Predicting Neurological Outcome at Discharge using Vital Signs and Lab measurements for Patients with Traumatic Brain Injury

Louis A. Gomez

[1]Stevens Institute of Technology, Hoboken, NJ

ABSTRACT

Objective: To test whether vital signs and lab measurements from the first two days of admission in an intensive care unit (ICU) are predictive of the neurological outcome at discharge.

Methods: We used data from patients in the MIMIC database with a diagnosis of traumatic brain injury and motor Glasgow Coma Scale (mGCS) recorded. Data included static, vital signs, and lab measurements from the first two days in the ICU. Motor GCS is used as a proxy to quantify neurological outcomes. This is mapped from a scale of 1 – 6: (i) 1, no response; (ii) 2, abnormal extension; (iii) 3, abnormal flexion; (iv) 4, flex to withdraw from pain; (v) 5, moves to localized pain; (vi) and 6, obeys command. We performed a binary classification task between the non-command following (1,2,3,4,5) vs command following (6) because this provides information about which patients could attain command following at discharge. We evaluate using recall, precision, and f1-score on three models: logistic regression, random forest, and XGBoost

Results: The best f1 score of 0.66 was achieved on both logistic regression and XGBoost. We attained the best average recall of 0.69 on logistic regression and the best average precision of 0.66 on XGBoost. Additionally, we found 7 features that are deemed important across all models used.

Conclusion: Our results show that predicting the neurological outcome is a challenging task. This work has shown that using only static, vital signs, and lab measurements from the first two days in the ICU, we achieve an f1 score of 0.66. Future work could involve using more information such as drugs administered and explore other ways to deal with missing values to retain more variables.

INTRODUCTION

Traumatic brain injury (TBI) is a severe condition caused by the disruption in normal brain functioning due to sudden trauma such as a bump, blow or jolt to the head [1]. In 2014, 2.87 million people in the United States were hospitalized or died due to TBI-related incidents [1]. Due to the nature of the injury, doctors are interested in the neurological outcomes after TBI. The Glasgow Coma Scale (GCS) [2] is a tool used in the Intensive Care Unit to measure the severity of TBI and neurological outcomes after TBI [2], [3]. The GCS is a behavioral assessment comprised of three different sub-scales: motor, verbal, and eye response. Prior work has shown that the total GCS score can be misleading [4]. Also, work on neurological outcomes has observed that the motor component of the score is often more predictive than any of the individuals' sub-scales or the total sum score [3]. In this work, we test whether patient data from the first two days in the ICU can be used to predict neurological at ICU discharge. The development of an accurate tool can provide medical doctors with more information and influence treatment plans that may involve devoting more clinical resources or intervening early on to improve patient outcomes.

METHODS

Dataset

In this study, we consider all patients admitted to the Intensive Care Unit (ICU) with an ICD9 diagnosis of traumatic brain injury. The definition and scope of traumatic brain injury were defined using the recommended ICD9 guidelines[1]. We excluded all patients without the motor GCS as this was our proxy for neurological outcome. The GCS is not recommended for use in children due to a lack of proper calibration [5], so we exclude all patients who are not older than 18. Additionally, we exclude all patients in the ICU for less than 48 hours as this is our time frame of interest. If a mGCS reading is absent on ICU discharge, we subsequently excluded that ICU admission. Additionally, if there are multiple mGCS assessments on the day of discharge, we select the last recorded score as the label. While patients may have multiple admissions to the ICU, we considered only the first admission to the ICU in this study. This brings the number of admissions we consider (with mGCS scores) to 452 admissions (am unique patients). The Motor Response is a scale from 1 – 6 described as: (i) 1, no response; (ii) 2, extension response in response to pain (abnormal extension); (iii) 3, flexion in response to pain (abnormal flexion); (iv) 4, withdraws in response to pain (normal flexion); (v) 5, purposeful movement to localized pain; (vi) 6, obeys command for movement. For the 452 ICU stays we consider, there is a large data imbalance in the number of labels per score. The largest difference is between the score of 6 with 75% and the score of 2 with 0.01% of the total number of labels. **Table 1** shows the grouping and the number of labels in each group.

Table 1: Description of each motor Glasgow coma scale score and number of labels present

Scores	Categories	Labels per category
1	No response	23
2	Extension response in pain	6
3	Flexion in response to pain	8
4	Withdraws in response to pain	25
5	Purposeful movement to localized pain	51
6	Obeys commands	339

Data Processing

Each patient has data recorded across different data modalities, but we only considered static, vital signs, and lab readings. To process our data, we follow the standardized approach introduced in the MIMIC Extract [6] to create reproducible data extraction and data representation pipeline. This pipeline includes clinical groupings, unit conversions, data filtering, and data aggregation. To begin, we combine itemids

[1] https://health.mil/Reference-Center/Publications/2015/12/01/Traumatic-Brain-Injury

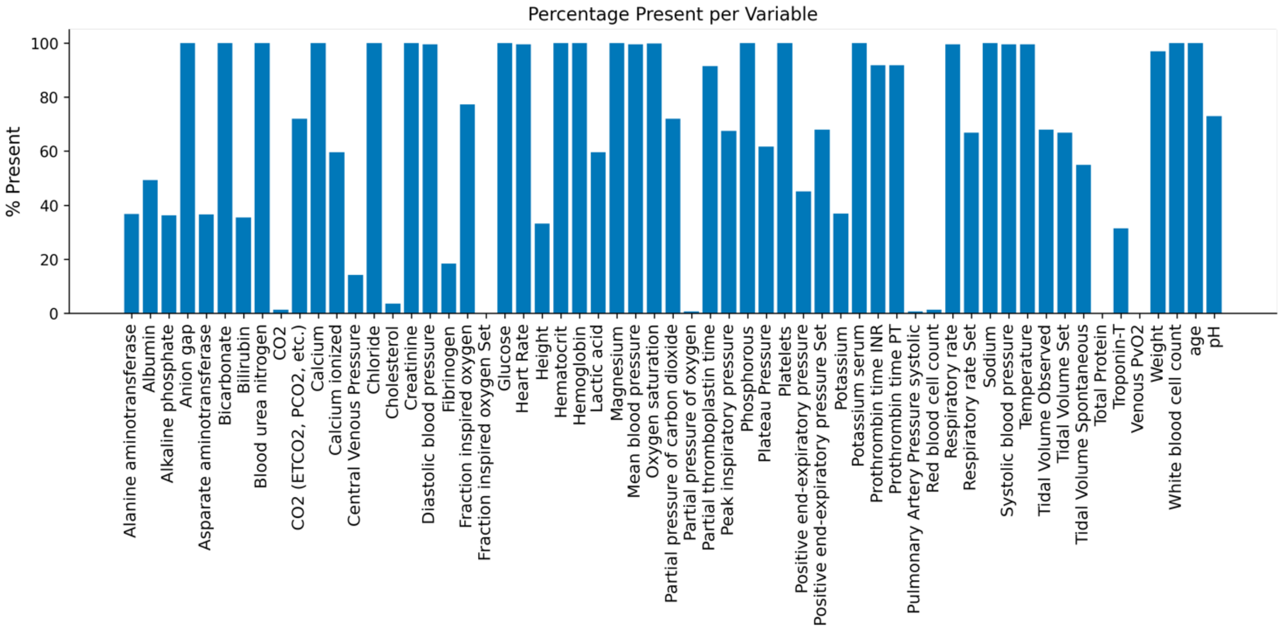

Figure 1: Bar chart showing the percentage of data measured in each clinical grouping

representing the same information at a high level into the same clinical grouping. This step is vital for two reasons. First, itemids for the same variable could be different in carevue and metavision. For example, heart rate has an itemid of 211 in carevue but an itemid of 220045 in metavision. Secondly, variables could have different itemids even in the same database source. For example, height has itemid 226707 and 226730 in metavision. Furthermore, this step helps to reduce our feature space to 59 variables. Detailed information on how these groupings were generated can be found in the data files from the MIMIC Extract GitHub repo[2] and paper. Next, we converted units for variables that are commonly measured in different units which included converting height to centimeters, weight to kilograms, and body temperature to Celsius. After unit conversions, we filtered all the variables to their clinically recommended ranges. We used data and code provided in supplemental materials of the MIMIC Extract to perform this operation[3]. This process involved the imputation of variables whose values were higher (or lower) than the "normal" range using the previous values. Values beyond the normal range were considered as errors and replaced with nans. To aggregate our data, we computed the mean for each variable over the two-day ICU period. After this step, each patient has a unique ICU stay and a single feature vector. Note that we do not include the total, verbal, or eyes scores in the feature set as these could be correlated with our label. **Figure 1** shows the amount of missing data present for all final clinical groups. At this stage, about 34% of the data is still missing as some clinically aggregated variables are not commonly recorded across patients. We fixed a threshold of 50% and dropped all features that did not meet the criteria. This step reduced the number of variables from 59 to 40 variables. See Appendix for the final 40 variables used in this study**.** Finally, we imputed our dataset using the mean imputation technique that replaces each nan value with the mean of that variable. This

[2] Clinical Groupings from MIMIC Extract
https://github.com/MLforHealth/MIMIC_Extract/blob/master/resources/itemid_to_variable_map.csv

[3] Variable Ranges from MIMIC Extract
https://github.com/MLforHealth/MIMIC_Extract/blob/master/resources/variable_ranges.csv

imputation strategy is in line with prior works using the MIMIC dataset [6]–[8]. We then normalized each variable to have a mean of 0 and a standard deviation of 1. The imputation and normalization were performed on the training set and then applied to the test set.

Experiments

We focus on the binary classification task between the non-command following (1,2,3,4,5) and command following (6). This is the largest classification group and provides information about which patient will attain command following at ICU discharge. We examined predictive performance between logistic regression, random forest, and extreme gradient boosted trees (XGBoost). We split our dataset into train and testing data with a 70/30 (316/136) stratified split to keep the percentage of per class samples equals in both train and test data. For each classification model, we used default model parameters for training. To mitigate the effects of our imbalanced dataset, we added higher class weights to the minority class. In XGBoost the weight is the ratio of the number of samples in the majority class to the minority class. In logistic regression and random forest, we used the 'balanced' option that readjusts the weight inversely proportional to the class frequencies. To evaluate our models, we computed the f1-score, precision, and recall and subsequently extracted the 20 most predictive features in each classification model. We examined if any features were deemed important across all models

RESULTS

Overall, we achieved a predictive performance of 0.55 using random forest and 0.66 on both logistic regression and XGBoost on the macro f1-score. The macro score is used here because our test set remains imbalanced. Results for all models on additionally average recall and precision are reported in **Table 2**. Both logistic regression and XGBoost have a marginal improvement over each other for recall and precision. Logistic regression has a better recall by 0.02 while XGBoost has a better precision by 0.01. The highest precision score of 0.70 and lowest recall score of 0.56 was on random forest. To better understand the recall and precision for individual classes, we plot each class for each model. This is vital because our test set is imbalanced and good average recall/precision performance does not convey how good the models perform in each class. **Figure 2** shows the recall and precision on a per-class level.

Table 2: Classification results across all three models

Classification Models	**Evaluation Metrics**		
	Avg. Recall	Avg. Precision	Macro F1-Score
Logistic Regression	0.69	0.65	0.66
Random Forest	0.56	0.70	0.55
XGBoost	0.67	0.66	0.66

In **Figure 2**, we observe the difference in recall and precision for each class. For the non-command following task, the highest recall is 0.65 using logistic regression while the lowest is 0.15 using random forest. Conversely, the highest precision is 0.62 using random forest while logistic regression and XGBoost achieve 0.44 and 0.47 respectively. For command following, the random forest has an almost perfect recall of 0.97, with 0.80 on XGBoost and 0.73 on logistic regression. For recall, the highest score is using logistic regression with a recall of 0.86 compared to 0.84 using XGBoost. The lower recall and precision in the minority class show that our model struggles with being able to classify the samples even after re-weighting is applied in training.

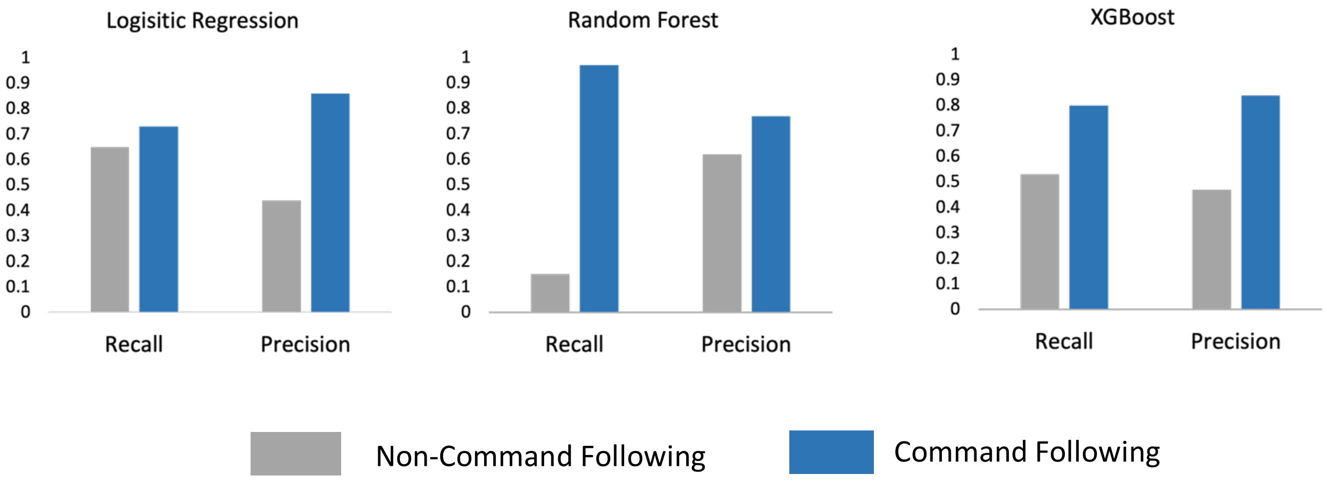

Figure 2: Bar charts for each model showing per class recall and precision.

Additionally, we aggregated the features that are deemed important for all models. We selected the top 20 features by magnitude since coefficients in logistic regression can have negative weights indicating an inverse relationship. Figure 3 shows the top 20 features across both models. Due to the differences in the way feature importance is calculated, (Gini coefficient in random forest, feature coefficients in logistic regression, and gain in XGBoost), each model may select a different set of features. Of the top 20 features selected in both models, there were 7 features in common for all models. The common features are pH, white blood cell count, magnesium, partial pressure of carbon dioxide, oxygen saturation, plateau pressure, and hemoglobin. See appendix for the bar charts with the top 20 features and their scores for each model.

DISCUSSION

In this project, we examined the predictive performance of using static information, vital signs, and lab measurements in the first two days of ICU admission to predict the neurological outcome at ICU discharge. Using three different models, we examined the recall, precision, and f1-score for classifying between non-command following (1,2,3,4,5) vs command following (6). We achieved the best f1 score of 0.66 performance on both logistic regression and XGBoost. Results for precision and recall on a per-class basis show that our model struggles to identify samples from the minority classes. Using random forest, the lowest recall of 0.15 is observed for the non-command following class. Additionally, we also found seven features in the top 20 most important features that were common to all models we experimented with. For processing our data, we followed the standard pipeline introduced in MIMIC Extract to make our dataset curation reproducible. The clinical grouping stage helped aggregated our data into high-level clinical groupings from itemids that convey the same information across metavision and carevue. This helped reduce our feature space and reduce the degree of missingness as well. While our results show that this classification task may be possible, we used only a small dataset of 452 admissions in this study which limits our generalizability. Larger datasets that include other kinds of patient diagnosis could be considered to create more general models to predict outcomes. Additionally, we did not explore the potential for multiclass classification between the mGCS scores as we had a limited number of labels in this cohort. In this work, we used the simple mean imputation method to impute our dataset. We did not consider other more sophisticated imputation techniques that could consider relationships and correlation across variables. Future work could examine the effects of imputation on predictive performance. An extension of this work could include other kinds of data available such as drug dosages and prescriptions, clinical notes, and waveform data.

CONCLUSION

Accurate classification early on during an ICU stay can lead to increased resources for patients with a higher probability of reaching full command following. In this work, we examined how predictive static, vital signs, and lab measurements from the first two days in the ICU. Future work could involve using more information such as drugs administered and explore other ways to deal with missing values to retain more variables.

REFERENCES

[1] A. Peterson, L. Xu, J. Daugherty, and M. Breiding, “Surveillance Report of Traumatic Brain Injury-related Emergency Department Visits, Hospitalizations, and Deaths - United States 2014,” Center for Disease Control and Prevention, 2019.

[2] G. Teasdale and B. Jennett, “Assessment of coma and impaired consciousness,” *The Lancet*, vol. 304, no. 7872, pp. 81–84, Jul. 1974, doi: 10.1016/S0140-6736(74)91639-0.

[3] M. Balestreri *et al*., “Predictive value of Glasgow coma scale after brain trauma: change in trend over the past ten years,” p. 2.

[4] S. M. Green, “Cheerio, Laddie! Bidding Farewell to the Glasgow Coma Scale,” *Ann. Emerg. Med.*, vol. 58, no. 5, pp. 427–430, Nov. 2011, doi: 10.1016/j.annemergmed.2011.06.009.

[5] F. Ghaffarpasand, A. Razmkon, and M. Dehghankhalili, “Glasgow Coma Scale Score in Pediatric Patients with Traumatic Brain Injury; Limitations and Reliability,” p. 2.

[6] S. Wang, M. B. A. McDermott, G. Chauhan, M. C. Hughes, T. Naumann, and M. Ghassemi, “MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III,” *Proc. ACM Conf. Health Inference Learn.*, pp. 222–235, Apr. 2020, doi: 10.1145/3368555.3384469.

[7] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, “Recurrent Neural Networks for Multivariate Time Series with Missing Values,” *Sci. Rep.*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-24271-9.

[8] B. Nestor *et al.*, “Feature Robustness in Non-stationary Health Records:,” *Proc. Mach. Learn. Res.*, pp. 1–23, 2019.

APPENDIX

Variables used in the Study

Anion gap, Bicarbonate, Blood urea nitrogen, CO2, Calcium, Calcium ionized, Chloride, Creatine, Diastolic blood pressure, Fraction Inspired oxygen, Glucose, Heart Rate, hematocrit, Hemoglobin, lactic acid, magnesium, Mean blood pressure, Oxygen saturation, partial pressure of carbon dioxide, partial thromboplastin time, peak inspiratory pressure, Phosphorus, Plateau pressure, Platelets, Positive end-expiratory pressure set, Potassium serum, Prothrombin time INR, Prothrombin time PT, Respiratory rate, Respiratory rate Set, Sodium, Systolic blood pressure, Temperature, Tidal volume observed, Tidal volume set, Tidal volume spontaneous, Weight, White blood cell count, age, pH.

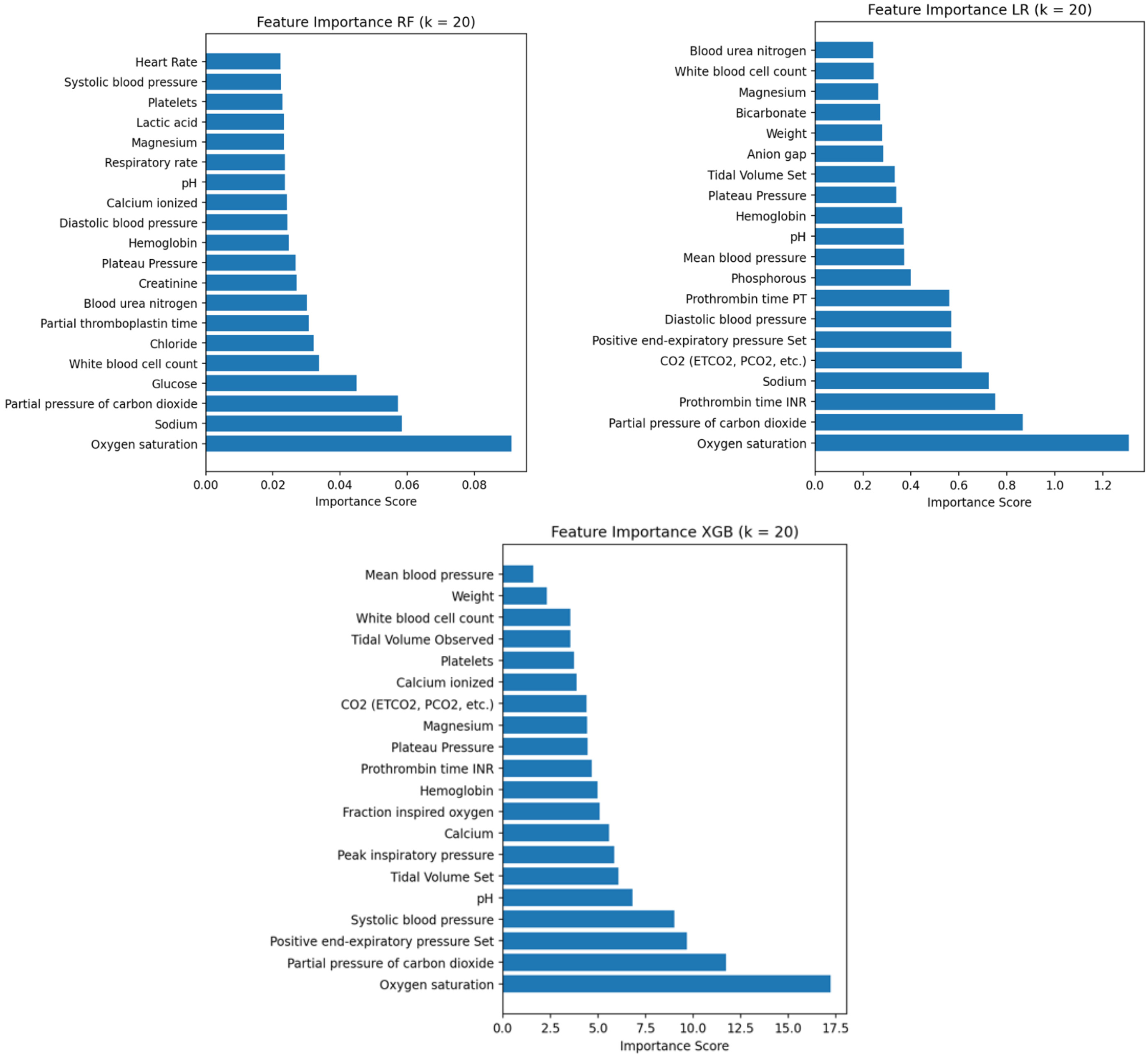

Figure A1: Bar charts showing the feature importance in random forest (upper left), logistic regression (upper right) and XGBoost (lower center). Note that in logistic regression, we show the absolute values of the scores