International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

MORTALITY PREDICTION IN THE ICU UTILIZING TOPIC MODEL AND BURSTINESS WITH MACHINE-LEARNING TECHNIQUES

Purshotam Yadav
Georgia Institute of Technology Atlanta, Georgia, USA

Swati Prasad
Georgia Institute of Technology, Atlanta, Georgia, USA

Raj Vansia
Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract

The study of patients in Intensive Care Units (ICUs) is a crucial task in critical care research. The mortality study of ICU patients is of particular interest because it provides useful indications to healthcare institutions for improving patients experience, internal policies, and procedures. This paper utilizes the MIMIC III dataset to develop mortality prediction for ICU patients. Our approach is to use topic modeling on clinical notes to extract features related to morbidity using LDA. Additionally, this paper applies a burstiness algorithm to identify topics that are trending within a given period of time which can signal an increased level in morbidity. These features will be used to train our model utilizing Logistic regression, Neural Network, Linear SVM and Non-Linear SVM. Receiver operating characteristic curves areas for predicting mortality was approximately 74%for SVM using LDA topics, burstiness parameter, and baseline features.

Keywords; mortality prediction, topic modeling, morbidity features, LDA (Latent Dirichlet Allocation), machine learning, ICU

I. INTRODUCTION

Modern electronic health-care records contain an increasingly large amount of data, and the ability to automatically identify the factors that influence patient outcomes. This paper examined the use of latent variables model to decompose free-text hospital notes into meaningful features and the predictive power of these features for patient mortality. This paper considers two prediction regimes

- Base line prediction: which used structured data available on admission
- Retrospective outcome prediction: which used all clinical text generated from hospital stay to supplement the baseline features

In general, latent topic features were more predictive than structured features and a combination of the two performed the best. Combining features extracted from the note notations with standard physiological measurements results in a more completed representation of patient's physiological state thus affording improve outcome prediction. Several severity scores have been developed with the objective of predicting hospital mortality from baseline patient characteristics, defined as measurements obtained within the first 24h after ICU admission. Severity scores, such as the SAPS II (Simplified Acute Physiology ScoreII) is the most widely used in clinical practice were subsequently developed using statistical modeling techniques.
This paper proposes different approach burstiness for studying the mortality rate of patients in ICU that is ortho gonal to the traditional analysis based on the Length of Stay (LOS). Our approach

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

is motivated by the fact that the temporal distribution of medical events may carry information about the severity of the illness. Intuitively, high values of burstiness indicate the presence of rapidly occurring clinical events.
In this thesis, this paper investigates different approaches to predicting mortality for critical care patients. Using over 20 variables to characterize a patient's stay, this paper compares a variety of different predictive modeling techniques. This paper considers commonly used linear and non-linear classification methods such as Artificial neural networks (ANN), support vector machine (SVM).Predicting mortality in patients hospitalized in intensive care units (ICU) is crucial for assessing severity of illness and adjudicating the value of novel treatments, interventions and health care policies.

II. LITERATURE SURVEY

Mortality prediction has been widely studied utilizing the MIMICII&III (Medical Information Mart for Intensive Care) databases. Common approaches have been logistic regression yielding a AUROC of 0.848 achieved by Fialho. Dybowski has demonstrated the use of artificial neural networks (ANNs) to achieve an AUROC score of 0.857. ANNs have showed flexibility in modeling patient physiology where non-linear techniques could not.
Bonomi and Jiang have utilized burst in ess of clinical events to help predict the mortality of patients. This accounts for clinical events that occur close to one another. The assumption is that an increased frequency of clinical events in a given period of time indicates there is an increase in morbidity.

Most approaches have utilized physiological waveform data and do not take into account other sources of data available. Instead of just utilizing physiological data researchers have used multi modal data, data from different sources such as genetic information. For example, genomic data and microbiology lab reports have been used to improve mortality prediction. This approach produced an AUROC score of 0.79 compared to traditional approaches which have yielded 0.69.
Free text mining has been used to help augment prediction by learning topics from unstructured clinical notes. Several recent works have used information from clinical notes in their model formulations and This paper are going to use existing preliminary result as an guideline. Saria combined structured physiological data with concepts from the discharge summaries to achieve a patient outcome classification F1 score of 88.3 with a corresponding reduction in error of 23.52%. Topic modeling was used to analyze clinical notes and Ghassemi has demonstrated the use of topic models created using (Latent Dirichlet Allocation) LDA with unstructured clinical notes and achieved the following mortality prediction retrospective models 0.96, 0.82, and 0.81 for in-hospital, 30 day, and 1-year.

III. SOFTWARE ARCHITECTURE

Baseline features are extracted from the database for every patient (e.g. age, sex, admitting SAPS-II score)

Each patient's burstiness parameter is calculated from its length of stay using formula as : $B = \sigma-\mu /\sigma+ \mu$ with standard deviation σ and mean μ. Then, the burstiness parameter B has values ranging between−1 and 1

Each patient's de-identified clinical notes are used as the observed data in an LDA topic model, and atotal of 50 topics are inferred to create the per-note topic proportion matrix. The topic number of 50 was chosen because that is what is assumed to get a unique distribution of different disease topics. Also, 50topics were chosen by Ghassemi et al from their paper.

Depending on the model and time window being evaluated, subsets of the feature matrix v and matrix q is combined into an aggregate feature matrix.

A linear kernel SVM is trained with gamma (0.1) and kernel (Non-linear) to create classification boundaries

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

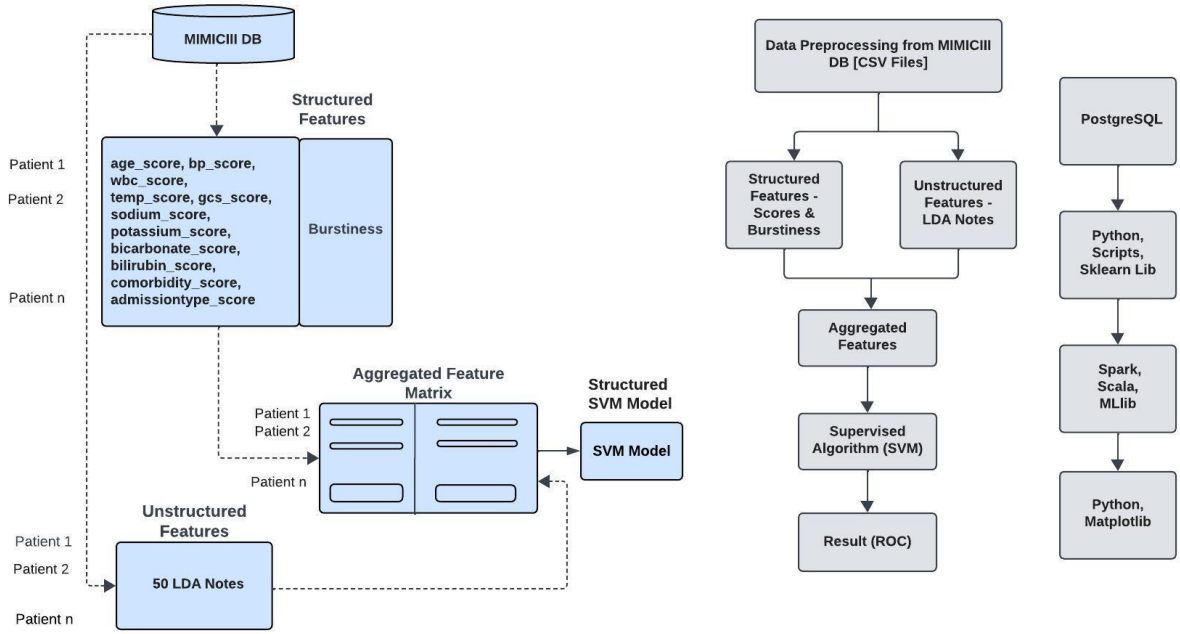


Fig1:Architecture

IV. DATA AND PREPROCESSING

- **Dataset Preparation and Preliminary Dataset**

Preparation: For the modeling and analysis, we relied on data extracted from MIMIC III database. While the MIMIC III database provides the rich collection of intensive care data. It is important to prepare data set to apply in machine learning models. These are the following steps to create variables and views from MIMICII tables

1. We installed MIMIC III database locally using given steps under https://mimic.physionet.org/tutorials/install-mimic-locally-ubuntu/

2. Saps II score were created using following reference https://github.com/MIT-LCP/mimic-code/blob/master/concepts/severityscores/sapsii.sql

3. This paper gets the baseline features variable using above reference and joining tables. This paper gets the mortality status with joining patients Table.

4. For burstiness, we used icu stays table. We created view locally with information number of days spent in ICU and joined with patients to get the mortality status.

5. To clean the data, we used python script based on sklearn libraries. Scores with na value was replaced with 0. Burstiness score were created for each patient applying mean and deviation on length of stay column.

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

Prelim datasets	Variables	MIMIC III tables
Baseline features	SapsII Score, age, sysbp, temp, PaO2FiO2, urine output, bun, wbc, potassium, sodium, bicarbonate, bilirubin,gcs, comorbidity, admission type_score	icuststay, chartevents, patients, admissions, diagnoses_icd, chartevents, icustays, services,
Notes	text, chartetime	Note events
Burstiness	Length of Stay, burstiness	patients, icu stays

Table 1 : Used Variables and Referenced MIMIC III tables

Using the table and variables above, this paper created the preliminary dataset. This process excludes patients with incomplete, missing data or poorly represented dataset. While most patients in MIMIC III dataset have only one visit of the hospital admission. The Mimic III dataset is used which includes 58,976 distinct hospital admissions, in which 46520 turns out to be Intensive care units (ICU) admissions with 43.8% (20399) of female and 56.1% (26121) to be male ICU admissions. This paper further worked on finding stats on mortality. This paper found 33.87% (15759) turns out to be mortality and 66.12% (30761) turns out to be alive from ICU admissions.

- **Extracting the variables**

First step is to translate the data in MIMICIII relational data base to suitable to form directly suitable for modeling. This paper extracted the required features and variables by creating views locally using tables referenced in Table 1. This paper further exported tables as csv and analyzed using python scripts.

- **Description of Extracted Variables**

Age, heart rate, systolic blood pressure, body temperature Glasgow Coma Scale, mechanical ventilation, PaO2, FiO2,urine output, BUN (blood urea nitrogen), blood sodium, potassium, bicarbonates, bilirubin, white blood cells, chronic disease (AIDS, metastatic cancer, hematologic malignancy) and type of admission (elective surgery, medical, unscheduled surgery).

SAPSII calculation is done further using the following scores

$$\begin{aligned} \text{SAPSII} = {} & \text{age\_score} + \text{hr\_score} + \text{sysbp\_score} + \text{temp\_score} \\ & + \text{PaO2FiO2\_score} + \text{uo\_score} + \text{bun\_score} + \text{wbc\_score} \\ & + \text{potassium\_score} + \text{sodium\_score} + \text{bicarbonate\_score} \\ & + \text{bilirubin\_score} + \text{gcs\_score} + \text{comorbidity\_score} + \text{admissiontype\_score} \end{aligned}$$

$$\text{SAPSII\_PROB} = \frac{1}{1 + e^{-(-7.7631 + 0.0737 \cdot \text{SAPSII} + 0.9971 \cdot \ln(\text{SAPSII}+1))}}$$

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

The Note events table in MIMICIII data base was used to capture the unstructured clinical notes. This text data was processed to remove any white spaces. Stop words were also removed from the text. Stop words include some of the following words("a","the","because")Each clinical note was treated as a document. This paper excluded discharge summary notes because they explicitly state the outcome. Patients were excluded if they had fewer than 100 non-stop words and under the age of 18.

Features	Overall Population	Dead at Hospital Discharge	Alive at Hospital Discharge
Age (newborns)	8100	0.827% (67)	99.17% (8033)
Age (adult)	50711	43.48% (22050)	56.51% (28661)
Gender (Female)	20399	35.46% (7235)	64.53% (13164)
Gender (Male)	26121	32.63% (8524)	67.36% (17597)
> Avg Length of Hospital Stay (11.32 days)	18129	12.76% (2314)	87.23% (15815)
> Avg Length of ICU Stay (4.93 days)	14059	17.11% (2406)	82.88% (11653)
> Avg SAPSII Score (32.71)	29203	58.3% (17026)	41.69% (12177)

Table 2 : MIMIC III Base features ~Mortality Chart

V. METHODS

This section summarizes our methodology for building and evaluating mortality prediction models.

Software/Hardware Environment Setup:

Hardware: The data was processed, and the model was run on a Mac in tosh OS with 16 GB RAM.
Software:

- Pre processing: Post gres SQ Lisusedin creating required tables for features. Further data processing and cleaning is done by using python with sklearn libraries.

- Processing: This paper used Scala & Sparkas big data environment to process the data. This paper used MLLIB SVM Supervised algorithm to predict the mortality.

1. This paper extracts the clinical baseline features including age, sex and SAPII score from the data base for every patient.

2. This paper also extracts each patient's de-identified clinical notes which includes 2,078,705 notes from the MIMIC-III database. Each patient's de-identified notes are used as input to LDA topic model and total of 50 topics are inferred to create the per-note topic proportion matrix. The process of generating topics from clinical notes is first aggregate all of the notes, these notes are converted from sentence to words using tokenizer, then stemming and lemmatization is applied then stop works are removed . Then for each patient's clinical note a topic model distribution is created that will be used as additional features.

3. This paper represents patient's longitudinal data as temporal sequences of medical events recorded overtime using burstiness parameter. This paper did the Burstciness Calculation using follows:
   This paper represents patient's longitudinal data as temporal sequences of medical events recorded over time. For example, in the case of a patient that has recurrent hospitalizations, each visit represents a medical event associated with a specific time stamp recorded by the healthcare institution. This paper defines temporal sequence X = {(Xi, ti)}n i=1 as an ordered sequence of n medical events. To capture information from the intervened time distributions, this paper uses the notion of burstiness. Let T = T1, T2...,Tn be an intervened time sequence with standard deviation σ and mean μ. Then, the burstiness parameter for T

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

is computed as: B = σ − μ /σ + μthe burstiness parameter B has values ranging between −1 and 1

B=−1 indicates a periodic sequence,
B=0 indicates a Poisson distribution in terevent sequence, B = 1 for extremely bursty sequences

4. Base line features, unsupervised LDA model output and burstiness parameter are combined in to an aggregate feature matrix.

VI. PREDICTION ALGORITHMS

a. The primary measure outcome was hospital mortality which came from ICU admission. 46520 turn out to be Intensive care units (ICU) admissions from MIMIC III. Individual mortality prediction for the SAPS II score was calculated as defined by its authors [1] using extracted features in section Data and Preprocessing.
b. A separate linear SVM, Logistic Regression, Non Linear SVM, Single Layer ANN & Multi Layer ANN is trained for Baseline prediction and each set of model features evaluated. This paper established a static baseline model using only structures features present at admission.
c. The Spark Mlib package of SVM library is used as supervised learning algorithm with 60:40 training and test ratio
d. Tuning of SVM: Gamma (0.1), iterations (400)
e. The Spark Mlib package was used to in put the clinical notes through the LDA model to generate a topic model. The model utilized 50 topics to be clustered.

Modelling:

1. Select a candidate set of features.
2. Define tuning parameters such as #of hidden layers, kernels and computational units in each supervised learning algorithm. (Start with a simple model.)
3. Obtain training and test data sets based on candidate features. Compare the models and pick the one with the best performance. If the result of the best performance model is satisfactory then stop.
4. Otherwise, if improvement is observed, go to step 3 to increase complexity of the model with more computational units and/or layers.
5. If no more improvement is observed go to step to redefine features(start all over again)

VII. DISCUSSION

The objective of this thesis was to build a mortality prediction model that could outperform current approaches. Key objective of this study was to get the predictive performance of the Supervised Algorithm uses SAPS II score. This comparison hinged on a variety of measures of predictive performance, described below. A mortality prediction algorithm is said to be adequately calibrated if predicted and observed probabilities of death coincide rather well. This paper assessed the predictability using the receiver-operating characteristic curve (AUROC).
This paper tried multiple supervised algorithms to get the best results. This is a binomial classification, This paper started with logistic regression and linear SVM. Logistic regression worked better as it adjust to linear anon linear functionality. Linear SVM didn't do well so This paper tried with non linear kernel to improve performance. This paper tried with single and multilayer ANN. There are numerous theoretical advantages of neural networks. Neural networks require no a priori assumptions or knowledge about the underlying frequency distribution (nonparametric); they have the capacity to model complex, nonlinear relationships; they do not require assumptions about the independence of variables and they are relatively robust and tolerant of missing data and input errors.
Topic modeling utilizing LDA was chosen to leverage clinical notes to generate extra features that can be used for prediction. Clinical notes are written by skilled clinical professionals that outline the health state of the patient during an ICU stay. The topics generated by the LDA model outline topics from the clinical notes utilizing unsupervised learning. Topic 1 shows a clustering of words

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

that create a topic related to cardiac health. These topics are used to add features by using the patient's clinical notes and indicating how the clinical note(document) is grouped to a topic.
In our evaluations, this paper considers two main variables: the length of stay (LOS) and burstiness parameter. This paper is interested in study the relationships between these variables and the mortality of patients in the MIMIC-III dataset. The dependency between these two variables is reported in Burstiness plot. From these results, this paper observes that the burstiness parameter provides orthogonal information about the mortality rate of the patients which is not directly correlated to the length of stay.

VIII. EXPERIMENTAL RESULTS

Initial ROC result of baseline, burstiness and Notes

Comparison between classifiers

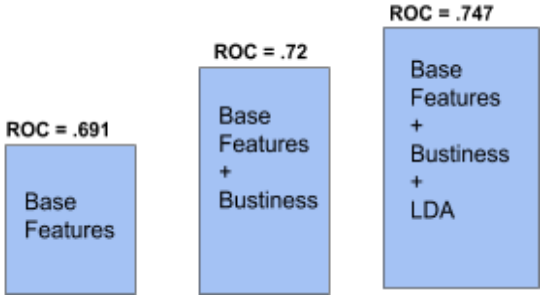
Classifiers	ROC Area
Single-layer ANN	.739
Two-layer ANN	.736
Linear SVM	.712
Non Linear SVM	.731
Logistic Regression	.72

Table 3: Analysis of Results

Topic	Keywords
Topic 1	heart, aortic, cardiac, ventricular, systolic, mitral
Topic 2	biopsy, cancer, chemo, tumor, metastatic, mass
Topic 3	cervical, fracture, lumbar, spine, ortho, bone
Topic 4	respiratory, airway, breathing, ventilation, lung

Top 4 topic models using LDA.

Non linear SVM Tuning
Tuning Parameter:
Gamma = 0.1
Iterations = 400
Training: Testing = 60:40 (43073:24613)



International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

Fig 2: ROC curves for Algorithms

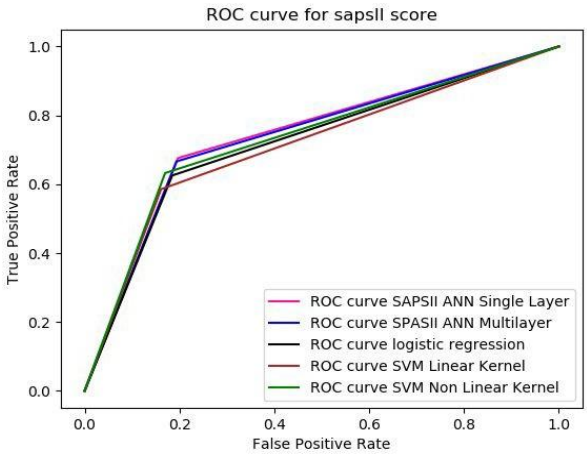
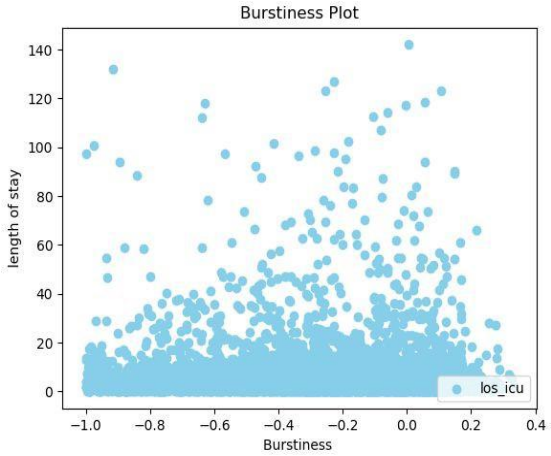


Fig 3: Burstiness Plot Result ROC = 0.747



IX. CONCLUSION

The objective of this thesis was to build a mortality prediction model that can outperform current approaches. This paper aimed to improve current methodologies in two keyways:

- By incorporating a wider range of variables and standard scores, notes. This paper introduced new parameter burstiness parameter for each patient to improve performance.

- By exploring different predictive modeling techniques beyond standard regression, linear SVM, nonlinear SVM, ANNs. This paper reached ROC .74 which is close to the ROC mentioned in Literature survey. However, this paper concludes that it is possible if care is taken to properly customize the model through extensive data preprocessing, integrating the note events & burstiness and tuning the algorithms.

X. FUTURE WORK

This paper achieved a ROC of 0.745, which is close to the existing ROC of 0.778. There is potential for improvement by considering the following:

- The paper utilized LDA notes, which include a large number of records. The running time could be optimized by using AWS for efficient processing of these notes.
- The results could be enhanced with more refined data and by employing different supervised algorithms.
- Further improvement could be achieved by applying a combination of machine learning algorithms.

International Journal of Core Engineering & Management
Volume-5, Issue-06, September-2018, ISSN No: 2348-9510

REFERENCES

1. Le Gall JR, Lemeshow S, Saulnier F (1993) A new simplified acute physiology score (SAPSII) based on a European/North American multicenter study. JAMA 270(24):2957–2963
2. Lehman LW, Saeed M, Long W,Lee J,Mark R. Risks tratification of ICU patients using topic models inferred from unstructured progress notes. In AMIA annual symposium proceedings 2012 (Vol. 2012, p. 505). American Medical Informatics Association.
3. Fialho AS, Celi LA, Cismondi F, Vieira SM, Reti SR, Sousa JM, Finkelstein SN. Disease-based modeling to predict fluid response in intensive care units. Methods of information in medicine .2013 Jun;52(06):494-502.
4. Dybowski R, Gant V, Weller P,Chang R. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. The Lancet. 1996 Apr 27;347(9009):1146-50.
5. Bonomi L, Jiang X. A Mortality Study for ICU Patients using Bursty Medical Events. InData Engineering (ICDE), 2017 IEEE 33rd International Conferenceon 2017 Apr19 (pp.1533-1540).IEEE.
6. Wiens J, Horvitz E, Guttag JV. Patient risk stratification for hospital-associated. diffasatime-series classification task. In Advances in Neural Information Processing Systems 2012 (pp. 467-475).
7. Saria S, Rajani AK, Gould J, Koller D,Penn AA. Integration of early physiological responses predicts later illness severity in preterm infants. Science translational medicine. 2010 Sep 8;2(48).
8. Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A, Szolovits P. Unfolding physiological state: Mortality modelling in intensive care units. InProceedings of the 20th ACMSIGKDD international conference on Knowledge discovery and data mining 2014 Aug 24 (pp. 75-84). ACM.
9. M. Ghassemi, M. Pimentel, T. Naumann, and T. Brennan, "A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data," presented at the AAAI Conference, Austin, TX, USA, Jan. 2015.
10. Johnson, J. Kramer, T. Clifford, L. P. Mark, and R. G. Mark, "Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 4, pp. 1097-1105, Jul. 2014.
11. R. P. Lippmann and D. M. Shahian, "Coronary Artery Bypass Risk Prediction Using Neural Networks," The Annals of Thoracic Surgery, vol. 63, no. 6, pp. 1635-1643, Jun. 1997. DOI: 10.1016/S0003-4975(97)00225-7
12. Secondary Analysis of Electronic Health Records, "Mortality Prediction in the ICU Based on MIMIC-II Results from the Super ICU Learner Algorithm (SICULA) Project," Chapter 20. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK543625/
13. Critical Care: Severity Scoring Systems," National Center for Biotechnology Information, U.S. National Library of Medicine. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK543625/#:~=Several%20severity%20scores%20have%20been,24%20h%20after%20ICU%20admission
14. To create view with SAPSII scores and other https://github.com/MIT-LCP/mimic-code/blob/master/concepts/severityscores/sapsii.sql