Prediction of Hypoxemia Through Deep Learning

Anuhya Vajapey
anuhyav@mit.edu

Marie-Laure Charpignon
mcharpig@mit.edu

Abstract

Hypoxemia is a severe condition that occurs in the Intensive Care Unit (ICU), that if undetected early can cause major complications in a patient's procedure. In this work, we use a deep learning approach with Long Short Term Memory Models to predict the onset of Hypoxemia 6 hours prior to the window it occurs in and evaluate the performance against a baseline logistic regression model.

1. Introduction

Hypoxemia is one of the most common critical illnesses in the ICU. It is a medical condition that occurs when a person's arterial blood oxygen partial pressure (PaO2) is less than 80 mmHg (PaO2 < 80 mmHg) and can have severe complications for patients.

There are three ranks of severity for Hypoxemia:

1. Mild: $Pa[O_2] < 80$ mmHg
2. Moderate: $Pa[O_2] < 60$ mmHg OR $Pa[O_2]/Fi[O_2] < 200$
3. Severe: $Pa[O_2] < 40$ mmHg OR $Pa[O_2]/Fi[O_2] < 100$

If mild hypoxemia is not detected or treated in a timely manner, it will transform into moderate or severe hypoxemia. The treatment for Hypoxemia consists of administering oxygen therapy to patients on mechanical ventilation.

1.1 Motivation

The goal of this study is to predict the onset of moderate Hypoxemia so that healthcare providers can be alerted for early intervention. Early detection of hypoxemia permits timely oxygen therapy, infusion control, and other appropriate treatments that can reverse aggravation without the use of mechanical ventilation.

There exists a set of physiological and biochemical indices from laboratory tests and vital signs that are known to be positively correlated with the occurrence of hypoxemia. We will use these indicators to predict the occurrence of moderate hypoxemia. Among the 44 covariates manually curated by Dr. Cong Feng, we used statistical analysis in addition to machine learning method to create a model with the most relevant indicators of hypoxemia. The model outcome will be a binary variable: 1 for higher risk of developing hypoxemia and 0.

1.2 Related Works

Several works have used various approaches to detect hypoxemia. One study used a deep learning model on data gathered from a fingertip sensor to predict impending hypoxemia [2]. Lundberg et. al created an ensemble-based gradient boosting model that uses EMR data to predict the near-term risk of hypoxemia [3]. Williamson et. all review the use of various machine learning and time series algorithms including Gaussian Mixture Models to predict hypoxemia in infants [6].

Deep learning has been used for a variety of clinical tasks. One study used deep learning methods to predict several severe complications in post cardiosurgical care in real time [1]. Another study built a real-time model to output the probability of discharging a patient to help hospital reprioritize patients[5]. A nonparametric model based on Gaussian process was developed in another study to predict a variety of diseases and subgroups[7]. Our model was heavily inspired by the real-time clinical intervention prediction model developed by Suresh et. al which consists of using neural networks to predict clinical interventions in a forward-facing manner[2].

2. Methods

2.1 Data Sources

We use data from the Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) v1.4 database[4]. MIMIC is publicly available and contains over 58,000 hospital admissions from approximately 38,600 adults. The databases have been de-identified; all PHIs have been removed and dates have been changed. We will be looking at ICU events ranging from 2001 to 2012.

2.2 Study Population

Only patients who are older than 15 years old at time of event and had ICU stays from 24 hours onwards are considered. Physiological factors of hypoxemia are known to be different for children and laboratory tests are also interpreted differently since the normal heart rate scale is shifted, among other reasons. Therefore, the study excludes any infants.

1. To predict early stages of hypoxemia, we include adult ICU patients of all diagnoses. In training phase, patients whose outcome is 0 did not develop hypoxemia nor required mechanical ventilation throughout their stay in ICU. Patients whose outcome is 1 must have had at least one PaO2 test result before the onset of hypoxemia and should be diagnosed with hypoxemia but not administered with mechanical ventilation, within the following 24 hours spent in ICU. To be specific, it means that any mechanical ventilation event preceding test or diagnosis will be excluded. Yet, if all of the above are satisfied and mechanical ventilation happens to be prescribed after the first day in the ICU, the record is not discarded.
2. We divided the selected cohort into positive and negative examples. The positive examples consisted of all patients undergoing mechanical ventilation that had $Pa[O_2]/Fi[O_2] < 200$. This resulted in 10,452 P/F ratios of 1. For the negative cohort, we selected all the patients on mechanical ventilation with P/F ratio >= 200 and all patients not on mechanical ventilation that did not die in the ICU. We ended up with a combined total of 37, 214 patients for the negative cohort. (See Appendix 1 and 4 for summary statistics about the cohorts.)

2.3 Preprocessing

First, we gathered a set of variables deemed clinically relevant by the clinician on our team, including laboratory tests and vital signs, along with binary features (vasopressors, crystalloids and comorbidities) and demographics. Across each hour, if multiple values were recorded for a patient for one variable, we found the average, min, and max for the hour and stored those three values for the variable.

In order to fill missing values, we used forward filling imputation and population mean. First we forward filled all missing values. Next, for each variable, we calculated the population mean. For all the missing values prior to the first recorded values, we filled them in with the overall population mean for that variable. The variables are listed at the end of this report.

2.4 Models

We used two approaches to predict hypoxemia. We used a logistic regression model to determine a baseline AUC for the task.

Along with establishing baseline performance metrics, the logistic regression model enabled us to understand which features were the most relevant. We adopted two approaches, which led to the development of two distinct models. The first method consisted in gathering data (laboratory tests, vital signs, vasopressors, crystalloids and demographics) from hour 6 to hour 11 and to predict the occurence of hypoxemia between hour 18 and hour 21. Yet, the difficulty is to be able to predict hypoxemia events in real time. That is why the second method consisted in aggregating all observation slices from hour 6 to hour 156 and to predict the outcome during a four-hour window 12 hours later. In both cases, we had train (size 6825 and 1002 respectively), validation (size 6826 and 1001 respectively) and test sets (size 6826 and 1001 respectively).

Our second approach was using Long Short Term Memory Models (LSTM). LSTMs are a state of the art neural network architecture that captures the temporal

dependencies in data over time. After preprocessing the data, we split the data into training and testing sets. The training test was further split into a validation set. We only used the testing set to evaluate our model's performance at the end. For each patient, we had an observation window, a gap window, and prediction window, as shown in Figure 1. We used the data in the observation window to predict what the outcome (1=hypoxemia) in the prediction window. If the outcome actually occured in the gap window, we discarded our predicted result and set the outcome to 'None' to ensure that we weren't predicting outcomes that occurred prior to the prediction window. The windows would then slide hour by hour till the end of the patient's stay in the ICU or maximum length of stay of 7 days. We initialized our window lengths to 6 hours for the observation and gap window, and 4 hours for the prediction window (see Appendix 1 and 2). Because hypoxemia is heavily class imbalanced, we manually tuned class weights in the LSTM model to ensure no overfitting. The results from a majority classifier (logistic regression) and our model are shown in the results section.

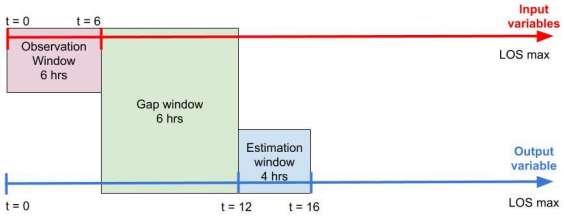


Figure 1: Window visualization and initializations.

3. Results

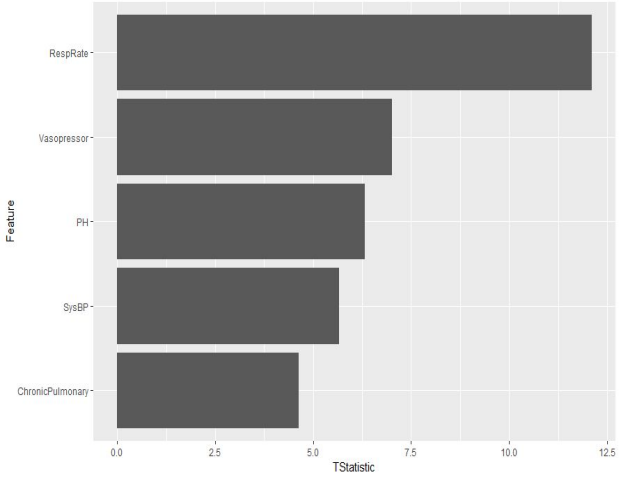
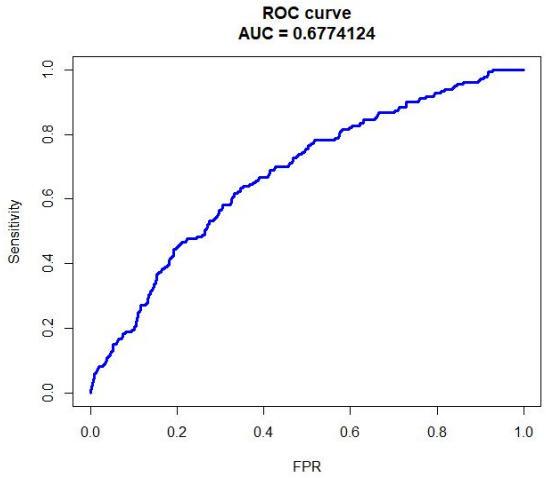
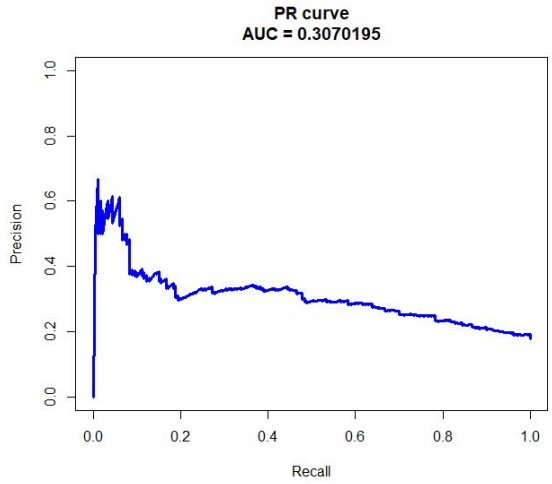


Table 1: Logistic regression model with one slice only (observation window: hour 6 - hour 12, prediction window: hour 12 - hour 18). Variable importance in terms of t-statistic absolute value.

Ranking the features by variable importance (in terms of the absolute value of their t-statistic) highlights three continuous variables and two binary features. Respiratory rate and systolic blood pressure are the two vital signs of importance. The pH value also matters a lot in the prediction of hypoxemia. Normal blood pH ranges from 7.35 to 7.45 (i.e. slightly to the alkaline side of the scale, centered at 7). If the pH is at the low end of the scale or if it is actually below 7.35, the condition is acidemia; if above, it is described as alkalemia. The prime buffer system to prevent pH imbalances is through carbonic acid and bicarbonate. But there is actually a third buffer system which establishes the link between respiratory rate and pH: when the first two systems are dysfunctioning, hyperventilation and hypoventilation can be used by the body to control pH. That systolic blood pressure appears among the top features also makes sense clinically. Hypoxemia is indeed associated with increased systolic blood pressure. Finally, previous research has shown that chronic hypoxemia is linked to elevated sympathetic activity and hypertension in patients with chronic pulmonary obstructive disease, a comorbidity that appears in the top five features. Note that the sympathetic nervous system itself can lead to an increase in heart rate and blood pressure.





Figures 2, 3: Logistic regression model with one slice only. Performance measured by AUC > 0.67. Performance in terms of Precision and Recall.

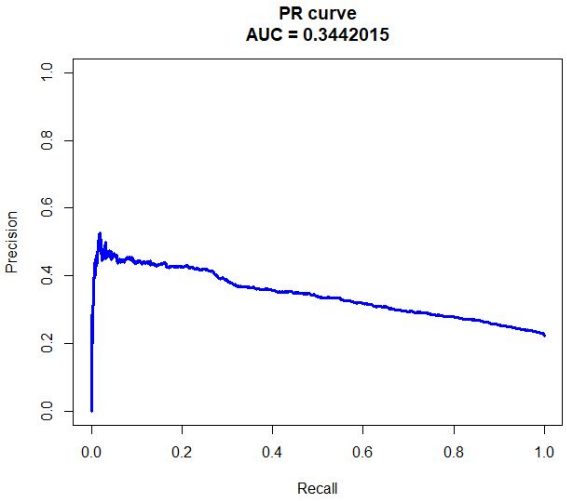


Figure 4: Logistic regression model with all slices. Performance in terms of Precision and Recall (+ 4 points in AUC compared to model with one slice only).

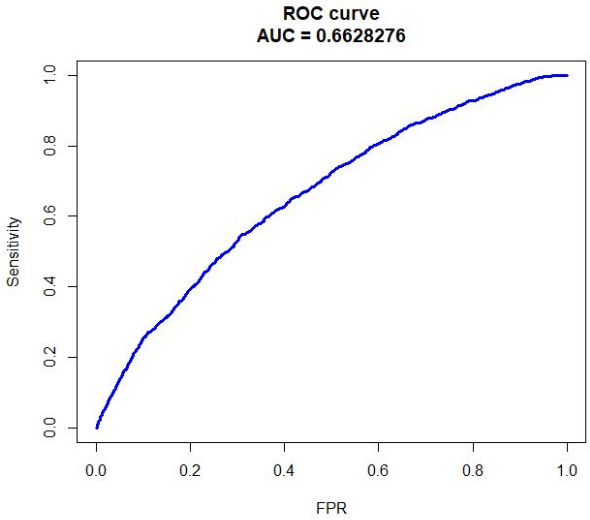


Figure 5: Logistic regression model with all slices. Performance measured by AUC > 0.66 (-1 point in AUC compared to model with one slice only).

Model	AUC*	ACCURACY
LR (t < 6 hr) only first observation window	0.71	0.71
LR (t < 168 hr) all 6-hour windows	0.66	0.58
LSTM	0.77	0.71

Table 2: Summary of performance for both logistic regression and LSTM models, in terms of AUC and accuracy.

We used several metrics for evaluating our model. Area Under the receiver operating characteristic Curve (AUC) is a popular measure of evaluating a model's in a binary classification problem. AUC reflects the ratio of false positives to true positives. The results are shown in Table 2.

4. Discussion

To the best of our knowledge, this study is the first to apply LSTMs to the prediction of hypoxemia. In this preliminary work, we established a baseline through a

logistic regression model and evaluated a real-time model through LSTMs.

4.1 Conclusion

With the baseline logistic regression model, we were able to get a high accuracy and AUC if we only consider the first window of data. However, a more clinically relevant result is building a real-time prediction model that can predict the onset of hypoxemia prior to the occurence. This task is much harder for a baseline logistic regression model to perform well on and resulted in accuracy of 0.58 and AUC of 0.66, which is significantly worse than when we only considered one window of data. LSTM model performed much better in this task with accuracy of 0.71 and AUC of 0.77.

A real time prediction model for detecting hypoxemia is useful for clinicians because hypoxemia often requires mechanical ventilation which could lead to multiple organ failure. Detecting it early would allow for oxygen therapy to be administered earlier and reduce the risk of further complications for a patient. Delivery of oxygen at the right time is absolutely crucial. Oxygen given too early is costly and wasteful. Given too late, and the patients might face complications.

4.2 Future Work

In the future, we want to explore different methods of imputation to see if AUC improves and also use other metric of analysis to analyze our model such AU-PRC which often utilized for class-imbalanced tasks. We can also try using different preprocessing methods. Instead of taking the averages of values, we plan on using a Z-scoring method as well which for each variable transforms it into a Z-score. This might help in reducing noise and make the data more learnable and differentiable between the positive and negative cases. We also plan on evaluating our model with different lengths for the observation, gap, and prediction windows to determine how early we can detect hypoxemia and how that correlates to accuracy of our model. We plan on determining the clinically relevant gap window. If only one hour is necessary to prevent hypoxemia, our model can evaluate with just gap window length equal to 1 hour, which might increase accuracy.

Various studies for time series predictions have also used other models than LSTMs. One interesting model could be a CNN with 1x1 convolutions. This would help in dimensionality reduction while allowing complex and learnable interactions of cross channel information. Exploring how Gaussian models or markov decision models perform on this task might also be useful in developing a faster model. LSTMs do take a lot of computing power and simpler models might be more useful for hospitals to run in order to actually use machine learning in practice.

Acknowledgments

We would like to thank Harini Suresh for collaborating with us for the LSTM methodology as described in her paper [2]. We also want to thank Aldo Arevalo and Dr. Lu Shen their guidance and contributions to the project. Thanks to Dr. Cong Feng for coming up with the clinical problem to solve. Thank you to all of the HST.953 course staff for answering our questions and mentoring us throughout the semester.

References

[1] “Articles: Machine Learning for Real-Time Prediction of Complications in Critical Care: A Retrospective Study”; Meyer, Alexander, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. (2018).

[2] “Anesthesiologist-level forecasting of hypoxemia with only SpO2 data using deep learning”; G. Erion, H. Chen, S.M. Lundberg, S.-I. Lee; 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[2] “Clinical Intervention Prediction and Understanding using Deep Networks”; H. Suresh, N. Hunt, A. Johnson, L.A. Celi, P. Szolovits, M. Ghassemi (2017).

[3] “Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery”; S.M. Lundberg, B. Nair, M.S. Vavilala (2017).

[4] “MIMIC-III, a freely accessible critical care database”; Johnson AEW, Pollard TJ, Shen L, Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG. Scientific Data (2016).

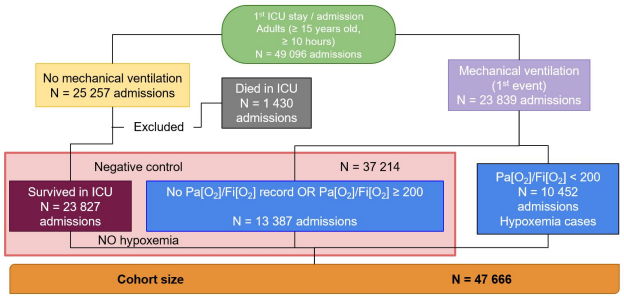
[5] “Real-Time Prediction of Inpatient Length of Stay for Discharge Prioritization”; Barnes, Sean, Eric Hamrock, Matthew Toerper, Sauleh Siddiqui, and Scott Levin.

Journal Of The American Medical Informatics Association: JAMIA 23, no. e1 (April 2016):

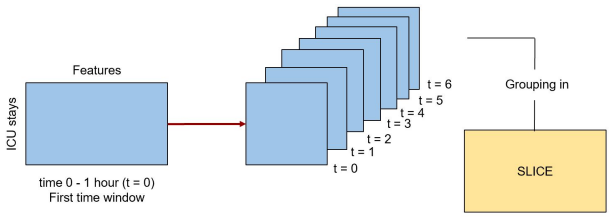
[6] "Review: Forecasting Respiratory Collapse: Theory and Practice for Averting Life-Threatening Infant Apneas"; Williamson, James R., et al. Respiratory Physiology & Neurobiology, vol. 189, Nov. 2013, pp. 223–231.

[7] "Sparse Multi-Output Gaussian Processes for Medical Time Series Prediction." Cheng, L.-F., Darnell, G., Dumitrascu, B., Chivers, C., Draugelis, M. E., Li, K., & Engelhardt, B. E. (2017).

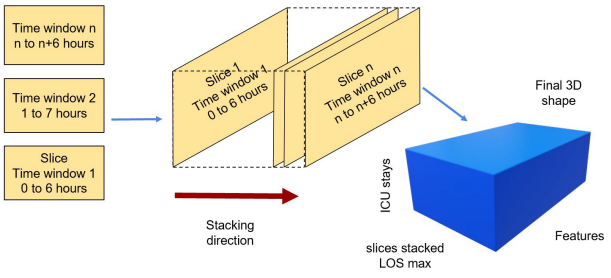
Appendix



Appendix 1: Cohort diagram.



Appendix 2: Visual representation for the creation of one slice, i.e. a six-hour observation window.



Appendix 3: Visual representation for the concatenation of all slices, across unique ICU stays, leading to 3D tensor.

	Train	Test
Patients	3974	3949
ICU Stays	4233	4216
Elective Admission	830	855
Urgent Admission	103	103
Emergency Admission	5892	5868
Clinical Referral	2401	2380
Emergency Room	1935	1935
Hospital Transfer	1389	1428
Physician Referral	1069	1057
Age <= 30	301	332
Age in]30,50]	1137	1073
Age in]50,70]	2982	2905
Age > 70	2405	2516
Black / African American	580	619
Hispanic / Latino	272	213
White	4915	4923
Female	2882	2925
Male	3943	3901
Vasopressor	459	435
Chronic Pulmonary	1627	1688
Fluid Electrolyte	3499	3531
Hypoxemia	1515	1515

Appendix 4: Summary statistics of train and test sets