Mortality Prediction in ICU

SACHIN CHANDRASEKHAR, Online Master of Science in Computer Science
[1]Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract;
In recent years, there has been a significant increase in the adoption of digital health record systems in hospitals. In the United States alone, the number of non-federal acute care hospitals with basic digital systems increased from 9.4% to 75.5% over a 7 year period between 2008 to 2014[1]. EHR data from Hospital's have increased exponentially. In order for us to effectively analyze those datasets and utilize them, we will have to use advanced big data technologies because the sheer volume of data is going to be huge.Early mortality prediction of hospitalized patients is important for assessing the severity of illness and deciding the appropriate treatment and interventions required. Several severity scoring models have been developed over the past decades, but still, early mortality prediction for intensive care unit patients remains a challenge.This study proposes a machine learning approach to address the task of predicting in-hospital mortality in the early stage of ICU stay using a 6-hour timeframe window for the first 3 timeframe windows which are very critical. The model would be useful to promptly identify high-risk patients who might be dead within hours or days since admitted to ICU. Data were extracted for the first 6-hour timeframe window for the first 3 windows since ICU admission for each ICU stay from MIMIC-III database. The extracted features includes 45 physiological variables such as heart rate, blood pressure, Glasgow coma scale, and demographic features such as gender, age, ethnicity. Although there are many missing values in the first 6-hour of ICU data, i have demonstrated in the study a feasible and novel framework to predict in-hospital mortality and death time.

Introduction;
Health care is considered to be one of the most exciting frontiers in data mining and machine learning.Patients who are typically admitted to an intensive care unit (ICU) suffer from serious emergency illness and are at a high risk of dying. Hence, they need to be constantly treated and monitored to ensure that they recover.The aim of the project is to repeat and improve previous study[2] in predicting in-hospital mortality and also try and predict at the early stage of ICU stay. In a modern-day ICU, there are several types of devices such as heart monitors, ventilators, catheter, arterial lines, etc. that continuously keep track of a patient's health records. If there are any irregular fluctuations, the nurse or the doctor is immediately notified.In healthcare, estimating the risk of mortality can be very crucial in terms of sorting and allocating hospital resources in determining appropriate levels of care needed. Lack of mechanism to fetch realtime data and performance issues due to the sheer volume of data has acted as hindrance in the past in mortality prediction in the early stages.
An ICU provides intensive treatment for patients with severe and life-threatening illness and injuries who requires extra care and attention. ICU is one of the most expensive care because it requires high staff to patient ratio for intensive patient monitoring and complex treatment. ICS's normally generate a massive amount of electronic healthcare records which are useful in predicting patients' disease status and the amount of healthcare needed. The Medical Information Mart for Intensive Care III (MIMIC-III)[3]-[5] is a freely-accessible ICU database comprising de-identified EHR of over 60,000 ICU stays for around 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. The database consists of rich information about patients' demographic characteristics, such as gender, age, ethnicity, admission type, and various in-hospital measurements, laboratory tests, procedures and medication of ICU patients over the time. The database provides data from two electronic healthcare record systems, namely the Carevue (from 2001 to 2008) and MetaVision (from 2008 to 2012), which collect and store data differently. Over the past few decades, several ICU scoring systems[6],[7]have been developed for mortality prediction using rule-based method or data mining approach. Some standard scoring systems include Acute Physiology And Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS) and Sepsis-related organ Failure Assessment Score (SOFA). APACHE[8] is a severity scoring systems designed to provide a morbidity score for a patient. A predicted mortality can be derived from this score. SAPS[9] was designed to predict morbidity for a particular patient by comparing the outcome with other patients or a group of patients by comparing the outcome with another group of patients. SOFA[10] provides a daily score to track a person's status during an ICU stay to determine the extent of a person's organ function or rate of failure.
This study aims to investigate the use of machine learning in predicting mortality in the early stage of ICU admission.

The goal of the study is to provide clinicians with timely information that can enhance their understanding of a patient criticality and act as a flag for poor outcomes. In this study, i have tried to capture the data at the early stage of ICU admission by splitting data into a 6-hour timeframe window for the first 3 timeframe windows which are very critical in identifying high risk patients and ensuring that they receive immediate and appropriate care and treatment.

Approach:

Most models in the literature were designed for at least 24 hours or 48 hours after ICU admission to provide real-time or retrospective prediction on patients' mortality. In this study, i propose to a two-phase model framework to address the task of predicting the mortality and also death hours in the early stage of ICU stay using 6-hour timeframe window for the first 3 windows. If a patient is predicted dead in the first phase, the model would further provide an estimate of death hours since ICU admission in the second phase. The aim is to identify high-risk patients who might be dead within hours or days since ICU admission. Data were extracted for the first 6 hours, 12 hours or 18 hours since ICU admission for each ICU stay from MIMIC-III database. Multiple models were trained on the extracted features of the study population for the specified timeframe. The model results were then compared and discussed in the study.The Model with the best performance will be chosen for each time frame window.With low cost high performance infrastructure setup like AWS Athena, near real time prediction and at the early stages of the ICS stay is possible. Apart from the primary metric of AUROC, accuracy, confusion matrix, precision, recall and F1-score are some of the other metrics which will be used to provide a full picture of model performance.

Exploratory Data Analysis:

MIMIC-III is a large, publicly available database comprising de-identified health-related data associated with approximately 60K admissions of patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.MIMIC supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. It is notable for three factors:

1) Dataset is publicly and freely available.
2) Dataset encompasses a diverse and large population of ICU patients.
3) Dataset contains high temporal resolution data including lab results, electronic documentation, bedside monitor trends and waveforms.

The current version of the database is v1.4.MIMIC-III v1.4[11] was released on 2 September 2016. It was a major release enhancing data quality and providing a large amount of additional data for Metavision patients. This will be the dataset which will be used in this project.Please find the schema definition of the MIMIC-III v1.4 dataset in the References[12]. The original dataset in MIMIC-III database consists of 61,532 distinct ICU stays of 46,520 unique patients. To form our study population, we will excluded ICU stays less than one hour to remove fuzziness in data due to unusual short stays and only consider adult patients with age between 16 and 89. The final study population will be 49,632 ICU stays out of 36,343 patients. Below table provides summary statistics of the study population.

Variables	Statistics
Age	Mean 62.61%
Gender F:M	42.21%: 57.79 %
ICU length of stay (days)	4.41
Hospital length of stay (days)	11.37
In-hospital mortality ratio	11.62%

Among 49,633 ICU stays, there are 5,766 in-hospital mortality. After filtering out the ICU stays with negative death time since ICU admission (which is likely an administrative error resulting in an incorrect ICU admission or incorrect death time), 5,718 in-hospital mortality were resulted. The average death time since ICU admission is 9.57 days, maximum death time is 206.38 days and minimum death time is 0 day

Feature Extraction:

My study aims to predict mortality and death time in the early stage of ICU stay. For each ICU stay, i have extracted data from the first 6 hours, 12 hours and 18 hours since ICU admission. First step was to extract patients' features in every hour of their ICU stay until they were discharged from the ICU and then to aggregate the feature values in the specified timeframe. There are altogether 123 extracted features covering 5 static variables and 40 physiological variables. Reference has been made to this code repository[13] when the features were constructed.
The static variables includes admission type, total number of previous and current ICU stays, and demographic features such as age, gender and ethnicity. The temporal data of physiological variables includes patients' vital signs such as heart rate and blood pressure, Glasgow coma scale, blood gases and chemistry values, laboratory results and urine output. Most of the temporal variables were aggregated by maximum, minimum, and average during the specified timeframe (6 hours, 12 hours or 18 hours), except that urine output was aggregated by sum. Table 2 provides the list of extracted features used in the study.

Categories	Variables	Extracted features
Demographic and static features	Age, Gender, Ethnicity, Admission type, Number of ICU stays	N/A
Vital signs	Heart rate, Systolic blood pressure, Diastolic blood pressure, Mean blood pressure, Respiratory rate, Temperature, Peripheral capillary oxygen saturation, Glucose	Minimum, Maximum, Mean
Glasgow coma scale and chemistry values	Glasgow coma scale (GCS), GCS components (motor, verbal, eyes) Partial pressure of oxygen, Partial pressure of carbon dioxide, pH, Ratio of partial pressure of oxygen to fraction of oxygen inspired, Total carbon dioxide concentration	Minimum, Maximum, Mean
Lab results	Anion gap, Albumin,Immature band forms, Bicarbonate, Bilirubin, Calcium, Creatinine, Chloride, Hematocrit, Hemoglobin, Lactate, Platelet, Potassium, Partial thromboplastin time, International Normalized Ratio, Sodium, Blood urea nitrogen,White blood cell count	Minimum, Maximum, Mean
Urine output	Urine output	Sum

Feature selection & Pre-Processing:

The average, minimum and maximum value of the selected features in the first 6 hours of ICU admission were computed and included as predictive features. Whether the patient died in the hospital after the ICU stay, or hospital mortality was used as a response variable.A total of 123 features were retained for further analysis.Data were assessed for missing values, and percentage of missing values for each feature by the response category (Deceased: Discharged) was calculated. Missing values were handled by setting up threshold at two levels: features and raw levels. Features with more than 60% missing were dropped.Missing values in the feature matrix were imputed with the column means.After missing values were handled, the data showed that only about 18% of patients died in the hospital. The ratio between our two classes, died in hospital to discharged from the hospital, is 18:82.The under-sampling technique was used to balance the data.Random under-sampling which allows to randomly select an equal number of records from the majority (discharged) class was used to balance the data.Continuous values such as age, weight, and

laboratory measures were scaled to a range between 0 and 1, as follows:

$$x_s = \frac{x_i - min(x)}{max(x) - min(x)}$$

Where xs is a scaled value, xi is the original value, and min(x) and max(x) are the minimum and maximum values of the particular feature. Scaling prevents variables from over dominating in the supervised learning algorithms.After data preparation is completed by addressing missing value, imbalance data and normalization , the final dataset was split into a 20/80 split; 20% for testing and 80% for training sets.

Enviornmental Setup:

The study consists two stages of implementation:
(i) Feature Engineering using spark and AWS Glue.To convert the original MIMIC-III CSV dataset to Apache Parquet[14], a data transformation job using AWS Glue was used. The CSV files were converted into Apache Parquet format ,which is optimized for modern data processing technologies such as Athena ,using AWS Glue.Apache Parquet stores data by column, so queries that fetch specific columns can run without reading the whole table.The data was then stored as tables in AWS Athena from where the data can be easily accessed by Python jupyter notebooks which were setup in AWS Amazon Sagemaker
(ii) Machine learning using Python 3.6 on AWS Amazon Sagemaker notebook instance 'ml.c5.4xlarge' was used to host and implement the ML model code. We also used Python packages such as Pandas ,Scikit-learn and other packages for efficient model testing, hyperparameter tuning and model evaluation

Model Architecture

I propose a two-phase model framework to predict in-hospital mortality and death time in hours. In Phase 1, a binary classifier was trained using the 123 extracted features identified to predict in-hospital mortality. In Phase 2, a multi-class classifier will be trained on the same set of extracted features to predict death time in hours since ICU admission for the predicted dead patients in Phase 1.
In Phase 1 the study response variable, mortality, is a binary variable. As such, we used a set of supervised classification algorithms to develop different predictive models. These included Logistic Regression, Support Vector Machine (SVM), Decision Tree, Neural Network and Random Forest. Finally, each model was tested with the test set, and performance metrics such as accuracy, precision, recall, F-score and AUC- ROC were calculated at the multiple time frame window like 6 hr ,12 hr and 18 hr.For each time frame window, the best model will be selected. Phase 2 Implementation is currently in progress where the data will be classified into 3 categories as below:

Class	Description	Number of Patients
Class 0	Death time $<=$ 1 day	865
Class 1	1 Day $<$ Death time $<=$ 1 week	2496
Class 2	Death time $>$ 1 week	2354

A multi-class classifier will be trained on the training set to predict the death time label using 6-hour, 12-hour and 18-hour data respectively. The model performance of the best classifier resulted from the grid search under 6-hour, 12-hour and 18-hour scenarios will be then compared and evaluated on the test set.

Experimental Results:

For Phase 1 model results, the below tables compares the model performance of the random forest classifiers separately trained using 6-hour, 12-hour and 18-hour data in Phase 1. The results show that data aggregation over wider timeframe gives slightly better result. Also we could see that for both 6-hr and 12-hr dataset, Random Forest classifier was the best predictor among all the chosen classifiers. But for 18-hour data set Neural network outperformed Random Forest classifier.The increase in the data would be one of the reasons for this increase in performance compared to 6 and 12 hour datasets.

Phase 1 Performance Metrics for 6 hr Dataset:

Algorithm	Accuracy	AUC	Precision	Recall	F1-score	Confusion matrix
Logistic Regression	0.94305	0.94907	0.99896	0.89934	0.94654	[840 108] [1 965]
Decision Tree	0.96238	0.96289	0.97722	0.94305	0.94969	[898 50] [22 944]
Random Forest	0.96238	0.96445	0.99378	0.93567	0.96385	[882 66] [6 960]
SVM	0.94200	0.94801	0.99792	0.89841	0.94556	[839 109] [2 964]
Neural Net	0.96133	0.96340	0.99275	0.93469	0.96285	[881 67] [7 959]

By Comparing the results for various models, we can see that for the first 6 hour data Random Forest Classifier has the best performance.

Phase 1 Performance Metrics for 12hr Dataset:

Algorithm	Accuracy	AUC	Precision	Recall	F1-score	Confusion matrix
Logistic Regression	0.94150	0.94693	0.99905	0.89491	0.94412	[956 124] [1 1056]
Decision Tree	0.95975	0.96022	0.97824	0.94257	0.96007	[1017 63] [23 1034]
Random Forest	0.96022	0.96174	0.99148	0.93238	0.96102	[1004 76] [9 1048]
SVM	0.93869	0.94466	0.99905	0.89038	0.94159	[950 130] [1 1056]
Neural Net	0.95648	0.95928	0.99810	0.92059	0.95778	[989 91] [2 1055]

By Comparing the results for various models, we can see that for the first 12 hour data Random Forest Classifier has the best performance.

Phase 1 Performance Metrics for 18hr Dataset:

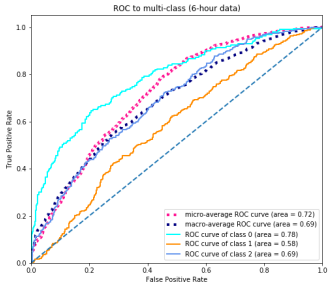
Algorithm	Accuracy	AUC	Precision	Recall	F1-score	Confusion matrix
Logistic Regression	0.95198	0.95748	1.0	0.91497	0.95560	[952 105] [0 1130]
Decision Tree	0.95930	0.95949	0.96637	0.95538	0.96084	[1006 51] [38 1092]
Random Forest	0.96799	0.96930	0.98849	0.95144	0.96961	[1000 57] [13 1117]
SVM	0.94878	0.95491	1.0	0.90982	0.95278	[945 112] [0 1130]
Neural Net	0.96982	0.97124	0.99115	0.95238	0.97137	[1001 56] [10 1120]

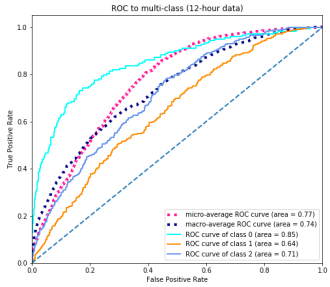
By Comparing the results for various models, we can see that for the first 18 hour data Neural Network Classifier has

the best performance.

As for Phase 2 model results, Table 5 compares the model performance of random forest multi class classifiers separately trained using 6-hour, 12-hour and 18-hour data in Phase 2. The results show that data aggregation over 18-hour timeframe gives slightly better result than those of 6-hour and 12-hour. Specifically, random forest classifier trained on 6-hour, 12-hour, 18-hour ICU data have micro-average AUROC 0.77, 0.79, 0.82 on the test set respectively. Figure 4 compares the micro-averge, macro-average and ROC curves for individual classes.

Models	Micro-average AUROC on test set	Macro-average AUROC on test set
Random forest multi-class classifier trained on 24-hour ICU data	0.72	0.69
Random forest multi-class classifier trained on 24-hour ICU data	0.77	0.74
Random forest multi-class classifier trained on 24-hour ICU data	0.79	0.76





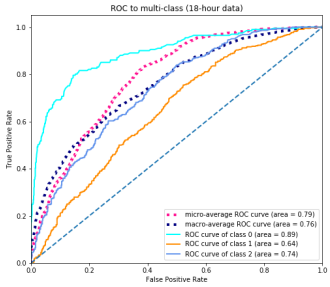


Table 1: ROC curves of random forest classifier evaluated on test set in Phase 2

Discussion:

In this report, the challenging problem of predicting early mortality in ICU using a machine learning approach is being addressed. Multiple classifiers were used in this prediction problem and compared them with standard machine learning matrices in order to identify the best Model. We could see that by combining static features such as patients' demographic information and dynamic features such as physiological variables measured in ICU, we could train an effective model to predict in-hospital mortality in the early stage of ICU stay. In Phase 2 a multi-class classifier was implemented in-order to predict death time for in-hospital mortality.As the goal of this study is to derive important takeaways for clinicians to use in the clinical setting, not only the accuracy of the model but also the interpretability of the model was equally important. A study[15] that explored early hospital mortality prediction using vital signs showed that Decision Trees provided the best interpretability and best accuracy out of the eight different learning strategies they tested. On the other hand,Ramon et al 2007[15] reported that naive Bayesian networks and Naive Bayesian networks performed better than Decision Tree. Similarly, Pirracchio et al.[16] 2015 reported that a Bayesian Additive Regression Tree (BART) is the best candidate, while Random Forests (RF) outperformed all other candidates when using transformed variables. Our results confirm some of these findings, the random forest was the best classifier in terms of accuracy.

Conclusion

MIMIC-III is a rich source of EHR comprising a diverse range of static data and high temporal data for ICU patients. We have proposed a two-phase models to predict in-hospital mortality and death time in hours since ICU admission for dead patients. The experimental results show that although the models trained on the 18-hour ICU data set give slightly better performance, the first 6 hours of ICU data already provides us enough information for in-hospital mortality prediction and a rough estimate of death hours since ICU admission. The proposed framework provides a base to promptly identify high-risk patients who might be dead within hours or days since ICU admission in their early stage of ICU stay, hence allow better resource allocation in the first few critical hours of ICU stays. Apart from this, our model framework provides a base for potential improvement. One possible way to enhance the performance is to fit the model with time-series data instead of aggregation of dynamic features.

References

1. http://www.ncmedicaljournal.com/content/77/2/112.full
2. H. Harutyunyan, H. Khachatrian, D. C. Kale, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. arXiv preprint arXiv:1703.07771, 22 Mar. 2017.
3. A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. Sci Data, 3:160035, 24 May 2016.
4. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/content/101/23/e215.full]; 2000 (June 13).
5. Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. The MIMIC Code Repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association (2017)
6. Soares M, Fontes F, Dantas J, Gadelha D, Cariello P, Nardes F, et al. (2004). Performance of six severity-of-illness scores in cancer patients requiring admission to the intensive care unit: a prospective observational study. Crit Care. 8 (4): R194203. doi:10.1186/cc2870. PMC 522839Freely accessible.
7. Strand K, Flaatten H (2008). Severity scoring in the ICU: a review. Acta Anaesthesiol Scand. 52 (4): 46778. doi:10.1111/j.1399-6576.2008.01586.x.
8. KnausWA,DraperEA,WagnerDP,ZimmermanJE(1985).APACHEII:aseverityofdiseaseclassificationsystem. Critical Care Medicine. 13 (10): 81829. doi:10.1097/00003246-198510000-00009. PMID 3928249. (This is the first published description of the APACHE II scoring system)
9. Jean-Roger Le Gall, MD; Stanley Lemeshow, PhD; Fabienne Saulnier, MD. (1993). A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA. 1993;270:2957-2963 This is the first published description of the scoring system
10. Vincent JL, Moreno R, Takala J, Willatts S, De Mendona A, Bruining H, Reinhart CK, Suter PM, Thijs LG. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med 1996 Jul;22(7):707-10.
11. https://mit-lcp.github.io/mimic-schema-spy/
12. https://physionet.org/content/mimiciii/1.4/
13. Github Repository available at https://github.com/alistairewj/mortality-prediction/tree/master/queries
14. https://aws.amazon.com/blogs/big-data/perform-biomedical-informatics-without-a-database-using-mimic-iii-data-and-amazon-athena/
15. Sadeghi, Reza, Tanvi Banerjee, and William Romine. 2018. “Early Hospital Mortality Prediction Using Vital Signals,” March. https://arxiv.org/abs/1803.06589.
16. Ramon, Jan, Daan Fierens, Fabian Guiza Grandas, Geert Meyfroidt, Hendrik Blockeel, Maurice Bruynooghe, and Greta Van den Berghe. 2007. “Mining Data from Intensive Care Patients.” Advanced Engineering Informatics 21: 243–56. https://doi.org/10.1016/j.aei.2006.12.002.
17. Pirracchio, Romain, Maya L. Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J. van der Laan. 2015. “Mortality Prediction in Intensive Care Units with the Super ICU Learner Algorithm (SICULA): A Population-Based Study.” The Lancet. Respiratory Medicine 3 (1): 42–52. https://doi.org/10.1016/S2213-2600(14)70239-5.