STAT 215A Final Project

1 Introduction

CT scans are vitally important tools for diagnosing traumatic brain injuries, particularly those that demand urgent intervention. Unfortunately, they come at a cost, particularly for young children. A 2007 publication by David Brenner and Eric Hall estimated that between 1 in 1000 and 1 in 5000 pediatric cranial CT scans result in a lethal malignancy, most frequently cancer.

In spite of these risks, however, CT scans are administered quite frequently. Many pre-verbal and crying children cannot properly communicate their symptoms to physicians, leading to an enormous amount of uncertainty when diagnosing. As such, many physicians err on the side of caution and administer a CT whenever such doubt exists, as failing to do so may result in missing serious and potentially deadly complications. In fact, Kuppermann et al. report that roughly half of American children admitted to emergency departments for head trauma receive CT, most of whom had ostensibly mild head trauma.

These statistics suggest that in spite of their utility, CT scans are largely overused in clinical practice, and have likely led to unnecessary pediatric mortality. Hence, better diagnostic screening tools would enable doctors to more accurately discern the risk of clinically-important traumatic brain injuries (ciTBIs) in children, thus mitigating the problem of unnecessary CT scans. In this report, we use machine learning algorithms to develop and validate rule lists for pediatric ciTBI prediction. Our work expands upon the methods introduced in the Lancet paper "Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study" (Kuppermann et al., 2009).

2 Domain problem to solve

Following the lead of Kuppermann et al., we chose to approach this problem using decision trees, as they are both effective and interpretable. Moreover, they tend to mimic the diagnostic process many domains experts will undergo when questioning a patient, as each node can be understood as a question and each branch an actionable item to their answer. However, our approach prioritizes predictability, so we also experiment with a logistic regression model as well. Thus, our primary objective is, in essence, to create a new set of Clinical Decision Rules substantiated by data as a supplement to conventional wisdom.

One judgment call that may improve the predictability of our model is to split our data by age. Rather than fitting a single model on the entire dataset, we can fit separate models for children under and over the age of 2. Kuppermann et al. do only the latter, whereas we attempt both techniques. Their reasons for doing so are young patients' "greater sensitivity to radiation, minimal ability to communicate, and different mechanisms and risks for traumatic brain injury." When we (and Kuppermann) split by age, each of the two models uses a different set of predictor variables.

We also considered the special case of children with injuries that clearly indicate a CT scan is necessary. According to Dr. Nathan Liu, most physicians will immediately assign a CT scan if the patient has a bulging anterior fontanelle or signs of a palpable or basilar skull fracture. Understanding this to be common practice, we can choose to omit these variables from our data set in order to better understand the relationship between less obvious symptoms and a ciTBI outcome. This is another judgement call that can be specified by the user. As such, these symptoms can be considered an implied fork in subsequent decision trees whereby their existence should immediately merit a CT scan. We note here that this omission runs counter to the methodology employed by Kuppermann et al., who fit each of their models with such variables.

Additionally, Dr. Liu estimated that most physicians will order a CT scan if they think the probability of a ciTBI exceeds 0.5%. Under Kuppermann et al.'s model, this is guaranteed to happen if at least one of six predictors is expressed. With their under-two model, for example, 1040 of 2216 children in the validation set would be assigned CT scans - 49.9% higher than the 694 who actually received CT. For children over the age of two, their model would lead to a 17.5% increase in the number of CT scans. Both of these scenarios constitute large increases in the number of false positives. Thus, using Dr. Liu's 0.5% criterion, their model would exacerbate CT overuse rather than mitigate it.

We attribute the excessively conservative nature of Kuppermann et al.'s model to its hyperfocus on minimizing the number of children with ciTBIs who aren't given CT scans. More technically, it trains to maximize the "negative predictive value" - the rate of children who aren't given CT that don't have ciTBIs - as well as "sensitivity," the rate of children with ciTBIs who are given CT. This objective fails to incorporate the health cost of CT scans, instead recommending that doctors assign significantly more CT scans than is actually necessary. To fix this issue, our models are tuned to maximize metrics like specificity and overall predictive accuracy as well as sensitivity and negative predictive value.

3 Data Collection

3.1 Data Overview

Our training data comes from the Pediatric Emergency Care Applied Research Network (PECARN) and is the same dataset as the one used by Kuppermann et al. In total, it contains samples from 43,399 children hospitalized with head trauma between June 2004 and September 2006, in 25 emergency departments across the United States. Overall, 35.2% of the children in the dataset had CT scans, and 0.89% were reported to have ciTBIs. A ciTBI is defined as one of the following four outcomes occurring:

1. Death from traumatic brain injury
2. Neurosurgery
3. Intubation lasting longer than 24 hours
4. Hospital admission of at least 2 nights.

With 125 variables recorded by medical personnel, we chose to subdivide our dataset into five key categories: demographics, symptoms, administration, hospital actions, and outcomes. Demographic data was defined as anything relating the individual in question to a population-level group, including age, race, ethnicity, and gender. It is worth noting that ethnicity and race were reported by the physician, not the parent or guardian, and it is not specified how gender was determined. These determinations, as such, may be prone to a significant amount of error.

Symptom data was classified as anything relating to the physical or mental state of the patient either before or during their admission to the hospital, including vomiting, altered mental status, and the severity of the incident, among others. Previous incidents (ex: when the patient last vomited) and mental status (ex: whether the patient was behaving abnormally) were reported by either the patient or their guardian, while any physical indicators (ex: skull fractures) were diagnosed by the attending physician.

While many variables were irrelevant for most patients, only 'Dizzy' was found to be missing from a significant amount of patient records (36.8%). Further analysis demonstrated that the average age of patients with missing values for 'Dizzy' was just 2.2, which is approximately 3 times lower than the average age of the overall dataset. As 'Dizzy' was a binary, yes/no variable that had to be answered by the patient, attending physicians likely left the question blank whenever the patient was pre-verbal or unresponsive.

We discuss our data cleaning techniques in depth in section 4.2.

3.2 Initial Exploratory Data Analysis

In order to better understand the contents of the data set, we performed exploratory data analysis. In doing so, we made several noteworthy discoveries that influenced both the data cleaning and model development and validation phases of this work.

First, we found several minor inconsistencies in the raw data relating to the Glasgow Coma Scale (GCS) scores and groupings assigned to certain individuals. First, though the data set was meant to contain only ‘mild’ head trauma cases, 984 instances of severe cases (GCS totals under 14) were reported. Second, 1,289 GCS totals were identified as not matching their associated scores. Of these, only 48 contained scores at all and only 12 suffered a ciTBI.

Second, to better understand current practice as it relates to CT scan prevalence in pediatric medicine, we looked at a variety of metrics relating the administration of a CT scan to other key variables. For example, while only 2.51% of patients scanned were reported to have a ciTBI, no patients with ciTBIs were reported as not having received a CT scan. This suggests that the data provided here consists of not only initial diagnostics, but long-term care updates as well. As we are attempting to create a quick and easy tool for doctors to utilize when first meeting a patient, it is therefore unlikely we will be able to meet this 0% false-negative metric.

Of the 35.3% of patients who were scanned, only 5.2% resulted in a TBI being found. Just under half of these patients (47.9%) with a TBI on their CT scan were later reported to have a ciTBI, meaning that many patients can have a TBI without having a ciTBI. Interestingly, 0.5% of patients with ciTBIs did not have a TBI shown on their CT scan. This suggests that there is also human error involved in the processing and interpretation of the CT scan itself.

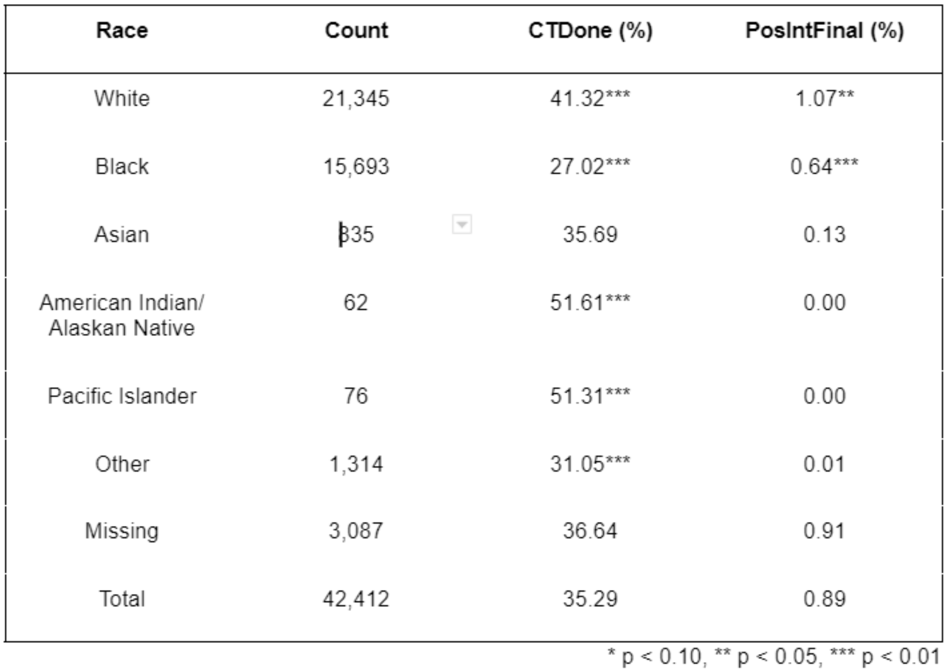
Similarly, we wanted to understand the relationship between TBIs and Other Serious Injuries (OSIs) that patients may have. We found that among the patients with OSIs that were given CT scans, only 3.26% had TBIs and 2.1% had ciTBIs. These are 37.3% and 16.3% lower than their respective values for the general population. Meanwhile, 58.3% of patients with OSIs were scanned, 65.2% higher than the general population. These differences hint that physicians may tend to suspect that trauma in one area is indicative of trauma in another - leading them to prescribe CT scans more frequently to patients who do not need them.

Our modeling ignored demographic features beyond age, due to the potential ethical and legal implications of basing medical decisions on factors like race or gender. That being said, we analyzed the prevalence of these features, as well as their relation to (a) whether a patient was given a CT Scan and (b) whether they were diagnosed with a ciTBI. This analysis is important because demographic trends could still influence the results of our models; simply being “blind” to these sensitive attributes is not sufficient. Results of our race-based analysis are shown in Table 1.

This table contains several notable findings. First, over half of our dataset (50.3%) consists of white and/or white-passing children. Second, these children, along with American Indian/ Alaskan Native children, are significantly more likely to be given a CT Scan than the average population. In contrast, Black children and children of ‘Other’ races are significantly less likely to be given CT. Lastly, white children are much more likely to have a ciTBI than the average population, while black children are significantly less likely to have one than the average population.

A study from 2016 found that doctors prescribed Black patients pain medication at a significantly lower rate. This likely stemmed from “racial bias in perceptions of others’ pain,” as well as “false beliefs about biological differences between blacks and whites” (Hoffman et al.). The relative under-prescription of CT scans found here may be due to similar biases. However, as demographic analysis is beyond the scope of this report, we will refrain from conjecturing as to why these specific trends exist. Similarly, Table 2 outlines the gender breakdown of the data set.

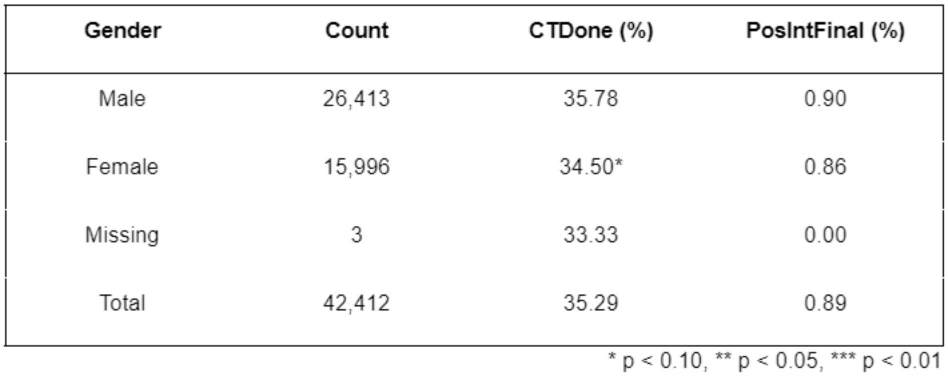
Unlike Table 1, Table 2 is not incredibly noteworthy. We see that 62.3% of the patients are male, and females are marginally less likely than males to be given a CT scan. Table 3 groups this information by age instead of gender.



Race	Count	CTDone (%)	PosIntFinal (%)
White	21,345	41.32***	1.07**
Black	15,693	27.02***	0.64***
Asian	335	35.69	0.13
American Indian/ Alaskan Native	62	51.61***	0.00
Pacific Islander	76	51.31***	0.00
Other	1,314	31.05***	0.01
Missing	3,087	36.64	0.91
Total	42,412	35.29	0.89

* p < 0.10, ** p < 0.05, *** p < 0.01

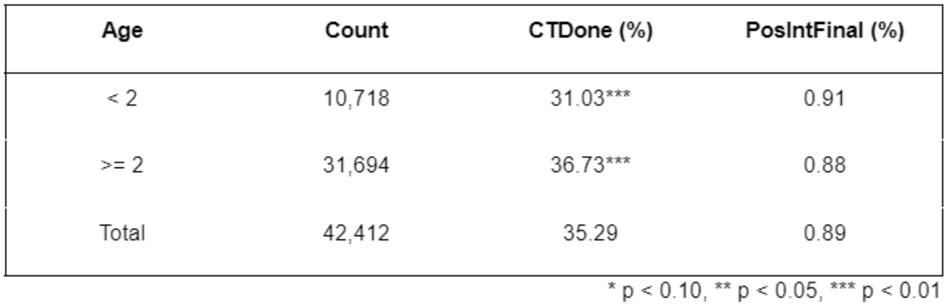
Table 1. Prevalence and Outcome Grouped by Race



Gender	Count	CTDone (%)	PosIntFinal (%)
Male	26,413	35.78	0.90
Female	15,996	34.50*	0.86
Missing	3	33.33	0.00
Total	42,412	35.29	0.89

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 2. Prevalence and Outcome Grouped by Gender



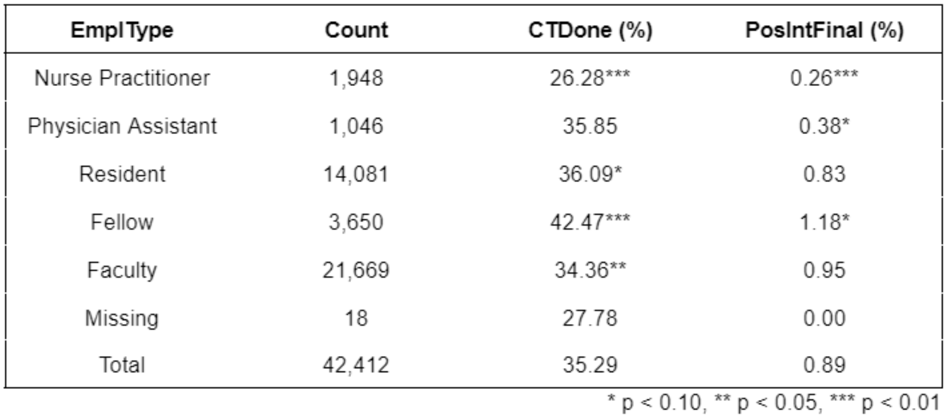
Age	Count	CTDone (%)	PosIntFinal (%)
< 2	10,718	31.03***	0.91
>= 2	31,694	36.73***	0.88
Total	42,412	35.29	0.89

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 3. Prevalence and Outcome Grouped by Age

Interestingly, CT scans were more likely to be given to patients over 2, even though they were no more likely to be diagnosed with a ciTBI. This group was also roughly three times more prevalent in the dataset than younger patients. Thus, the dataset is biased across both racial and age groups in terms of representation, likeliness of CT scanning, and rate of ciTBI diagnosis.

While administrative variables were also omitted in modeling, several findings are worth noting. Table 4 displays the same trends discussed with demographic data divided by the employment type of the physician filling out the datasheet.

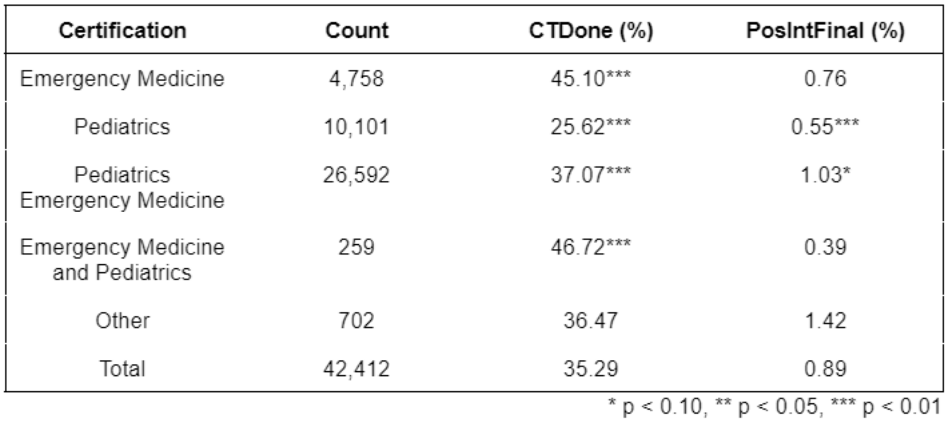


EmplType	Count	CTDone (%)	PosIntFinal (%)
Nurse Practitioner	1,948	26.28***	0.26***
Physician Assistant	1,046	35.85	0.38*
Resident	14,081	36.09*	0.83
Fellow	3,650	42.47***	1.18*
Faculty	21,669	34.36**	0.95
Missing	18	27.78	0.00
Total	42,412	35.29	0.89

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 4. Prevalence and Outcome Grouped by Employment

Notably, Nurse Practitioners were the least likely to give CT scans and to see patients that ended up having ciTBIs. This result, however, is not terribly surprising: Intuitively, patients thought to be in obvious medical danger should likely be seen by an MD instead of, or along with, a nurse. Additionally, Fellows were the most likely to prescribe a CT scan and have a ciTBI. Again, this is a sensible finding, as they are more likely than other medical practitioners to see patients in serious or critical conditions. Similar results are shown in Table 5, which breaks down the dataset by specialty of the attending physician filling out the datasheet.



Certification	Count	CTDone (%)	PosIntFinal (%)
Emergency Medicine	4,758	45.10***	0.76
Pediatrics	10,101	25.62***	0.55***
Pediatrics Emergency Medicine	26,592	37.07***	1.03*
Emergency Medicine and Pediatrics	259	46.72***	0.39
Other	702	36.47	1.42
Total	42,412	35.29	0.89

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 5. Prevalence and Outcome Grouped by Certification

Curiously, those trained in Emergency Medicine (separately from Pediatrics) were significantly more likely than other physicians to give a CT scan despite seeing patients no more likely than average to have a ciTBI. Pediatrics doctors, on the other hand, send significantly fewer patients to get a CT scan. These disparities are likely more attributable to the differences in training each receives, as opposed to the type of patients they attend to.

4 Meaning

4.1 Variables

We did not incorporate all 125 variables into our model, since most were not relevant to our modeling task. Rather, we began by taking a subset of roughly 50 features, most of which concerned patients' symptoms. These variables were subsequently passed into a feature extraction algorithm that returned the ones with the highest predictive efficacy. We discuss the meaning of these variables in this section, as well as the variables used by Kuppermann et al. The feature selection procedure itself will be described in section 10. Note that the selected features differ based on preprocessing decisions. Here, we focus on the ones extracted with default judgement calls:

1. Features can have up to 10% of their values missing
2. Missing values for any variable except the outcome are imputed with the median
3. Do not exclude patients with clear signs that a CT scan is necessary
4. No splitting by age

Our model selected the following features:

AMS_1: Whether the patient has an altered mental status (1 is true, 0 is false)

Clav_0: Whether the patient has evidence of trauma (including laceration, hematoma, and abrasion) above the clavicles

Drugs_1: Whether the patient was suspected for drug or alcohol use

FontBulg_1: Whether the patient has a bulging anterior fontanelle

GCSEye_3: Whether the patient opens their eye in response to speech

GCSMotor_6: Whether the patient is capable of following motor commands

HA_verb_1: Whether the patient has a headache at the time of the evaluation

HemaLoc: If the patient has a hematoma, whether it is frontal (1) or either occipital or parietal/temporal (2_or_3)

High_impact_InjSev_3: Whether the injury mechanism was severe

- Motor vehicle collision with patient ejection, death of another passenger, or rollover
- Pedestrian or bicyclist without helmet struck by a motorized vehicle
- Falls of > 5 feet for patients 2 yrs and older
- Falls of > 3 feet < 2 yrs
- Head struck by a high-impact object

LOCSeparate_1_or_2: Whether the patient has a history - suspected or confirmed - of loss of consciousness

NeuroD_1: Whether the patient has a neurological deficit (other than mental status)

SFxBas_1: Whether there are signs that a basilar skull fracture has occurred.

SFxPalp_1_or_2: Whether the exams for a palpable skull fracture were either positive or inconclusive

SeizOccur_2_or_3: Whether the patient had a seizure more than 30 minutes after the event

SeizLen_3_or_4: Whether the patient had a seizure lasting over 5 minutes

Seiz_1: Whether the patient had a seizure

VomitNbr: Whether the patient vomited once (1) or multiple times (2_or_3)

VomitStart: Whether the patient started vomiting within (2) or after (3_or_4) an hour of the event

The Baseline model used the following variables:

AMS_1: Whether the patient has an altered mental status

ActNorm_0: Whether the patient is acting abnormally

AgeTwoPlus_1: Whether the patient is under 2 years old

HASeverity_3: Whether the patient has a severe headache

HemaLoc_2_or_3: Whether the patient has either an occipital or pariental/temporal hematoma

High_impact_InjSev_3: Whether the injury mechanism was severe

LocLen_2_3_4: Whether the patient lost consciousness for longer than 5 seconds

SFxBas_1: Whether the patient has signs of of a basilar skull fracture

SFxPalp_1_or_2: Whether the exams for a palpable skull fracture were either positive or inconclusive

Vomit: Whether the patient vomited after the injury

All of these variables are equal to 1 if they are true, and 0 otherwise. Note that every variable is categorical, making them an excellent choice for decision trees.

4.2 Data Cleaning

To better address our specific domain problem, a series of data cleaning steps were performed. First, as we are concerned only with diagnosing ciTBIs in patients considered 'borderline' by physicians, samples were limited to children with head trauma deemed 'mild.' This was done using the Glasgow Coma Scale (GCS), which is defined as the sum of the patient's Eye, Motor, and Verbal scores. Higher GCS values indicate less serious head trauma, with a maximum score of 15 signifying that each of these three areas is performing normally. Per the advice of Dr. Liu, we omitted all 2,255 patients whose total GCS scores were less than 14, as they would automatically be given a CT scan regardless of other symptoms. In addition, one of our judgement calls excludes 923 patients with clear signals that a CT scan is necessary.

Our chosen response variable, PosIntFinal, signifies whether or not the patient has a ciTBI. Recall that ciTBIs are defined as either death, neurosurgery, intubation longer than 24 hours, or hospitalization for at least 2 nights. We cross-referenced these four outcome variables to find that PosIntFinal had 185 false negatives and 3 false positives, which we fixed. It also had a number of missing values, 8 of which remained after this imputation. We removed the 8 rows corresponding to these values.

We excluded features like dizziness that had too many missing values. Of course, what constitutes "many" is subjective, so we designated it as a judgment call. We let 10% be the default amount, but also fit models for a 1% threshold.

We also removed a number of features without direct relevance to the task at hand. We excluded variables pertaining to other substantial injuries (OSI) - that is, injuries to regions other than the head. Intuitively, whether or not someone broke their arm doesn't tell us valuable information about the extent of their head trauma. In addition, we did not use any features categorized as "administrative," "hospital actions," and "outcomes." These variables are either irrelevant to our domain problem or cannot be used during initial diagnosis.

After removing unwanted rows and columns, we preprocessed our data by imputing missing data. This was another judgment call, for which we implemented three options: No imputation, imputation with the median value, and imputation with nearest-neighbor values. To accelerate computation for nearest-neighbor imputation, we only computed similarity between each patient and 10 randomly selected others. This, however, still took much longer than imputation with the median value. For this reason, we removed it from our dictionary of judgment calls when considering all possible perturbations.

Subsequently, we encoded all categorical features as binary variables. Here, each binary variable denotes whether or not the patient had some particular value for that feature. This schema required representing features with more than two levels as multiple column variables. In that case, each row has a 1 for exactly one column in the set of binary variables for a given feature. Finally, we split our data by age, if specified by a judgment call.

5 Relevance

Our collected data can predict ciTBI risk in part, but ideally, more data points would be available. In particular, it would be helpful to know other clinical variables such as height, weight, heart rate, and blood pressure. Doctors are given this data from the physical examination, meaning it is available in practice but not to us.

In addition, a patient's medical history plays a critical role in the doctor's decision. For example, a child with multiple previous ciTBIs should be far more likely to receive a CT scan, even if the other predictors suggest otherwise. While predictors like "history of loss of consciousness" explore medical history, it would be better if more features were available. The accuracy levels of the baseline models presented in the paper are quite low (around 60%), so their results based on this data do not wholly answer the substantive question. Even our improved models have accuracy scores of less than 90%.

However, a large part of this is due to limitations in the modeling and quantity of data points. It is well known that interpretability often comes at the expense of predictive accuracy, and this situation is no exception. It is highly likely that a neural network with more of the same type of data points would achieve much higher accuracy scores. Future work in fields like interpretable deep learning may allow future medical research to fully utilize the predictive power of these more complex methods while still maintaining the trust of the actual medical personnel who will use them.

6 Translation

The issue of unnecessary CT scans stems from doctors needing to quickly weigh the immediate risk of an undiagnosed TBI against the long-term effects of radiation exposure and other side effects. How can we use data to decrease the number of unnecessary CT scans while still limiting the number of false negatives? This question is best formulated as a prediction problem. In order to help physicians better determine whether a given patient with mild head trauma is in need of a CT scan, we must create a model that is capable of predicting the probability a given patient is at risk of a ciTBI. Ideally, accurate predictions would enable doctors to more confidently determine that a scan is not needed.

While providing accurate predictions is the primary goal of our work, it is also reasonable to ask if the data can provide insight into which of the 100+ variables are actually useful when making this decision. The two translations are not mutually exclusive; our logistic regression model is primarily geared towards prediction and our decision tree is more useful for inference, but both provide probabilities of a ciTBI and select meaningful features.

When producing models that are intended for medical settings, it is essential to incorporate elements of both translations. If prediction was our only goal, less transparent methods like random forests and deep neural networks would likely perform better at the cost of interpretability. For obvious reasons, few doctors would welcome such algorithms into the examination room. On the other hand, too much emphasis on inference produces models that are equally ill-suited for practical applications. The information gained from these models would help doctors reconsider their own thought processes, but every inaccurate prediction has a direct impact on a real patient's life.

7 Comparability

While cleaning the data, it became apparent that many of the variables could not be directly compared in their raw form. Most were categorical and 4 were nonbinary. Those that were binary also differed at times between 0 and 1 choices and 1 and 2 choices based on the nature of the data being collected. With these differences in mind, and as discussed in the Data Cleaning section of this report, all categorical variables were one-hot encoded and any remaining numerical variables were removed from the data set per the lab instructions. As a result, all categorical variables have identical units after cleaning and can easily be compared.

Additionally, there existed three sets of variables which contained redundant information: age in months and age in years, mechanism of injury and severity of injury (as the injury mechanism was used to determine the severity), and the total GCS score and the GCS Group assigned to the patient. To avoid any instability that may result in our models should these redundancies remain, age in years, mechanism of injury, and the GCS Group of each patient were removed from the cleaned data set.

More importantly, however, each of the one-hot encoded variable groupings was highly dependent on one another. For example, if a patient was given a score of 1 for AMS_0, they were, by definition, also given a score of 0 for AMS_1. As such, prior to training predictive models, the first value from each group was dropped from our training set in order to ensure that the resulting coefficients were, in theory, not entirely dependent on one another. As each of the variables were binary, no additional normalization was needed to prep the data for model creation.

8 Visualization

As the majority of the data used for our analysis was binary, we did little in the way of visualizing the range and distribution values of our features. However, Tables 1-5 in the Data Exploration section detail the prevalence of samples across various demographic and administrative groups as well as their tendency to be given (or give) CT scans and the frequency with which they are diagnosed with (or diagnose) a ciTBI. Furthermore, Figures 1-2 demonstrate the correlation values inherent in our selected features, while Figures 3-5 provide graphical representations of the three models developed for this report.

9 Randomness

As discussed in the Data Exploration and Data Cleaning sections of this report, there were quite a few significant trends underlying this dataset. White patients were more likely to have ciTBIs than average, black patients were less likely to be administered a CT scan, and the physician's title and certification were heavily correlated with both the outcome and whether their patients were given a CT scan. As such, it is reasonable to assume that biases exist within the dataset.

However, it would be difficult to conduct a truly randomized experiment that relies so heavily on patient data. Certain categories are tilted heavily toward one demographic, as mentioned in Section 3, but so are certain populations. Moreover, while these disparities do exist, there is no scientific evidence indicating that ciTBIs differ along such lines. In other words, the patients seeking medical attention for a TBI can be assumed to be a random sampling of the overall population unless evidence to the contrary is provided, regardless of the degree of care provided or the ultimate outcome of their condition.

That being said, this implies that we are making an inherent assumption, not about who can get TBIs, but who seeks medical help when confronted with one. It is not unlikely that certain individuals would refrain from seeing a physician due to cost, societal norms, or mistrust in the medical community, nor is it unreasonable to assume that such hesitations are more prevalent in certain demographics than others. Therefore, while we can assume that the data outlined above is as close to a random sample as is feasible for such a study, there is by no means a guarantee that it properly reflects the population-level dynamics of the area.

10 Stability

While we considered many off-the-shelf modeling techniques including Bayesian rule lists, greedy rule lists, logistic regression, and RuleFit, we ultimately settled on a decision tree and logistic regression models for the high degree of interpretability and accuracy they provide. Each will be discussed in more detail here.

10.1 Feature Selection

Features were selected for our decision tree and logistic regression models using Python's sklearn package. Each of the relevant variables was fed into a feature selection algorithm to determine which were most important in determining our response variable. Those below a certain threshold were deemed ineffective and removed. This process was then repeated with the important variables until fewer than 30 remained. All variables that remained were then fed into the decision tree algorithm to create our final model. A correlation matrix of these variables is shown in Figure 1.

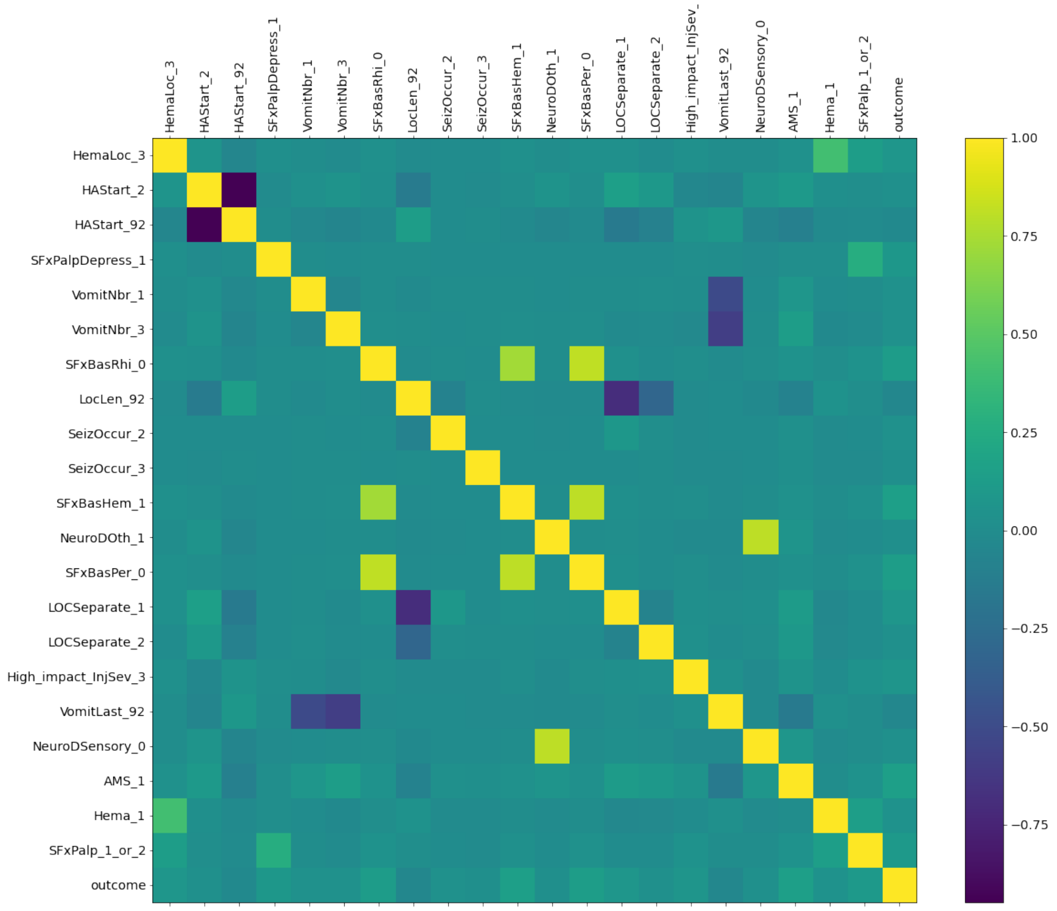


Figure 1. Correlations Between Decision Tree Features

For the logistic regression model, however, an additional step was implemented. As this model was made with multiple-choice, not true/false, questions in mind, features were recombined with their associated groupings prior to training the model. This means that if HA_verb_1 was deemed important, HA_verb_2 would also be included in the model. This adjustment was made to allow more nuance and interpretability of logistic regression coefficients produced. As a result, the features used to train this model were slightly different than those used in the other two, as can be seen by the correlation matrix shown in Figure 2.

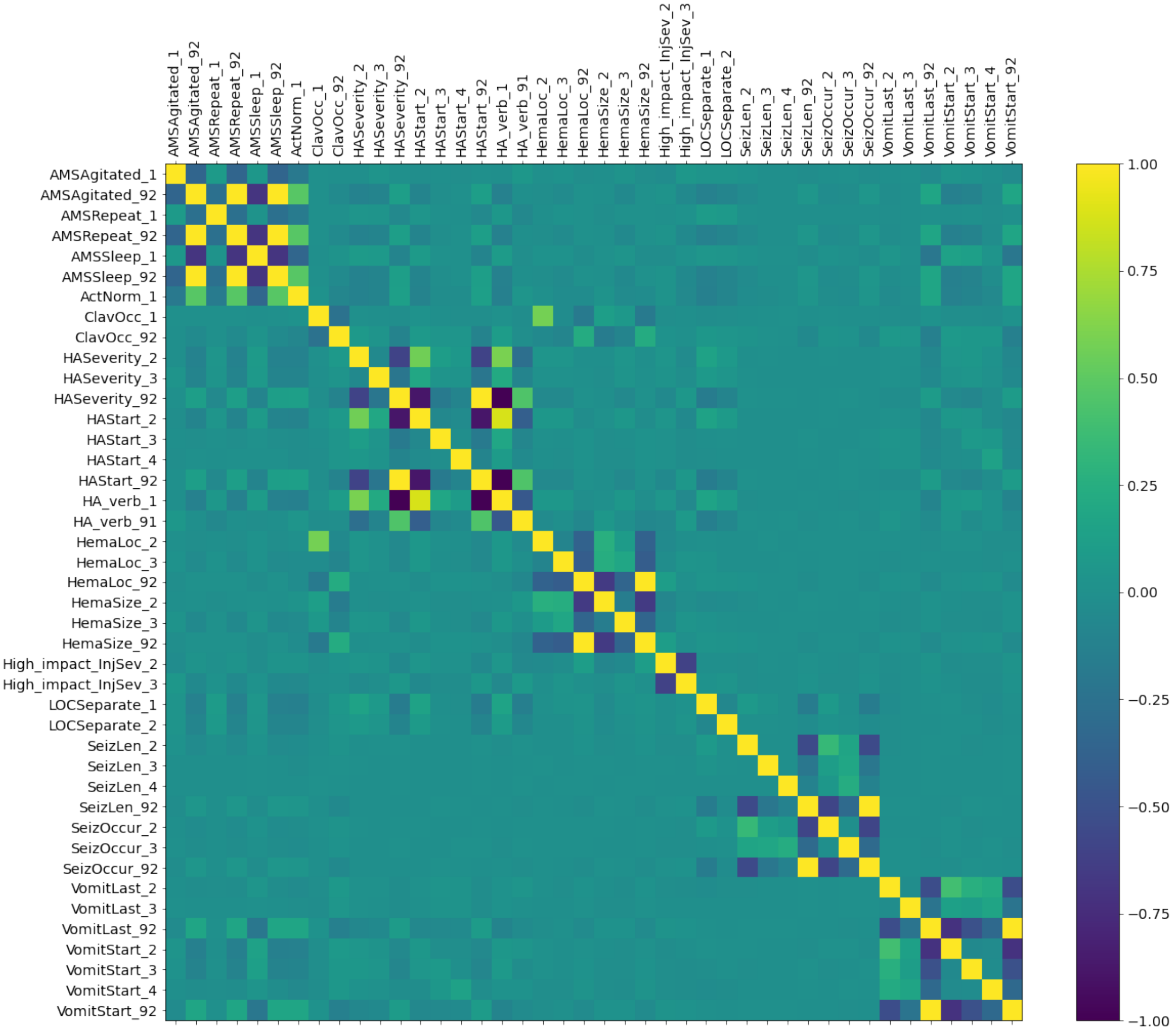


Figure 2. Correlations Between Logistic Regression Features

10.2 Greedy Rule List (Baseline)

The models used in the paper by Kuppermann et al. are greedy rule lists, decision trees forced to follow a linear path. As a result, the rules are easy to read and interpret. In an effort to replicate their results, we followed their data cleaning and feature selection process as closely as possible when setting up the baseline models (these processes were separate for our new models). However, there were several instances where our workflows diverged. For example, we choose to eliminate entries with NA values for any of the six predictors while the authors keep them in the model. While we use the same splits and features, our training algorithm chose to arrange them in a different order. Finally, our randomized test and training split certainly differs from that used in the paper. All of these factors together allow us to perform a massive stability check on the original results.

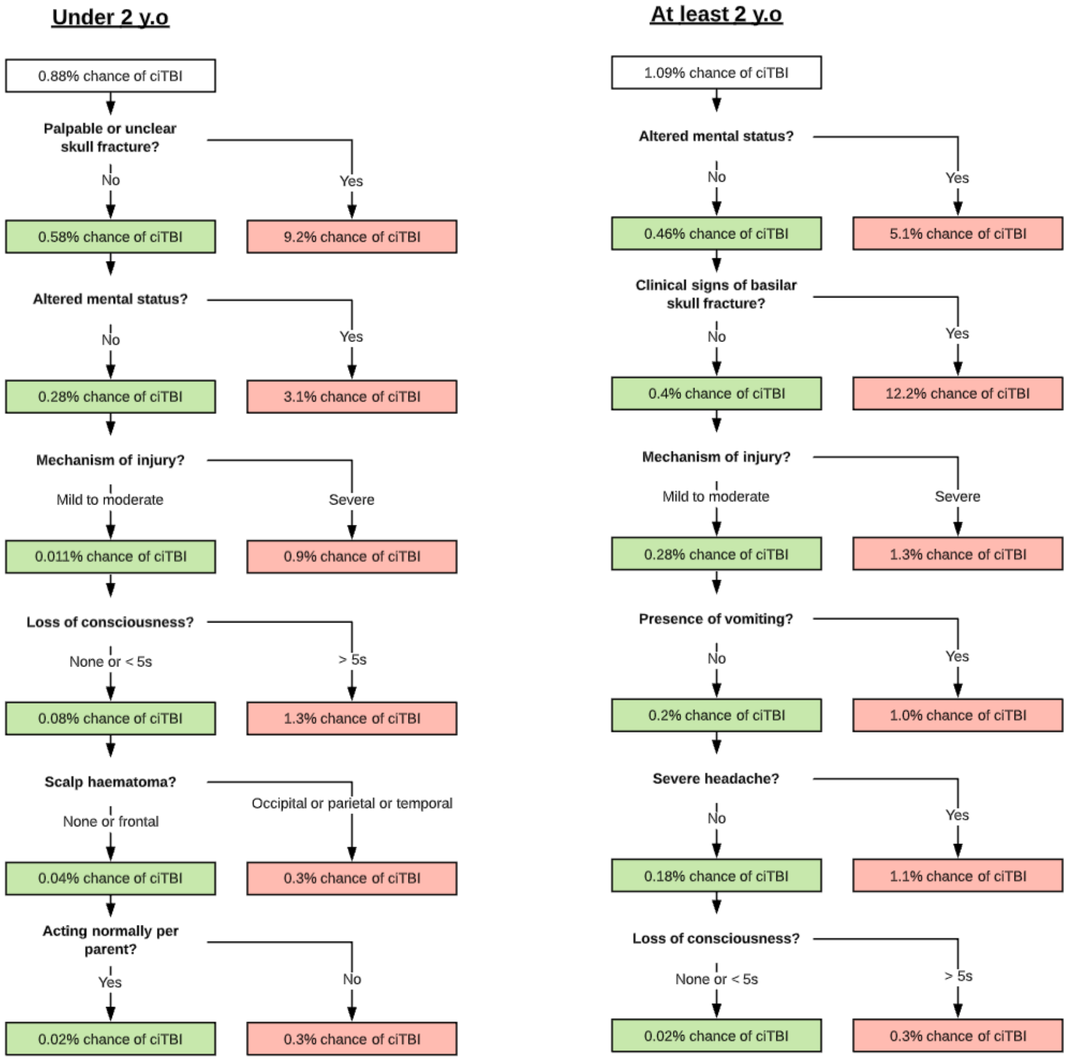


Figure 3. Results of Baseline Decision Tree Replication

As discussed in the first section, the baseline models are overly conservative. In patients younger than 2 y.o., our evaluation metrics are similar to those in the paper. On the validation dataset, accuracy is 0.669, specificity is 0.667, sensitivity is 0.886, and the negative predictive value is 0.998. While this model performs exceptionally well when only minimizing false negatives, it actually assigns many more unnecessary CT scans.

In patients at least 2 y.o., cracks begin to show in the baseline models. On the validation dataset, accuracy is 0.324, specificity is 0.327, sensitivity is 0.08, and the negative predictive value is again 0.998. The reasons for these interesting numbers are twofold. First, the low number of positive observations means that metrics like sensitivity can vary significantly when even a few patients are misclassified. Second, the use of two models and a training/testing data split leaves very few of these positive signals to be evaluated. Both of these problems could be remedied by collecting larger amounts of data or by not splitting the data into two, as we will see later. Decision trees are inherently unstable and sensitive to changes in the training data, and this simpler model is even more so.

Our new models place higher importance on accuracy and specificity while also maintaining high sensitivity and negative predictive values. This approach is necessary to solve the problem of unnecessary CT scans in children. While the authors of the paper provided an important starting point, they are still too hesitant to mislabel a child who actually has a ciTBI. While understandable, this is the same sentiment that leads to nonessential CT scans in hospitals across the country.

10.3 Decision Tree

As mentioned previously in section 4, we didn't put all the variables into the model and only used 21 variables which represent 18 original variables, and restricted our focus on the patients with GCSscore 14 or 15. We implemented the DecisionTreeClassifier function in the sklearn library with maximum depth equal to 8 and weighted the class because of the imbalance. After talking with the doctor and reading the original paper, we thought that though the decisions along the clinical process didn't always form a tree, we can derive rules from the branches of a tree. The RuleFit algorithm also adopts similar ideas by training random forests and weighting the extracted rules from the forest. We selected the depth of the tree by comparing the accuracy for different depths using cross-validation on the tuning data set. When it equals 8 the average accuracy is around 0.82, which is relatively higher than 7. When trained to a depth of 9, the decision tree performed well on the training data but generalized poorly. This is a sign of overfitting, so we kept the decision tree with a maximum depth equal to 8. The first two levels of this model are shown in Figure 4.

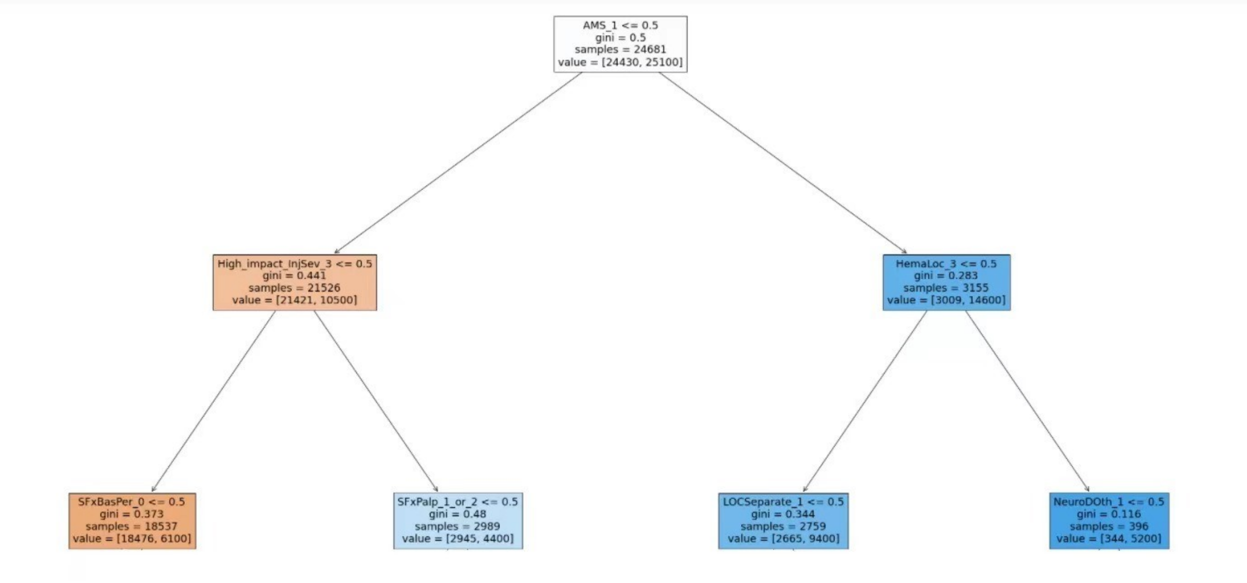


Figure 4. First Two Layers of Decision Tree Model

No matter how we specify the hyperparameters, "Altered Mental Status" is always selected as the root of the decision tree. Since we already split the patients by GCSscore as well as other mildest trauma, this AMS can further indicate whether the patient has suffered any other noticeable trauma. The second important feature is 'High_impact_InjeSev_3' which stands for injury severity based on the injury mechanism classification. This may overlap with 'AMS_1' in meaning, but reflects more detailed information about patients. For those patients who didn't have altered mental status or severe injury mechanisms, it comes to the skull fracture diagnosis. As we get deeper into the tree, there are many more options. For example, the third layer is mostly about various kinds of skull fractures. Additionally, high vomiting numbers, loss of consciousness, and hematoma location play vital roles in the decision-making process.

We then evaluate our decision tree training results on the test data set. The negative predictive value of the decision tree is 0.996, sensitivity is 0.77, specificity is 0.81, and the accuracy is 0.81. These results match our expectation of achieving higher values in these four aspects.

In the data cleaning process, we have four critical judgement calls to extract our final training data, tune data, and test data. By running all the combinations of different choices, we had 24 versions of data groups, each of which contained a training set, tune set, and test set. We re-train the decision tree with a maximum depth equal to 8 and weight the classes by the approximated proportions found in the dataset. We then evaluate the 24 trees based on negative predictive value, sensitivity, specificity, and accuracy. In terms of sensitivity, none of them exceeded 0.76 and most of them are around 0.5. The negative predictive values are all around 0.99. The accuracy is between 0.7 and 0.9 and specificity has a maximum of 0.9. We found that some combinations may have some advantages in one or two aspects, but they performed worse in the other assessments. Thus, our model is stable and produces useful results.

Moreover, the training process of the decision tree contains some randomness, and setting different random states does not significantly influence the values we care about (+0.001).

10.4 Logistic Regression

Features were also passed into a Logistic Regression algorithm provided by the python library sklearn. Unlike with previous models, we ensured the interpretability of our results by omitting any specialized weighting metrics for a given outcome. This omission guarantees that the resulting coefficients properly represent percentage scores and can be interpreted as such. In other words, given our model, physicians would be capable of quickly and accurately tying a percentage likelihood for ciTBIs to each patient they see.

Once trained, the intercept and coefficients of the model were stored in a yaml file for easy access in the future. Coefficients, grouped by variable, are shown in Figure 5. Variables omitted to achieve linear independence were stored with a coefficient value of 0. External storage also allowed us to predict patient outcomes without retraining the model for each iteration, leading to a significant decrease in runtime.

In order to compare this model to both the original baseline decision tree and our own, in-house decision tree, its accuracy, negative predictive value, sensitivity, and specificity were calculated to be 0.725, 0.998, 0.878, and 0.723, respectively, using a 0.5% threshold for the administration of a CT scan.

Furthermore, when applying this model to 25 bootstrapped versions of our original dataset, none of these values changed by more than 1 percentage point, suggesting the model is not only very accurate, but highly stable as well. Additional tests with altered judgement calls in both feature selection and model creation further validated this assessment.

To better demonstrate the efficacy of this model, we also created a simple, yet powerful, app using the python Flask library along with the cloud platform Heroku. This site is available to the general public and, at the time of writing this, consists of a list of just 16 multiple choice questions tied directly to the features we used to build our models. The questions were crafted in such a way that anyone, physician or otherwise, can answer them and receive a probability score denoting the likelihood that they have a ciTBI. To ensure full transparency, the app also provides a disclaimer encouraging patients to seek medical help if they believe they have a TBI and links to see the variables and weights we used to develop and run our model. This app can be found *here*.

11 Conclusion

Therefore, while no model will ever be perfect, we believe that the models outlined here signify a marked improvement in the predictability and classification of ciTBIs in children. Our combination of interpretability, real-world domain knowledge, and standard data science models creates a powerful, stable, and easy-to-use set of tools for pediatricians to add to their arsenal moving forward. Moreover, the generalizability and standardization of our approach demonstrates that such a strategy, if implemented correctly, can likely be

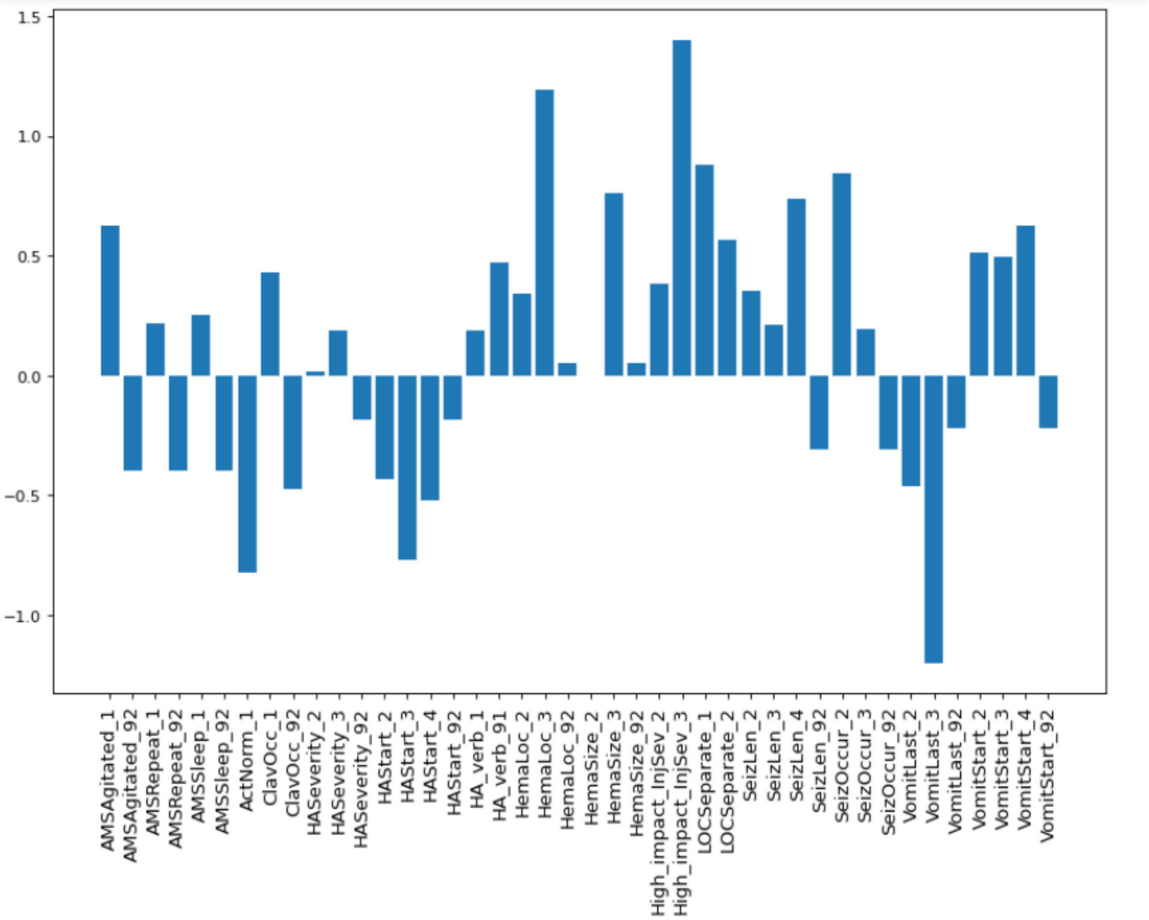


Figure 5. A Breakdown of the Coefficients Used in the Logistic Regression Model

used in other clinical applications as well, thereby helping to usher in a new generation of better equipped, better informed, and more confident physicians moving forward.

12 Statement of Academic Integrity

Dear Bin,

All work shared in this document that is not entirely our own is properly cited in the References section. As you know, we believe that academic integrity and honesty amongst researchers is paramount to the success of scientific endeavors everywhere, from the classroom to the lab, as it empowers us to explore, develop, and invent without fear of copyright or usurpation. Without it, few findings can be shared, virtually no collaboration can occur, and little true progress can be made. Therefore, we will continue to do our level best, both this semester and beyond, to give credit where it is due and ensure that we create work that is entirely original.

13 References

Brenner, David J., and Eric J. Hall. “Computed Tomography — An Increasing Source of Radiation Exposure.” New England Journal of Medicine, vol. 357, no. 22, Nov. 2007, pp. 2277–84. DOI.org (Crossref), https://doi.org/10.1056/NEJMra072149.

Hoffman, Kelly M., et al. “Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs about Biological Differences between Blacks and Whites.” Proceedings of the National Academy of Sciences of the United States of America, vol. 113, no. 16, Apr. 2016, pp. 4296–301. PubMed Central, https://doi.org/10.1073/pnas.1516047113.

Kuppermann, Nathan, et al. “Identification of Children at Very Low Risk of Clinically-Important Brain Injuries after Head Trauma: A Prospective Cohort Study.” The Lancet, vol. 374, no. 9696, Oct. 2009, pp. 1160–70. DOI.org (Crossref), https://doi.org/10.1016/S0140-6736(09)61558-0.

14 Acknowledgements

We would like to give a special thanks to Dr. Nathan Liu, Dr. Bin Yu, Omer Ronen, and Chandan Singh for their support and guidance in understanding, characterizing, and addressing this incredibly important domain problem.