Sepsis Prediction in ICU patients

Dhruv Patel, Arti Chauhan, Thomas Weldon, Satish Gupta
Georgia Institute of Technology, GA
https://youtu.be/aly31d5nTvU

1. ABSTRACT

Accurate detection of sepsis-onset is a challenging problem as sepsis is a heterogeneous syndrome and its diagnosis involves considerable subjectivity. This issue is compounded by ever-changing clinical and regulatory guidelines. In this proposal, we discuss a system that will leverage time-variant feature-set to determine sepsis onset time, without explicitly relying on ICD codes. We will build upon prior work from ML-researchers and Physicians on this topic to create more informative feature-set to improve specificity and sensitivity of model.

2. INTRODUCTION / MOTIVATION

Sepsis is one of the leading factors of decomposition and mortality in hospitals in the United States. Every year, approximately 1.7M adults develop sepsis resulting in 270,000 deaths. Mortality rates are highest amongst patients who develop septic shock[1]. Despite the existence of several medical interventions which can reduce the mortality rate, symptoms which signal the onset of sepsis are often obscured due to comorbidity and other factors resulting in a failure to diagnose.

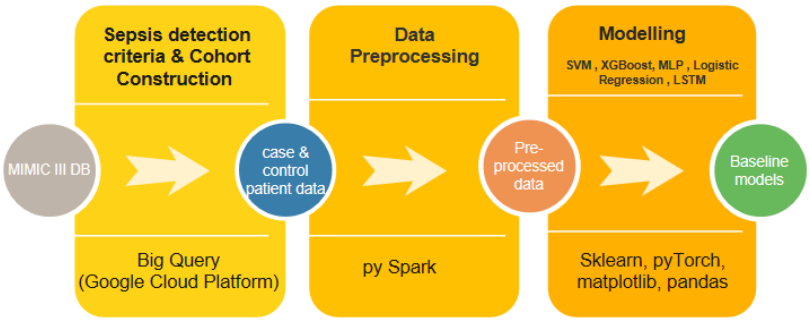
3. PROBLEM FORMULATION

Given the potential impact new machine-learning based models could have on both healthcare outcomes and hospitalization costs, we propose to add a clinical prediction benchmark task using data from the Medication Information Mart for Intensive Care (MIMIC-III) to predict, in near real time, the onset of sepsis. Once established, this baseline model will help the research community to evaluate the precision and sensitivity of new model, along with the performance gains of new big data processing frameworks.

Our over-arching goal is to develop a model that can predict the onset of sepsis with high accuracy in ICU (adult) patients *x* hours prior to clinical recognition.

4. APPROACH & IMPLEMENTATION

High Level workflow



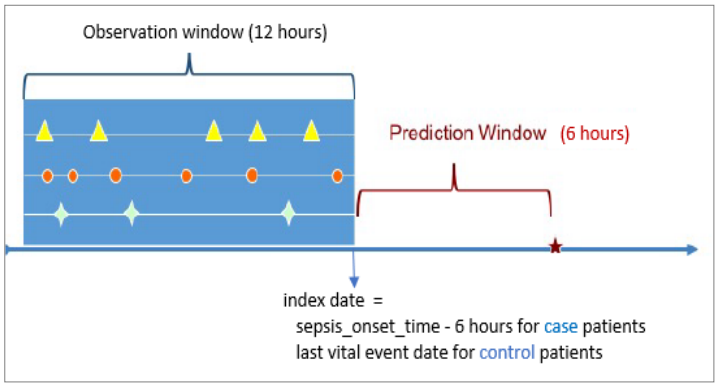
1. Identify case and control patients : Sepsis onset time was determined using SOFA>= 2 criteria as described in [1]. Using this information, ICU patients (> 15 years) were split into case and control group.

2. Define observation and prediction window : Observation window was set to 12 hours and observation window was set to 6 hours. These time periods were chosen based on distribution of ICU stay length, which is typically 24 hours (Fig-1). '*Sepsis Management Bundle'* were revised in 2012 and changed to '6 and 3-hour-bundle'. This led us to believe 6 hours prediction window a good starting point.

3. Feature engineering : Our research showed that vitals and age have high predictive power for this task and hence were chosen as base-feature set. To further enhance model's performance, we explored other data available in MIMIC database and finalized on feature-engineering based on a) GCS b) Lab results c) Comorbidity d) Note-Events
Though Comorbidity and Note-Events were promising features, we couldn't utilize it for reasons explained in section 6. Hence, we focused on combining features extracted from remaining set with standard physiological measurements to improve prediction outcome. These measurements were then temporally binned (with bin-width of 1 hour) and averaged within a bin.

4. Data cleanup : Missing data was imputed using a 'Forward/Back Fill' method where most recent bin value was propagated to subsequent / previous empty bin. Outliers were handled using winsorization technique, where most extreme values were set to 98th percentile value of feature in question.

5. Feature reduction and model creation
    1. Following set of supervised models were developed to compare and contrast predictive power of each ML-Algorithm for this dataset.
        - SVM , Logistic Regression , XGBoost, MLP
    2. Since MIMIC data is timestamped, we leveraged this information to train a RNN network as well.



Sepsis onset time criteria

Sepsis detection criteria is based on Desautels's paper (SOFA>=2) and was implemented in Big Query.

Taking the initial time of the earliest culture draw or antibiotic administration as the time of suspicion of infection, a window of up to 48 hours before this time (limited by ICU in−time) and 24 hours after this time

(limited by time of departure from the ICU) is defined. The SOFA score at the beginning of this window is compared with its hourly value throughout this window.If this hourly value is ≥ 2 points higher than the value at the start of the window, we take first such hour as the onset of sepsis.

Filtering Rules

Case patients

1. sepsis_onset_time > icu_in_time and
2. number of hourly records in observation window >= 3 and age > 15

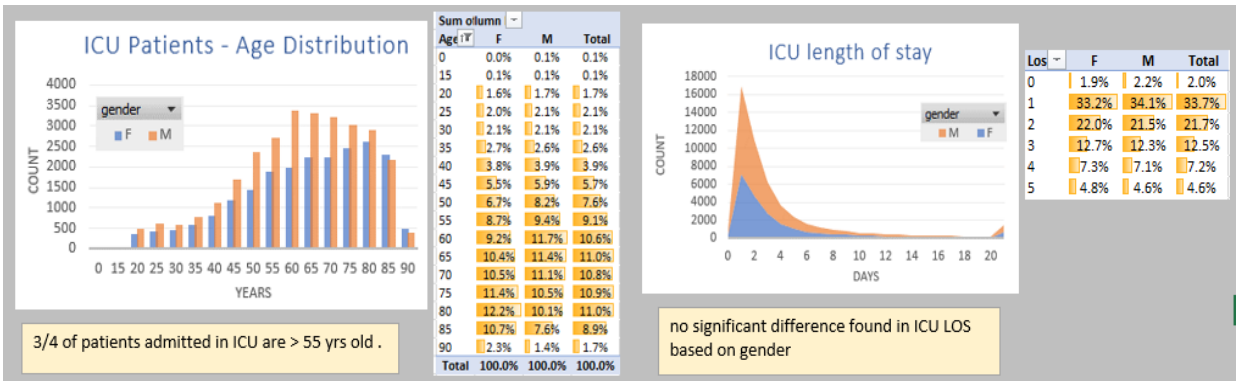
Control patients

1. Select patients that don't appear in any sepsis criteria (i.e. in sepsis_onset_time.csv or sepsis_superset_patientIDs.csv) and
2. Patient's icu_out_time - icu_in_time > 18 hours and
3. number of hourly records in observation window >= 3 and age > 15

This resulted in 546 case patients. To keep dataset balanced, similar number of control patients were randomly selected after applying above rules.

Exploratory Data Analysis

The distribution of ICU patients as well as their gender and length of stay are shown in Figure-1 below. It is important to note the ICU stays are relatively short to make predictions using longitudinal data. This will have a substantial impact on the width of the observation window we can use to make predictions.



Age	F	M	Total
0	0.0%	0.1%	0.1%
15	0.1%	0.1%	0.1%
20	1.6%	1.7%	1.7%
25	2.0%	2.1%	2.1%
30	2.1%	2.1%	2.1%
35	2.7%	2.6%	2.6%
40	3.8%	3.9%	3.9%
45	5.5%	5.9%	5.7%
50	6.7%	8.2%	7.6%
55	8.7%	9.4%	9.1%
60	9.2%	11.7%	10.6%
65	10.4%	11.4%	11.0%
70	10.5%	11.1%	10.8%
75	11.4%	10.5%	10.9%
80	12.2%	10.1%	11.0%
85	10.7%	7.6%	8.9%
90	2.3%	1.4%	1.7%
Total	100.0%	100.0%	100.0%

Los	F	M	Total
0	1.9%	2.2%	2.0%
1	33.2%	34.1%	33.7%
2	22.0%	21.5%	21.7%
3	12.7%	12.3%	12.5%
4	7.3%	7.1%	7.2%
5	4.8%	4.6%	4.6%

Figure 1. *All ICU patients*

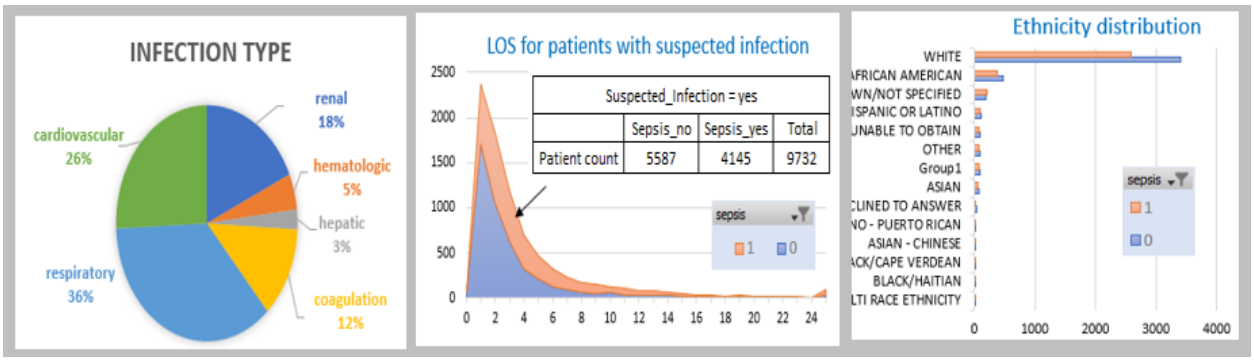


Figure 2. *ICU patients with suspected infection*

5. EXPERIMENTAL EVALUATION

Data was split into train (60%) , validation (10%) , test (20%) set. Models were trained using K-fold CV. Model performance was evaluated based on ROC-AUC. Tuning of hyperparameters was done using Grid search.
Features composed of Mean and Ptp (difference between Max and Min value) of vitals, GCS and lab-results in observation window for each icu_stay. This gave 24 new features to experiment with , in addition to 8 base features.

It is worth pointing out that while during draft-phase we focused primarily on tuning models for a balanced dataset, in this iteration we tested robustness of our model against imbalanced dataset. This is crucial because in real world , there will be many more control than case patients. Hence two sets of experiments were conducted using

a) Not-so-balanced dataset : Case to Control ratio was kept at 1:3 (via random sampling)
b) Highly imbalanced dataset : this dataset included all Control patients. (case to control ratio = 1:60)

5a. Not-so-balanced dataset (1:3)

Since in medical world, not all data is available for a given patient, it is important to evaluate to what degree model performance suffers in absence of a given feature. To evaluate that, base features were enhanced sequentially, first with Neurological measures (GCS) and then with Lab results. Table below summarizes performance of SVM, Logistic Regression, MLP and Gradient Boost classifier for different feature-sets. This dataset had 1:3 case to control ratio.

a) Vitals only

	GB	SVM	LR	MLP
Accuracy	0.8218	0.7822	0.7896	0.7871
F1 score	0.4545	0.2281	0.3411	0.3175
Recall	0.3333	0.1444	0.2444	0.2222
Precision	0.7143	0.5417	0.5641	0.5556
AUC	0.6476	0.5547	0.5952	0.5856

b) Vitals + GCS

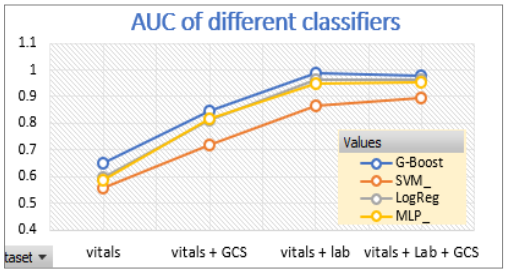
	GB	SVM	LR	MLP
Accuracy	0.8985	0.8564	0.8886	0.8837
F1 score	0.7684	0.5915	0.7305	0.7283
Recall	0.7556	0.4667	0.6778	0.7000
Precision	0.7816	0.8077	0.7922	0.7590
AUC	0.8475	0.7174	0.8134	0.8182

c) Vitals + Lab

	GB	SVM	LR	MLP
Accuracy	0.9926	0.9356	0.9777	0.9653
F1 score	0.9832	0.8354	0.9497	0.9222
Recall	0.9778	0.7333	0.9444	0.9222
Precision	0.9888	0.9706	0.9551	0.9222
AUC	0.9873	0.8635	0.9659	0.9500

d) Vitals + GCS + Lab

	GB	SVM	LR	MLP
Accuracy	0.9876	0.9455	0.9777	0.9678
F1 score	0.9718	0.8675	0.9497	0.9274
Recall	0.9556	0.8000	0.9444	0.9222
Precision	0.9885	0.9474	0.9551	0.9326
AUC	0.9762	0.8936	0.9659	0.9516



Results Discussion - We tested and validated a machine learning based approach for predicting onset of sepsis, using data that is commonly available for ICU patients. Using retrospective MIMIC-III data and Sepsis criteria 'SOFA >= 2' , several ML models were trained and tested. Feature importance was evaluated using Random Forest Classifier (Fig-4). Neurological measure (GCS-verbal and GCS-eyes) are amongst top-5 most important feature , followed by Albumin, Lactate , Sodium from Lab results. Top-5 features explained 55% variance and Top-10 features explained 75% variance in the model. PCA was used to visualize data along the first three Principal Components. Though we don't see clear separation for all datapoints , there are two distinctly visible clusters (Fig-5)

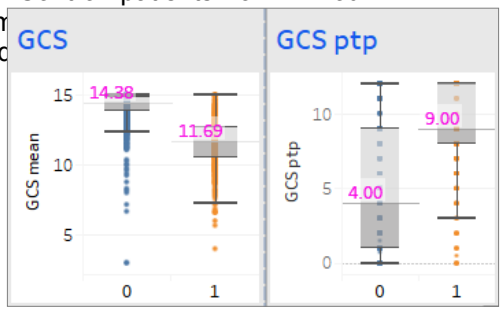
Fig-3 presents comparative view of SVC , Logistic Regression, MLP and Gradient Boost models.

- Gradient Boost classifier outperformed all other models and Support Vector Classifier shows the least favorable performance. Ensemble methods reduce variance without increasing bias. Gradient Boost classifier employs boosting , an iterative technique which adjust the weight of an observation based

on last classification. GB classifier worked significantly better on this dataset compared to other models.

- Using only vitals as features, gives accuracy of 0.82 but AUC is low (0.65). This difference between accuracy and AUC can be explained by class label imbalance (1:3).
- Model performance improves considerably when Neurological measure (GCS) is added to feature-set, improving AUC from 0.65 to 0.84.

Figure on right shows distribution of GCS for Case and Control patients for 12-hour observation window. GCS-Ptp , which is difference between m observation window shows clear distinction between Case and



- Adding lab-results to feature-set gave the best performance, improving AUC to 0.98. From Fig-3 we can also observe that adding GCS to 'Vital+Lab' doesn't bring much to the table.
- Adding GCS and lab-results to feature-set significantly improved model predictions. However, this improvement relies on that fact that these measurements are available for patient in question. In our dataset, >70% of patients had at least 1 GCS measurement.
- Fig-6 presents normalized confusion matrix for all four models (using 'Vitals+GCS' features) and helps visualize Type-I & Type-II error model is making. SVC is making more Type-II error when compared to other models.

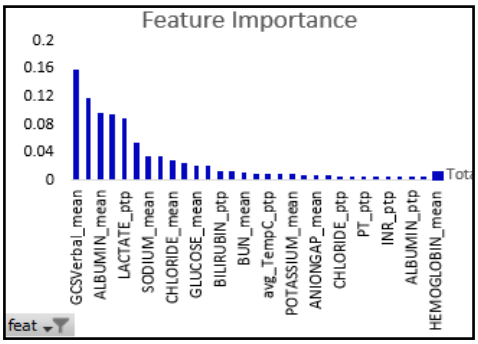


Fig-4 : *Feature importance (using RF classifier)*

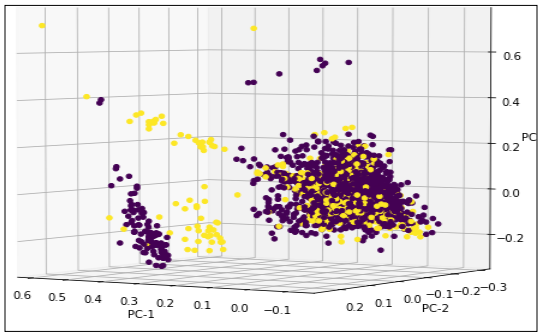
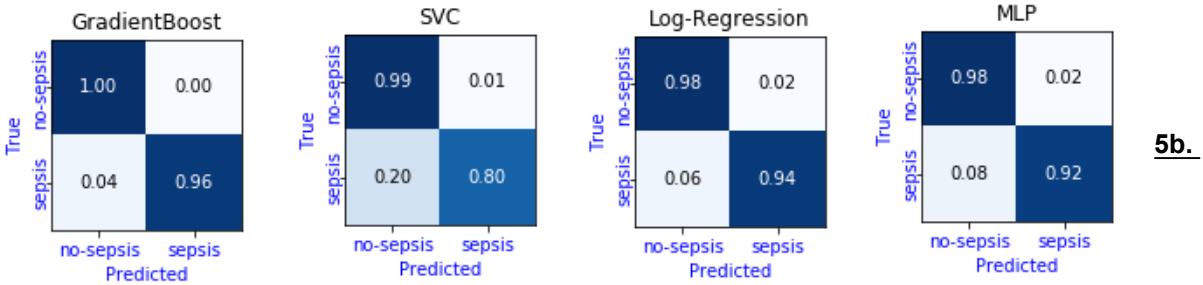


Fig-5 : *Training set visualized along first 3 PCs*



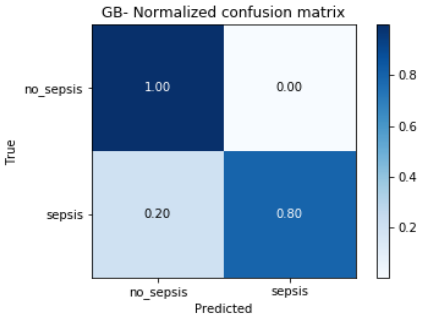
5b.

Highly imbalanced dataset (1:60)

Goal of this experiment was to determine how well model performs in face of highly imbalanced dataset. This dataset had 452 case patients and 27027 control patients (1:60 ratio). Comparing Fig-3d to Fig-4, performance of all models degraded a bit. e.g. AUC for Gradient-Boost model changed from 0.97 to 0.89. However, it is significantly better when using vitals alone as feature-set. Confusion matrix shows model makes no Type-I mistake but commits Type-II (20%) mistake, leading to low Recall.

Vitals + GCS + Lab *(entire cohort)*

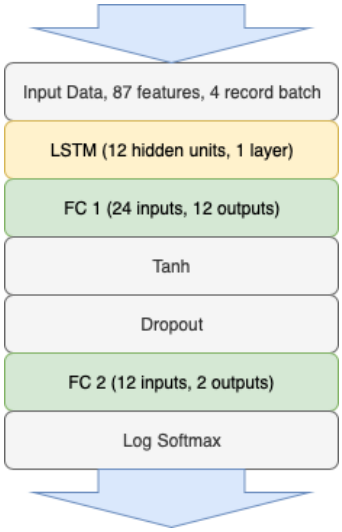
	GB	SVM	LR	MLP
Accuracy	0.9956	0.9944	0.9962	0.9955
F1 score	0.8571	0.8121	0.8743	0.8485
Recall	0.8000	0.7444	0.8111	0.7778
Precision	0.9231	0.8933	0.9481	0.9333
AUC	0.8994	0.8715	0.9052	0.8884



LSTM Results

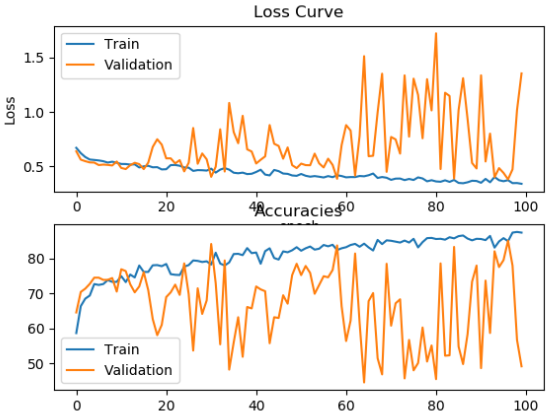
We also developed a Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) model to learn trends from the longitudinal data. In this iteration, we:

- expanded the input dimension to 87 features by enriching the dataset.
- added bidirectional weights to the network
- added an additional fully connected layer
- added dropout layer to help mitigate variance bias



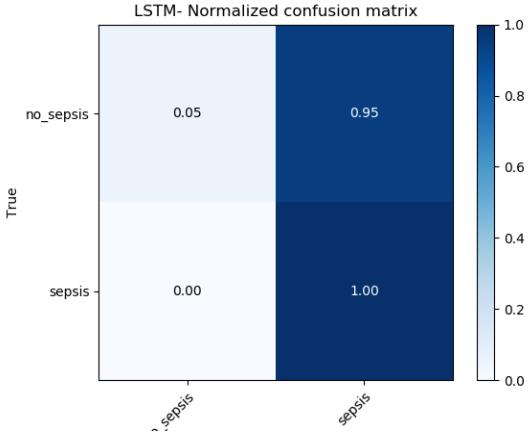
Next the model was trained using a mini batch size of 4, over 100 epochs. There were experiments performed to try and increase both the hidden size and the number of layers of the network to boost accuracy but unfortunately those modifications only boosted training time. Data were not collected on these experiments.

Looking at the loss curve below, the high variance of the validation curve indicated that the model was over training. Therefore, regularization was added to the architecture in the form of a dropout layer which randomly



The training data consisted of 34,173 patient records with 744 case patients and 1,339 control patients. We downsampled the control patients to 744 to provide a balanced dataset to the learning algorithm. The sampling was performed randomly.

The table below shows a summary of the scoring differences between the new model compared to the baseline model. In both cases, the models had very similar AUC scores with very different scores in the other categories.



LSTM	Baseline	Enriched	Percent Diff
Accuracy	0.8333	0.4480	-60.15%
F1 score	0.1951	0.6012	101.98%
Recall	0.1176	1.0000	157.89%
Precision	0.5714	0.4298	-28.30%
AUC	0.5497	0.5274	-4.14%

There is however a remarkable difference when comparing the models. The baseline model, had very low recall and there were a large number of false negatives in the test prediction. In the new model, the classifier scored perfectly on recall, meaning it captured all the sepsis cases.

Code to reproduce above experiments can be found here . Link to preprocessed here

6.CONCLUSION

Key Contribution

This study presents a machine-learning based approach to predict sepsis-onset-time using commonly available features such as vitals, GCS and lab-results for ICU patients, without explicitly relying on ICD code.

Study clearly quantifies predictive power of each feature-set. This help with interpretability of model as clinicians can gauge which feature is adding most value and hence can reason about it using their domain knowledge.

This study evaluated robustness of model in presence of imbalanced dataset, which is common is real world setting.

Challenges and Learnings

- *Sepsis detection criteria* - Detecting if the patient have sepsis or not is a challenging task. Ever-changing clinical guidelines for sepsis detection makes it even harder. Guided by our research, settled on using SOFA score >=2 as detection criteria.

- *ETL challenges* - Setting up a pyspark environment was very challenging. We have to go through the setup many environment setting before we could run code in pyspark. Cleaning date and getting the date difference in pyspark was challenging due to its syntax. replacing or removing the columns after the join was challenging.

- *Imbalanced Data* – As we performed exploratory data analysis of MIMIC-III dataset, it was clear that class labels are highly imbalanced. Hence, we must evaluate our approach to understand and evaluate impact of this on our model's performance. We started off with balanced dataset using equal number of case and control patients but randomly sampling control patients may not produce deterministic results, that generalize to entire dataset. To overcome this problem, we then optimized our models for not-so-balanced dataset (1:3) and highly-imbalanced dataset (1:60).

- *Control patient selection* - Given the highly imbalanced dataset, it was crucial for us to select control patients in an impactful way. Few different techniques such as selection based on `*length-of-ICU stays*` , `*number of records in observation window*` were tried but the one that proved to be most effective was eliminating control patients that are flagged by any other sepsis criteria such as Angus, Martin , CDC or explicit ICD code.

- *Text Feature extraction* – In addition to using non-textual features in electronic healthcare records (EHR), we had initially proposed to apply Natural Language Processing (NLP) techniques to both fill the gap for missed diagnosis and add new features to the dataset. It has been reported that sepsis is frequently underdiagnosed in the hospital despite being mentioned in the discharge notes. However, we ran into issues while extracting text features – part of the problem was that note-events in MIMIC-III data doesn't have 'icustay_id', which was our primary-key for other non-text features. Secondly, notes showed up for only 25 patients from our cohort, including only 1 case patient when a 12-hour (observation window) filter was applied. For these reasons and lack of time , we were not able to include text-features in our model but intend to pursue it in our future work. Our project includes code to extract features from chart events for others to use as well in their research.

- *Comorbidity* : Comorbidity is another promising that we couldn't utilize because this info didn't have temporal information (i.e. time at which comorbidity was detected for a given patient) and hence couldn't merge with other temporal features such as Vital , GCS etc. .

- *RNN* : One of the biggest surprises was how well the aggregated classification models performed compared to a deep learning model. It was highly anticipated that the LSTM network would be able to produce for better scoring metrics than the other models. Having limited experience with deep learning frameworks, this was a very informative exercise in learning how to

do the tensor math and design the model architecture. In future, I would invest heavily on finding a pretrained model suited to the task rather than building one from scratch.

At this stage, we have developed several classification models using features that are generally available for ICU patients. Given that there is always some nondeterminism in physiological processes, there is an open question of what the max theoretical limit is for this problem.

TEAM CONTRIBUTION

Project schedule was developed for draft and final phase and can be accessed here. Schedule has details on task distribution , deliverables timeline and work-item owners. Snapshot of task distribution is shown below.

	Work Item	Owner(s)
1	Implement sepsis detection criteria	Arti
2	create base features	Satish
3	Data pre-processing	Dhruv
5	Enrich dataset - NLP	
6	Model creation (XgBoost , MLP)	Arti
7	Model creation (SVM , LR)	Arti
8	LSTM model	Thomas
9	Create Project draft	Team

Task Distribution – Draft Phase

		Work Item	Owner(s)
1		Review litratrure and brainStorm new features	Team
	a	produce balanced dataset with Random sampling	Satish
		Combine Vitals with GCS and Lab.	
2		Extract new features (textual)	Thomas
		LDA – create features based on topic modelling and enhance dataset from step-1.	
3		Hyper-tuning / Feature Engineering	
	a	*SVM , MLP*	Arti
	b	*Gradient Bosoting , Logistic Regression*	Arti
	c	*LSTM*	Thomas
4		Project - Final Report	Arti / Thomas
5		PPT + video presentation	Dhruv
		1. PPT slides to summarize your paper *2.Put the Youtube link of 5-min presentation*	
6		SW packaging and documentation	Team
		zip Project report, code, Readme, data , best models and PPT using the format of team #-topic and submit on Canvas	

Task Distribution – Final Phase

References

1. T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh,D. Shimabukuro, U. Chettipally, M. D. Feldman, C. Barton, D. J. Wales, and R. Das. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. JMIR Med Inform, 4(3):e28, 30 Sept. 2016.

2. M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, pages 75–84, New York, NY, USA, 2014. ACM.

3. K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria. A targeted real-time early warning score (trewscore) for septic shock. Science translational medicine, 7(299):299ra122–299ra122, 2015.

4. Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2565–2573. ACM, 2018.

5. S. Nemati, A. Holder, F. Razmi , M.D. Stanley, G.D Clifford GD, T.G Buchman. An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. Critical Care Medicine. 46(4):1. December 2017

6. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the third International Consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016;315(8):762-774.

7. Mortality Prediction Model of Septic Shock Patients Based on Routinely Recorded Data-Computational and Mathematical Methods in Medicine Volume 2015, Article ID 761435

8. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. (February 2016). "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". JAMA. 315 (8): 801–10. doi:10.1001/jama.2016.0287. PMC 4968574. PMID 26903338.

9. 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.