, comparing to some common approaches such as logistic regression. This finding demonstrates that it is important to perform sequential analysis of clinical data, because an abnormal change in a key physiologic measurement may signal potential clinical deterioration, even if the absolute value is not in a critical zone yet. Our research work sheds new light to empower deep learning into health-related project.

Although in this work we used the same features as SAPS-II calculation, it is worthwhile to mention that identifying efficient features to predict ICU survival is not trivial. This therefore remains to be an important direction of future research.

This present study also has several other limitations. First, the MIMIC-III dataset is collected from a single intuition so our finding may therefore not be generalizable to other clinical or geographic settings. The data

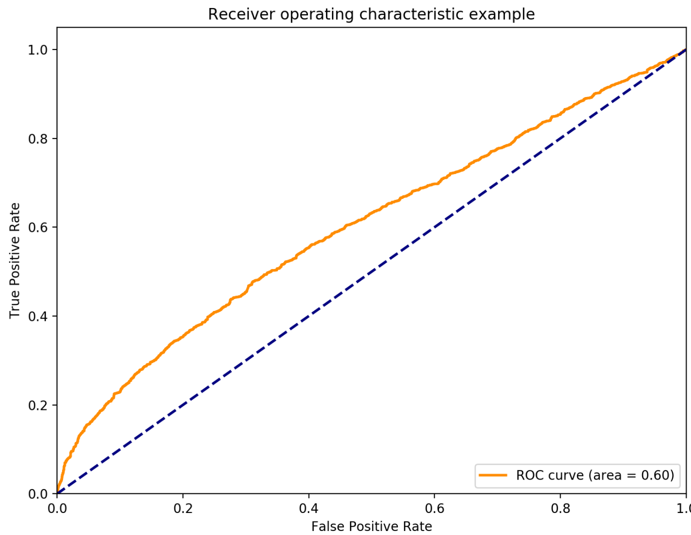

Figure 2: The ROC of the RNN-LSTM model.

from a medical ICU may not be applicable to other ICU categories. Second, some other data which are available from the MIMIC-III dataset, such as fluid balance and monitor data, have not been incorporated into our model. Future work will focus on aggregating these additional data and quantifying their impact on prediction accuracy. Third, the RNN-LSTM model implemented in this work has only three layers which lacks capability to efficiently capture both sequential and non-sequential features.

8 Conclusion

To conclude, in this work we propose to apply deep learning models into mortality prediction of ICU patients on MIMIC-III (v1.4) dataset. We preprocess data and extract features which have been used in SAPS-II. These features includes both sequential and non-sequential data which better reflects patients' psychological conditions. Then we implement and train a basic RNN-LSTM model, and compare its perdiction performance with that of a logistic regression model. Our result shows that the basic RNN-LSTM model can stably exceed the accuracy of a ”traditional” logistic regression model. The significance of our deep learning model includes: 1) by effectively capturing fluctuations in time-series features, it could give clinicians an early sense of the patient's mortality status; and 2) it could be used to help allocate ICU resources more efficiently.

In the future, our work can be extended in several directions, for example: 1) more sophisticated data preprocessing steps and deep learning models will be conducted to capture the characteristics of the massive MIMIC-III datasets; and 2) more extensive ICU datasets will be employed to evaluate and improve our models.

References

[1] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[2] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[3] GS Doig, KJ Inman, WJ Sibbald, CM Martin, and JM Robertson. Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 361. American Medical Informatics Association, 1993.

[4] C William Hanson and Bryan E Marshall. Artificial intelligence applications in the intensive care unit. *Critical care medicine*, 29(2):427–435, 2001.

[5] Álvaro Silva, Paulo Cortez, Manuel Filipe Santos, Lopes Gomes, and José Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artificial intelligence in medicine*, 36(3):223–234, 2006.

[6] Romain Pirracchio. Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sicula) project. In *Secondary Analysis of Electronic Health Records*, pages 295–313. Springer, 2016.

[7] Joon Lee, David M Maslove, and Joel A Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PloS one*, 10(5):e0127428, 2015.

[8] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376, 2017.

[9] Filip Dabek and Jesus J Caban. A neural network based model for predicting psychological conditions. In *International conference on brain informatics and health*, pages 252–261. Springer, 2015.

[10] 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, 2018.

[11] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[12] Wendong Ge, Jin-Won Huh, Yu Rang Park, Jae-Ho Lee, Young-Hak Kim, and Alexander Turchin. An interpretable icu mortality prediction model based on logistic regression and recurrent neural networks with lstm units. In *AMIA Annual Symposium Proceedings*, volume 2018, page 460. American Medical Informatics Association, 2018.

[13] Yao Zhu, Xiaoliang Fan, Jinzhun Wu, Xiao Liu, Jia Shi, and Cheng Wang. Predicting icu mortality by supervised bidirectional lstm networks. In *AIH@ IJCAI*, pages 49–60, 2018.

[14] Hanzhong Zheng and Dejia Shi. Using a lstm-rnn based deep learning framework for icu mortality prediction. In *International Conference on Web Information Systems and Applications*, pages 60–67. Springer, 2018.

[15] Zhengping Che, David Kale, Wenzhe Li, Mohammad Taha Bahadori, and Yan Liu. Deep computational phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 507–516. ACM, 2015.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.