Lab 1 - PECARN TBI Data, STAT 214

Ruiwen Liu
February 22, 2025

Introduction

While CT scans are an important tool for diagnosing brain injuries, oversue can lead to unnecessary radiation exposure and increased healthcare costs. The primary goal of this report is to identify important clinical features to predict which patients with traumatic brain injueries (TBI) are at very low risk of clinically-important TBI (ciTBI) and can avoid unnecessar CT scans. This report is based on the study of "*Identification of Children at Very Low Risk of Clinically-Important Brain Injuries After Head Trauma: A Prospective Cohort Study*" by Nathan Kuppermann et al. (2009). The study developed evidence-based guidelines (PECARN rules) to help emergency departments decise when a CT scan is necessary for patients with head trauma.

Understanding this data is important because it has clinical implications: reducing unnecessary CT scans can lower radiation risk for patients while still ensuring that ciTBI are properly identified. By exploring the dataset, we can assess how well different clinical factors predict ciTBI and if a data-driven approach can enhance clinical decision-making.

In the rest of the report, we will firstly describe the dataset and our preprocessing steps. Next, we will conduct exploratory data analysis (EDA) and visualization to uncover important patterns and relationships between variables. Based on the findinsg, we will then build interpretable model to predict which patients are at low risk of ciTBI and check the stability and interpretability of the model. Finally, we will discuss the limitations of our analysis and potenial areas of improvement in applying these finding to real-orld settings.

Data

The dataset used in this analysis comes from the study of Kuppermann et al. (2009), which was conducted across 25 emergency departments in the Pediatric Emergency Care Applied

Research Network (PECARN). The dataset includes patient demographic features as well as a variety of clinical factors. The dataset also tracks results from CT scans and hospitalization. To be more specific, we can divide the entire dataset into 6 sections:

- Mechanism of injury
- Clinical variables: history and symptoms
- Clinical variables: physical examination findings
- Other information collected on case report form
- Clinically-important traumatic brain injury (ciTBI)
- Traumatic brain injuro on CT

This dataset is relevant because our goal is to predict patients who do not need a CT scan to reduce the exposure to radiation. By analyzing clinical factors we mentioned before, we can identify which factors are more important in predicting ciTBI status. Then, we can identify patients with low risk of ciTBI who do not need a CT scan. This will help emergency departments improve clinical decision-making and protect patients from overuse of CT scans.

Data Collection

Patient history, injury mechanism, and clinical symptoms were recorded by trained site investigators and emergency department physicians using a standardized data from before they know the CT results. This make sures that CT findings didn't influenced the pre-CT factors. ciTBI was defined by death from TBI, neurosurgery, intubation for more than 24 hours, or hospitalization for at least two nights due to TBI.

To ensure the dataset focuses on mild TBI cases, patients with Glasgow Coma Scale (GCS) < 14 were exlcuded. Also, cases involving trivial injueries, penetration trauma, pre-existing nuerological conditions, ventricular shunts, or bleeding disorders were removed. Any patients with missing ciTBI should also be exlcluded.

For clinical and injury-related features, except for GCS score, most of them were recorded as categorical variables. Injury mechanisms were categorized and classified by severity. Symptoms such as vomiting and seizures were recorded as binary values. Other symptoms were recorded as categorical scales.

Data Cleaning

The two most significant challenges in this dataset were unrelated columns and missing values. The original dataset contained 125 columns, but not all were relevant to our goal of predicting whether a patient needs a CT scan. Based on domain knowledge and the PECARN clinical prediction rule, we removed variables that did not contribute to this decision-making process.

Removing Unrelated Variables

- **CT Findings and Results**: Any variable related to CT scan results was excluded because these findings are only available after a scan is performed and thus cannot be used for pre-CT decision-making.
- **Post-Diagnosis Hospitalization and Disposition**: Variables related to hospital stay, discharge status, and final outcomes were removed since they are not part of the initial risk prediction.
- **Demographic Variables**: Variables such as gender, race, and ethnicity were excluded as they do not have strong predictive value for ciTBI. However, age was retained since it is a key factor in PECARN risk stratification.

Removing Highly Correlated Variables

Before handling missing values, we also want to removed some highly correlated variables to reduce redundancy and improve model efficiency. Figure 1 and Figure 2 demonstrate the correlation between some clinical features. For example, `GCSEye`, `GCSVerbal`, and `GCSMotor` are all strongly correlated with `GCSTotal`, since GCSTotal is a sum of these three components. Instead of keeping all four columns, we retained only `GCSTotal` as it provides the same information in a more compact form. This approach helps prevent multicollinearity, simplifies the dataset, and ensures that our model focuses on the most informative features.

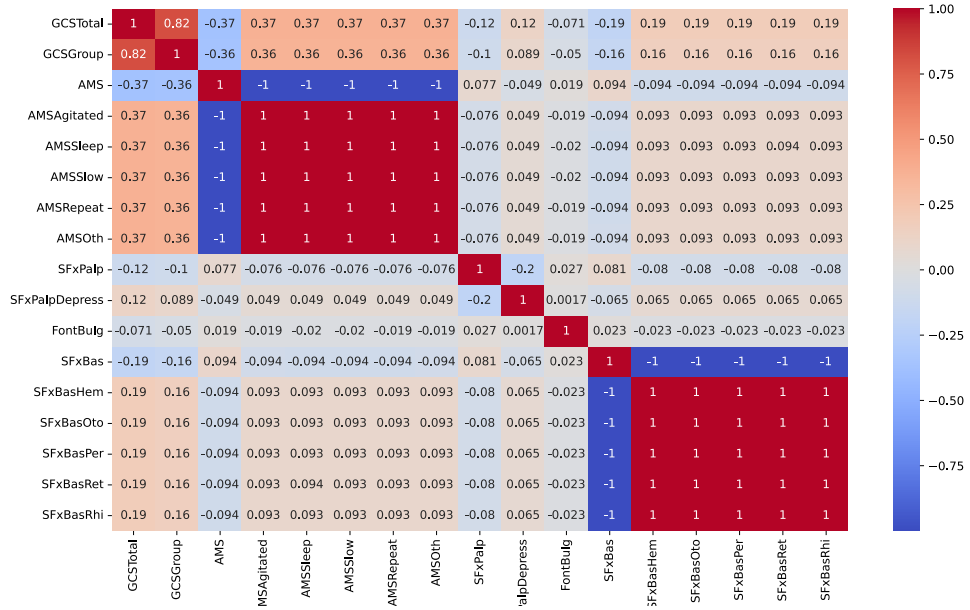

Figure 1: Correlations Between Some Features

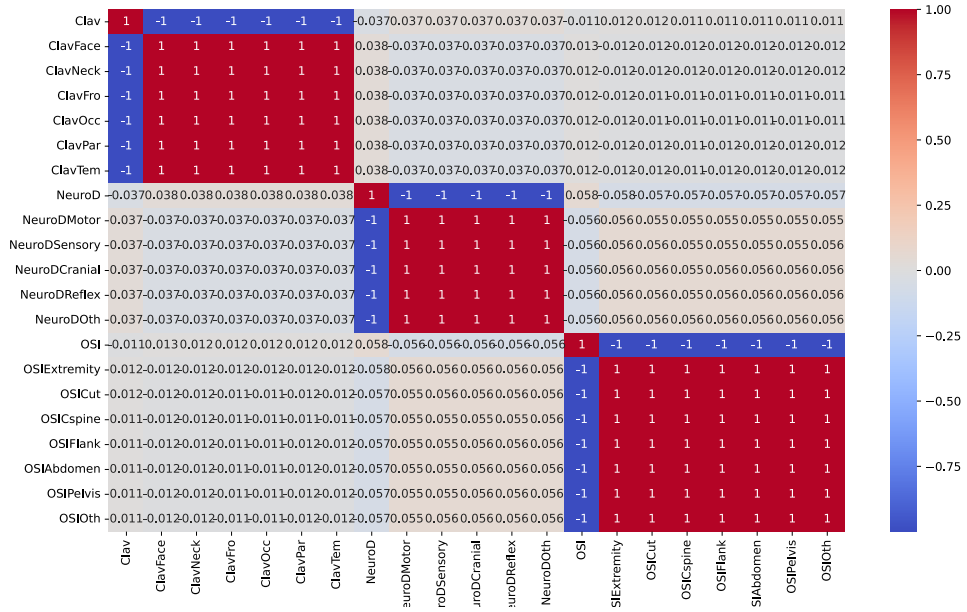

Figure 2: Correlations Between Other Features

After filtering out these variables, we were left with 35 key features relevant to our analysis.

Handling Missing Values

After reducing the dataset to 35 variables, we observed that many features still contained missing values. To analyze the extent of the problem, we generated a histogram showing the distribution of missing values across all variables (Shown in Figure 3). Based on the distribution, we can see most of variables have missing values from 0 to 500. There are several variables with missing values more than 1000. There's an outlier with missing value more than 15000, which is `Dizzy`.

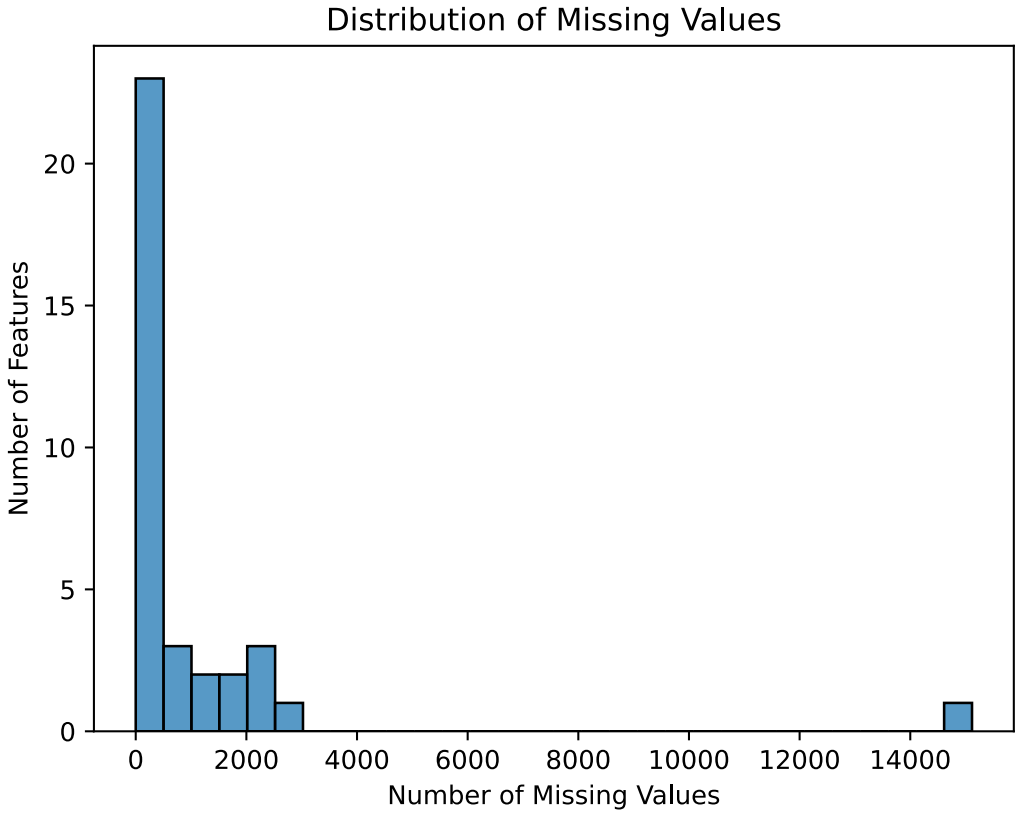

Figure 3: Distribution of Missing Values

To handle missing values effectively, we used different imputation strategies depending on the nature of each variable:

1. Domain-Specific Inputation:

- Some values recorded 91 as a category for pre-verbal/non-verbal patients (e.g., young infants who cannot respond to certain questions). If a variable was missing for a patient younger than one year, we assumed the missing value was probably due to the child being pre-verbal and thus assigned 91 to replace NA.
- Some variables were dependent on other variables. If the primary variable was no or missing, the dependent variable would be considered not applicable recorded as 92. We replaced such cases with 92 to keep consistency.

2. Mode-Based Imputation for Related Variables:

- If a variable was closely related to another feature (e.g., symptoms or injury severity), we replaced missing values using the most frequent value within each category. This helped to preserve the relationships between variables.

3. Unrecoverable Missing Data:

- For other cases when a missing value could not be inferred, we replace NA by 999. The reason we chose 999 is that tree-based machine learning algorithms can easily distinguish an extramely large value from other normal values. Thus, NA values can be separated from other meaningful data, reducing the risk of bias due to inputations.

A table summarizing the specific imputation approach for each variable is attached at the end of report.

Data Exploration

To better understand the dataset and uncover potential patterns, we conducted exploratory data analysis (EDA) by visualizing feature correlations, distributions, and group comparisons. These analyses provide insights into how different clinical variables relate to clinically-important traumatic brain injury (ciTBI), help us identify potential challenges, and guide the development of our predictive model.

Firstly, the dataset is highly imbalanced, as shown in Table 1 below. Out of 42,412 total cases, ony 376 case (approximately 0.89%) were identified as ciTBI, while the vast majority (approximately 99.1%) did not meet the ciTBI criteria. The significant imbalance presents a challenge for modling, which may requre us make some adjustment to prevent algorithms from favoring the majority class.

	Count	Percentage
No ciTBI (0)	42036.0	99.1134584551542
ciTBI (1)	376.0	0.8865415448457984

Table 1: Imbalance in ciTBI Cases

Given that the response variable is highly imbalanced, we also want to explore if this imbalance is consistent across different age groups. The reason why we are interested in this ciTBI prevalence across age is because one of the key considerations in the PECARN prediction rule is that different algorithms are used for patients under 2 years old and those aged 2 and older. The bart chart in Figure 4 illustrates the distribution of ciTBI across two age groups. While the number of cases in both groups is relatively low, the proportion of ciTBI cases appears similar. However, since there are some differences in clinical features and symptom significance between patients in different age groups, we still consider developing two separate models to capture the difference.

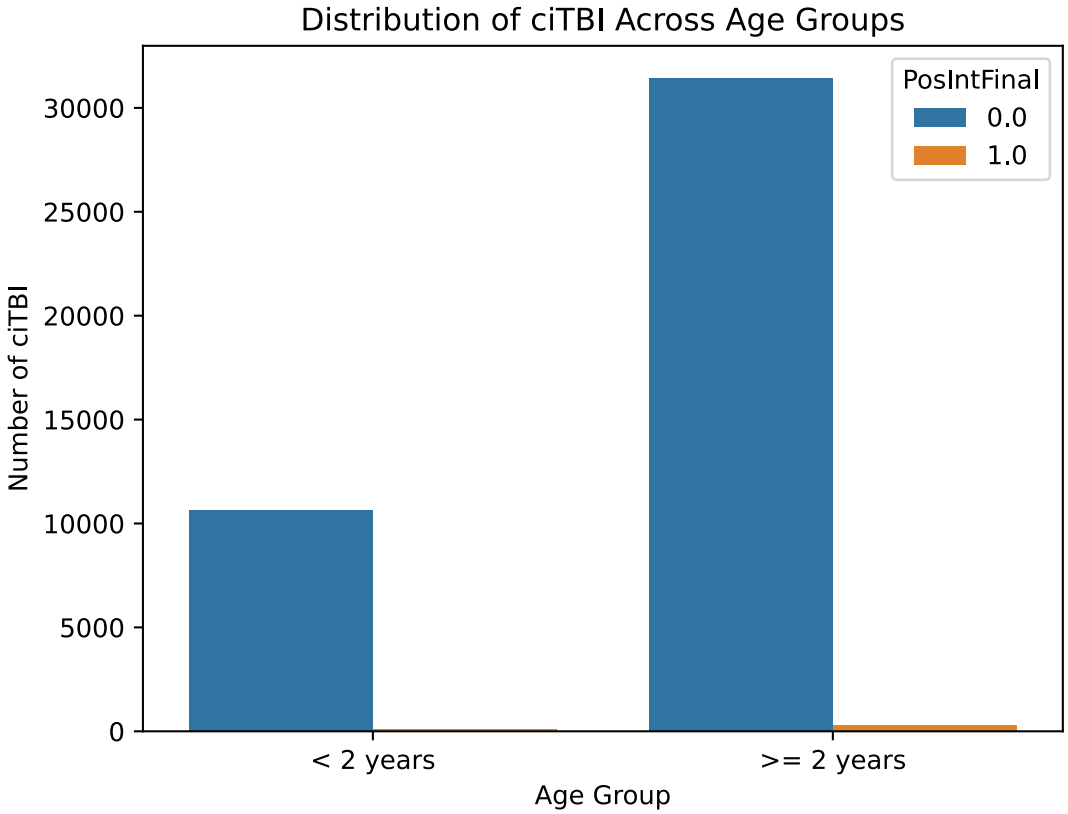

Figure 4: Distribution of ciTBI Across Two Age Groups

To examine how different variables relate to ciTBI, we computed the correlation between individual features and the outcome variable. Figure 5 shows the histogram displaying correlation betwen ciTBI and other factors. We can see that most of fatures have weak correlation with the outcome, with correaltions centered around zero. This may suggest that ciTBI is influenced by a combinatino of multiple clinical features rather than a single strong predictor. This finding indicates the importance of using multivariate models rather than relying on individual clinical indicators.

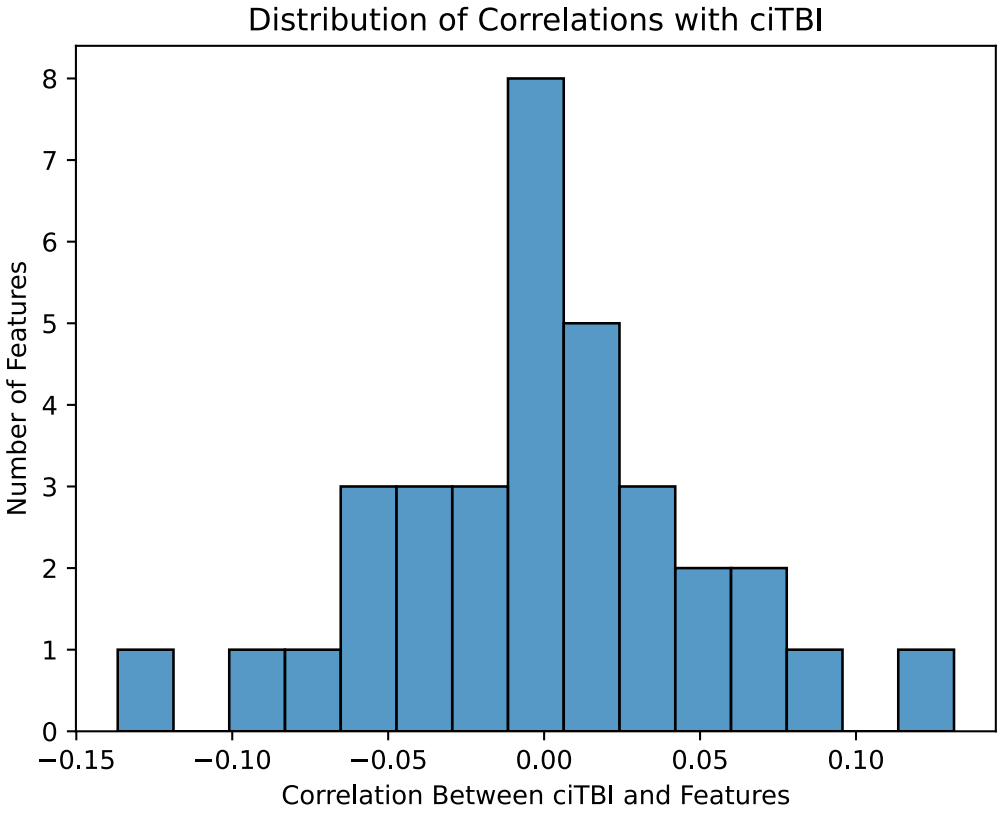

Figure 5: Correlation with ciTBI

Therefore, our main takeaways from data exploration include: the dataset is highly imbalanced, with less than 1% of cases classified as ciTBI. Age-based stratification is necessary for model development. No single feature strongly predicts ciTBI, highlighting the importance of a multivariable predictive model.

Findings

First Finding: Feature Correlations

While most features exhibit weak individual correlations with ciTBI, certain features, including GCS Total Score, Normal Behavior, Basilar Skull Fracture, Neurological Deficits, Palpable Skull Fracture, and Loss of Consciousness, demonstrate notable associations. Additionally, some clinical features show intercorrelations.

The correlation heatmap in Figure 6 provides an overview of how different features are related to each other. Notably, GCS Total Score (`GCSTotal`) shows a negative correlation with ciTBI, with correlation coefficient of -0.52, indicating that patients with lower GCS scores are more likely to have ciTBI. This aligns with clinical expectations, as lower GCS scores suggest greater neurological impairment. Another feature showing a negative correlation with ciTBI is Normal Behavior (`ActNorm`), with coefficient -0.17, meaning that patients who are reported as acting normally are less likely to have ciTBI. This suggests that parental or clinical observations of abnormal behavior may serve as an early warning sign for serious head injuries.

On the other hand, several variables show positive correlations with ciTBI. This means that their presence increases the likelihood of a serious brain injury. These variables include: Basilar Skull Fracture (`SFxBas`), Neurological Deficits (`NeuroD`), Palpable Skull Fracture (`SFxPalp`), and Loss of Consciousness (`LOCSeparate`). The result is reasonable because these features are documented risk factors for serious brain injury and are key factors in the PECARN prediciton rule.

Beyond their relationships with ciTBI, some clinical features also exhibit correlations with each other. For example, Amnesia (Amnesia_verb) and Headache (HA_verb) are positively correlated, suggesting that patients experiencing amnesia are also more likely to report headaches. Similarly, GCS scores are correlated with normal behavior (ActNorm), meaning that patients with lower GCS scores are more likely to exhibit abnormal behavior. Another notable correlation is between GCS and Neurological Deficits (NeuroD), which makes sense as lower neurological deficit is often accompanied by a lower GCS score. These correlations suggest that some features capture overlapping clinical information, which should be considered when building a predictive model.

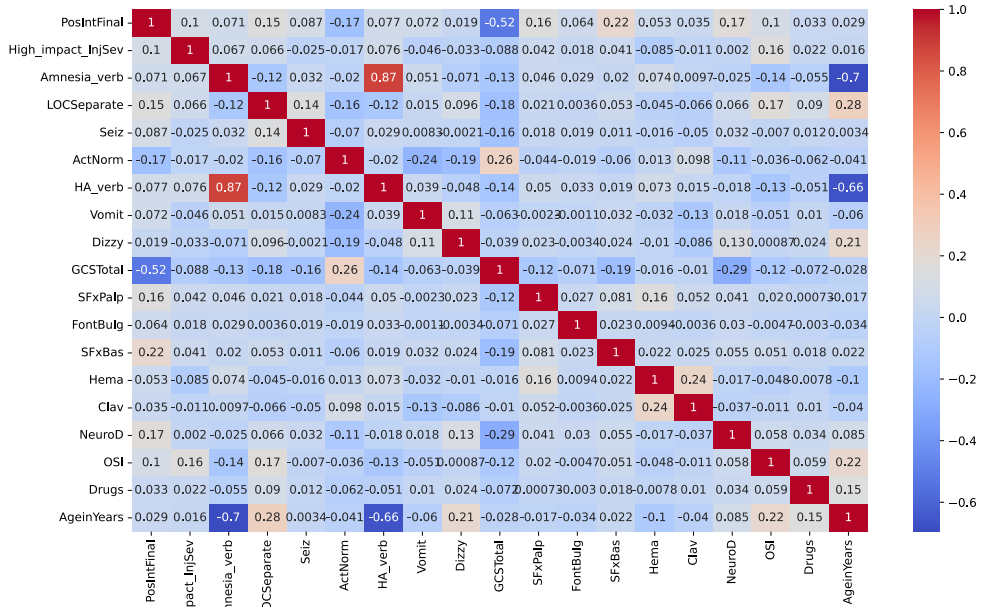

Figure 6: Feature Correlations

Second Finding: Relationship Between GCS and ciTBI

Lower GCS scores are strongly associated with ciTBI, even within the mild TBI population, emphasizing the importance of GCS assessment.

Figure 7 demonstrates the relationship between GCS score and ciTBI risk. A clear trend emerges: while the total number of ciTBI cases is higher among patients with a GCS score of 15, the proportion of ciTBI cases is higher among patients with lower GCS scores. It indicates that patients with lower GCS scores are more likely to have ciTBI. Since the dataset excludes patients with GCS < 14, our analysis confirms that even within mild TBI cases, a small reduction in GCS score can be a meaningful predictor of serious injury. This finding underscores the importance of incorporating GCS into our predictive model for CT scan decision-making.

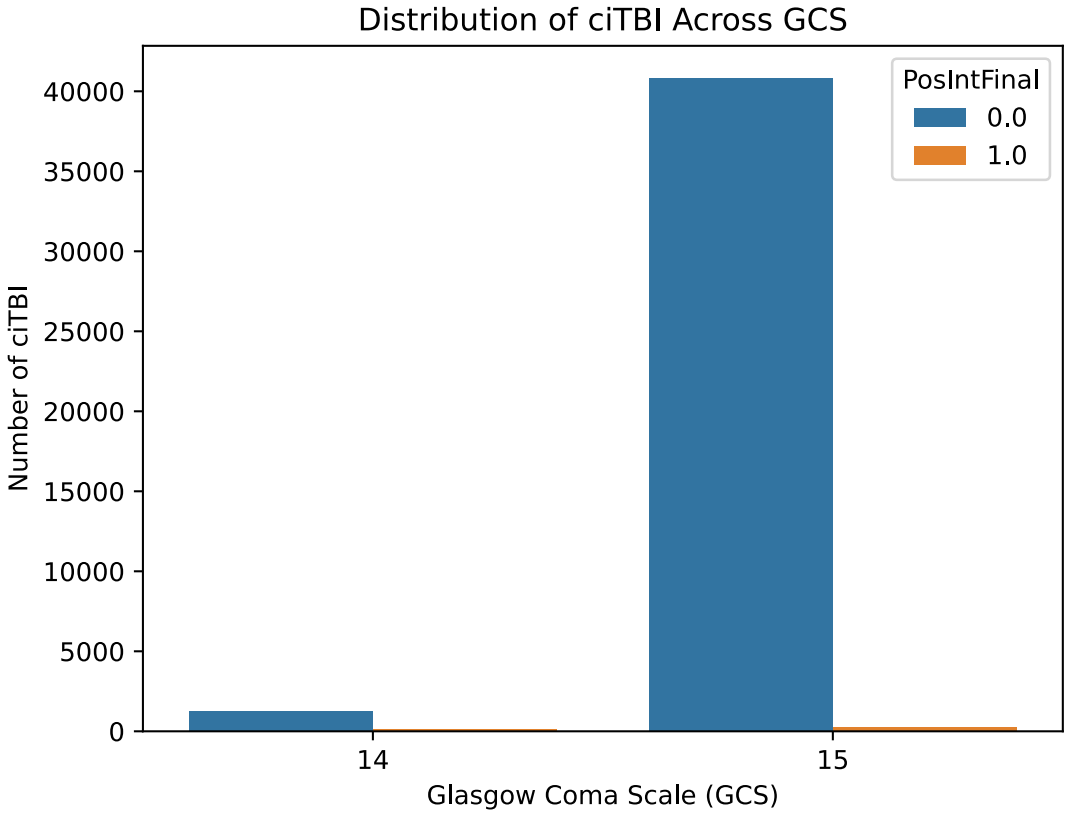

Figure 7: Relationship Between GCS and ciTBI

Third Finding: Injury Mechanism and ciTBI Risk

Falling from an elevation poses a higher risk of ciTBI across both age groups, while others injury mechanisms are more relevant for specific age group.

The relationship between injury mechanism and ciTBI risk reveals notable differences across age groups. Based on Figure 8, in younger patients (<2 years), the most common injury mechanisms include fall from elevation, other mechanism, and occupant in motor vehicle collision, with fall from elevation being the most frequent among ciTBI cases. Similarly, in older patients (2 years), fall from elevation also accounts for the largest number of ciTBI cases, followed by occupant in motor vehicle collision and pedestrian struck by moving vehicle, which are more prevalent in this age group compared to younger patients.

These patterns suggest that certain injury mechanisms pose a higher risk across both age groups, while others may be more relevant for specific age categories. This suggests that injury mechanism should be interpreted differently based on age. It also supports the need

for age-stratified risk assessment, as younger patients experience different injury patterns than older children, which could impact the likelihood of ciTBI.

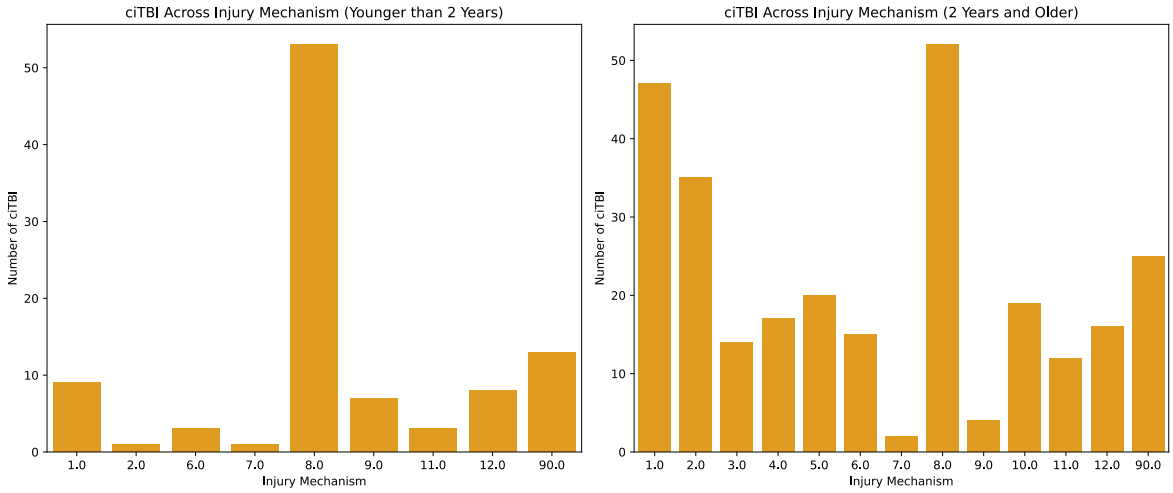

Figure 8: Injury Mechanism and ciTBI Risk

Reality Check

To ensure the validity of our cleaned dataset, we compared it to real-world expectations and domain knowledge from the PECARN study findings. Firstly, our dataset reports the proportion of ciTBI cases among all patients as 0.89%, which is very close to the prevalence rate mentioned in the original study. In addition, our analysis also relfects the relationshipsamong clinical features: patients with lower GCS scores are more likely to have serious brain injury, and ccting abnormally (`ActNorm`) is also negatively correlated with ciTBI, both of which are consistent with existing real-world expectations and domain knowledge.

Injury mechanism distributions also match real-world patterns, with younger patients (<2 years) experiencing more falls from elevation, while older patients (2 years) have higher rates of sports- and vehicle-related injuries compared to younger patients. All these findings suggest that our cleaned dataset remains representative of actual clinical cases and matches with real-world expectations. While minor differences may exist due to missing value imputation, they do not appear to significantly impact overall trends. Therefore, we are confident that our cleaned data passes reality check and represent original dataset appropriately.

Stability Check

To ensure that our findings regarding injury mechanism and ciTBI risk are not overly sensitive to small variations in the data, we conducted a stability check by perturbing the dataset.

Specifically, we randomly sampled 50% of data in both younger (<2 years) and older (2 years) patients and compared the results before and after sampling.

The result in Figure 9 shows that the overall pattern remains consistent despite the random sampling. Injury mechanisms differ across age groups, but fall from elevation is the mechanism with the highst ciTBI prevalence rate.

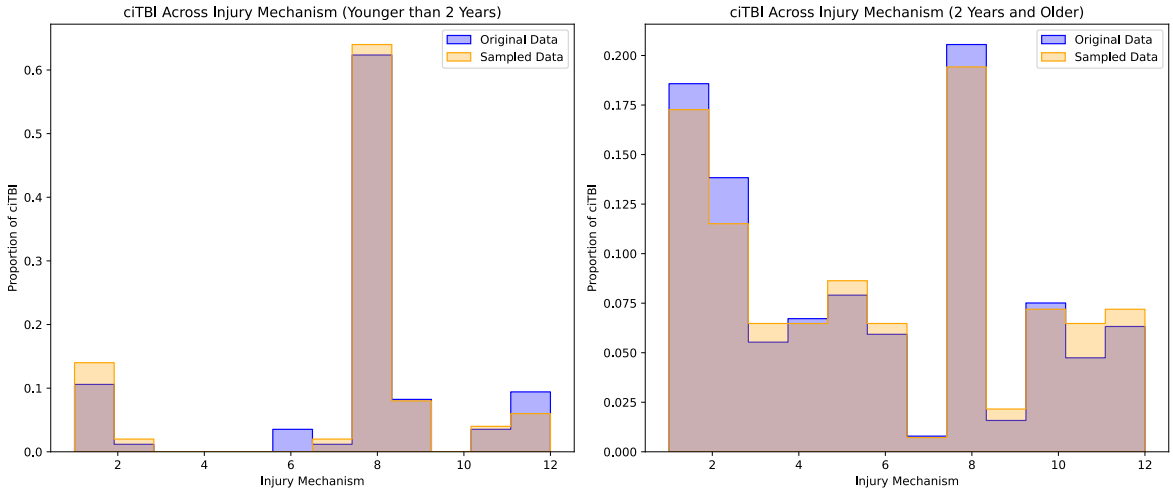

Figure 9: Injury Mechanism and ciTBI Risk Before and After Sampling

Modeling

Implementation

To develop a predictive model for ciTBI, we trained two separate decision tree models—one for younger patients (<2 years) and another for older patients (2 years). This aligns with the age-stratified approach we discussed before. For each age group, we split the dataset into 80% training and 20% testing sets, ensuring that the model was trained on a sufficient number of samples while keeping a testing set for evaluation. We selected Simple Decision Tree as our modeling approach because it has some important benefits: It is interpretable, allowing clinicians to understand how decisions are made; it handls non-linear relationships well, making them effective for clinical data; it naturally performs feature selection, focusing on the most important predictors.

To account for class imbalance (since ciTBI cases make up less than 1% of the dataset), we adjusted the class weight parameter so that the model did not over-prioritize the majority class. This ensures that ciTBI cases are given more weight during training, improving sensitivity to rare but critical cases. To avoid overfitting, we also made adjustments on hyperparameter including maximum depth, minimum sample leaf, and minimm sample split.

- Class Weight = 1 for non-ciTBI and 80 for ciTBI
- Maximum depth = 5
- Minimum sample leaf = 500 for younger and 800 for older
- Minimm sample split = 800 for younger and 3000 for older

Aftertraining, the models achieved accuracy of 0.81 for both younger patients and older patients.

Interpretability

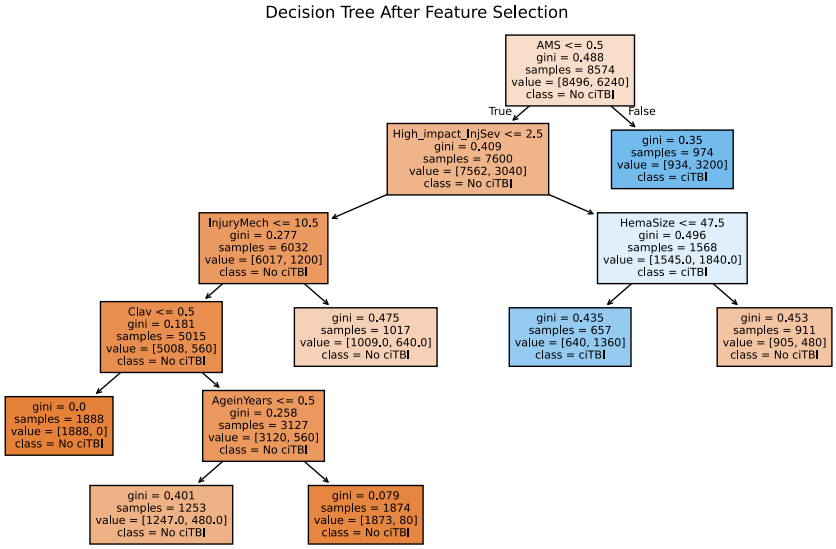

Figure 10: Tree Model for Younger Patients

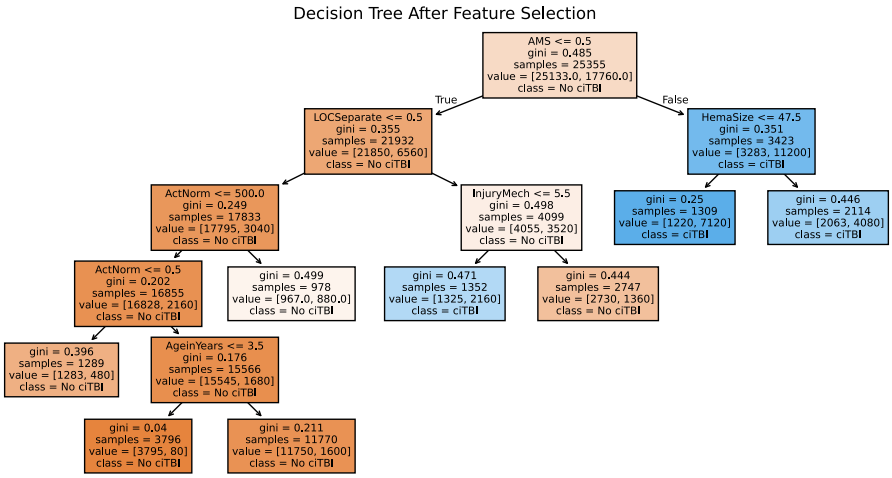

Figure 11: Tree Model for Older Patients

Discussion

Despite the strengths of our analysis, several limitations should be acknowledged. First, the dataset is highly imbalanced, with ciTBI cases comprising less than 1% of total observations, which may impact model performance despite using class weighting.

Second, missing data handling relied on assumptions, such as imputing values based on age or related features, which may introduce bias.

Conclusion

In this study, we explored the prediction of clinically-important traumatic brain injuries (ciTBI) in patients using a dataset from the PECARN study. After thorough data preprocessing, including handling missing values, removing redundant features, and addressing class imbalance, we developed two separate decision tree models, one for younger patients (<2 years) and another for older patients (2 years), to reflect differences in injury mechanisms and clinical presentation. Both models achieved accuracy scores of 0.81, demonstrating reasonable predictive capability while maintaining interpretability for clinical use. Through exploratory analysis, we found that GCS scores, abnormal behavior, and skull fractures are strong indicators of ciTBI, while injury mechanisms differ between age groups but consistently show that fall from elevation is the most frequent among ciTBI cases. A stability check confirmed that

these patterns hold even with perturbations in the data. This analysis can assist in reducing unnecessary CT scans while ensuring patient safety.

Academic Honesty Statement

I affirm that the work in this report is my own and that all sources, including original article and contributions from classmates, are properly cited.

Collaborators

I discussed how to deal with missing values with Junya Tsuneishi. I wrote up the report completely on my own.

Bibliography

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, 2009, pp. 1160-1170. https://doi.org/10.1016/S0140-6736(09)61558-0.