Manuscript/Abbreviated Title: Spinal cord injury AIS predictions using machine learning

https://doi.org/10.1523/ENEURO.0149-22.2022

Cite as: eNeuro 2022; 10.1523/ENEURO.0149-22.2022

Received: 10 April 2022
Revised: 21 October 2022
Accepted: 18 November 2022

This Early Release article has been peer-reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.eneuro.org/alerts to receive customized email alerts when the fully formatted version of this article is published.
Manuscript/Abbreviated Title: Spinal cord injury AIS predictions using machine learning

Authors and Affiliations:
Dhruv Kapoor
Georgia Institute of Technology, College of Computing

Clark Xu
Georgia Institute of Technology, College of Computing
University of Wisconsin – Madison School of Medicine and Public Health, Department of Medicine

Author Contributions: All authors contributed to research design/performance/analysis and paper writing.

Correspondance: To be addressed to Clark Xu at 608-890-5508 and/or cxu@medicine.wisc.edu.

Figure Count: 2
Table Count: 4
Multimedia Count: 0
Abstract Words: 222
Significance Statement Words: 36
Introduction Words: 604
Discussion Words: 759

Acknowledgements: N/A
Conflict of Interest: Authors report no conflict of interest.
Funding Sources: N/A

The authors wish to be considered as joint first-authors.
Spinal cord injury AIS predictions using machine learning

Abstract

Objective: To use machine learning to predict AIS scores for newly injured SCI patients at hospital discharge time from hospital admission data. Additionally, to analyze the best model for feature importance in order to validate the criticality of AIS score and highlight relevant demographic details.

Design: Data used for training machine learning models was from the NSCISC database of United States SCI patient details. 18 real features were used from 417 provided ones, which mapped to 53 machine learning features after processing. 8 models were tuned on the dataset to predict AIS scores and Shapely analysis was performed to extract the most important of the 53 features.

Participants: Patients within the NSCISC database who sustained injuries between 1972 and 2016 after data cleaning (n = 20,790).

Outcome Measures: Test set multi-class and aggregated Shapely score magnitudes.

Results: Ridge Classifier was the best performer with 73.6% test set accuracy. AIS scores and neurologic category at admission time were the best predictors of recovery. Demographically, features were less important but age, sex, marital status, and race stood out. AIS scores on admission are highly predictive of patient outcomes when combined with patient demographic data.

Conclusion: Promising results in terms of predicting recovery were seen and Shapely analysis allowed for the machine learning model to be probed as whole, giving insight into overall feature trends.

Significance

The research is intended to introduce the use of machine learning to enhance predictive capabilities of spinal cord injury recovery, to validate previous motor-sensory classification work, and to extract important deciders of recovery from constructed models.
Introduction

Spinal cord injury (SCI) profoundly changes a patient’s life. Effects range from impaired motor function, up to and including paralysis of the limbs, as well as mental health effects such as depression or suicide. Patient outcomes are highly sensitive to where and to what extent the injury is on the spinal cord. In general, an injury closer to the brain stem has a greater impact. Among the impaired motor functions are: paralysis, loss of sensation, increased chance of developing pressure ulcers, bladder dysfunction, neurogenic bowel, muscle atrophy, autonomic dysreflexia, and impaired sexual function (Sezer et al., 2015).

The American Spinal Injury Association Impairment Scale (AIS or ASIA Scale) classifies the motor-sensory abilities of a patient with a SCI (Ho et al., 2007). There are five letter-grade categories: AIS A is a complete injury with no retention of motor control or sensory function below the point of injury, and AIS E is an injury with minimal impact on the patient. Clinicians use AIS to classify SCI and quantify SCI recovery, for example, an improvement from B to C, or a deterioration from D to C.

One key question among SCI patients is how the severity of their injury, as measured by the AIS, will improve or deteriorate during the course of SCI recovery. This varies by the type of injury and demographic differences between patients.

In a traditional clinical setting, the main factors identified for prognostication of SCI recovery include patient age, patient gender, length of inpatient stay, type of inpatient discharge, type of SCI, time to procedure, procedure type, and comorbidities (Chay and Kirshblum, 2020). SCI prognosis is primarily carried out by either standard of care diagnostics including heuristic bedside evaluation and magnetic-resonance imaging, or traditional clinical analysis such as odds-ratio statistic (Burns et al., 2012).
Thus, there is an opportunity to support SCI recovery by adding a machine learning based framework of SCI prognosis using big data and precision medicine as one of the clinician’s tools for improving SCI patient outlook.

Researchers have developed many tools on SCI treatment; however, after literature review, it was seen that there is a wide gap in the use of machine learning algorithms to predict SCI recovery in a contemporary precision medicine context, especially with regard to feature importance and using a very large dataset (Snoek et al. 2004; Munce et al., 2014). One paper attempted to predict discharge location using an ensemble model and used Area Under the Curve (AUC) as an outcome (Fan et al., 2021), another made use of Convolutional Neural Nets (CNNs) on Magnetic Resonance Imaging (MRI) charts to achieve an accuracy of 71.4% (Okimatsu et al., 2022), while two other studies greatly limited the complexity to specific AIS scores of A (Buri et al., 2022) and D/E (Inoue et al., 2020). The authors in (Chou et al., 2022) carried out a similar study to the one presented here but their sample size consisted of 74 patients.

The research carried forward was based on the National Spinal Cord Injury Statistical Center (NSCISC) database which includes patient details across the United States (Chen, n.d.). Several different machine learning models were employed to predict AIS level upon patient discharge for data recorded between 1972-2016 and the best model was further examined to extract feature importance information. The ground truth AIS scores at discharge were supplied as part of the dataset.

The analysis of feature importance serves two purposes. One is to verify the importance of AIS classification at admission time as a critical feature, using a data-driven approach. And the other is to identify demographic features which also play a crucial role in determining recovery.
Methods

Computational implementation was carried out using Python v3.8, shap v0.40.0, and scikit-learn v1.0.1.

Data Preparation

The NSCISC database comprises more than 29,000 traumatic SCIs since 1973 for patients treated at any regional model SCI system within the first year of injury and who have signed a consent form for inclusion (DeVivo et al., 2002). The patient details within the database have also been stripped of all HIPAA-defined identifiers. The NSCISC dataset was loaded in from the published CSV format, and it had 417 raw features. A custom data mapper was used to translate the raw data headers and values into more recognizable features. For example, the label AWghtRhb was translated to Weight at Admission. Any non-recorded values in the dataset were assigned a value of Unknown.

There were far more features in the dataset than are relevant to the machine learning models designed, so the data used for training was limited to patient information which is known at or before hospital admission. Exploratory data analysis showed that some features had over 90% missingness and these were excluded from consideration as model inputs. The other reasons for excluding certain variables within the dataset was if they were specific to a certain area of the body or spine such as sensory level of the left side during admission. Including all of these would’ve greatly increased the sparsity of the eventual feature vector, leading to risks involved with the curse of dimensionality. Univariate analysis was also performed on possible features to look at max/min values, counts, and outliers. Table 1 shows the final features chosen as well as their mapped versions for input into machine learning model construction, along
with imputations performed for missing values.

Generally, imputation followed the format of using the mode as the chosen mapped value or an Unknown label was assigned instead if it was found in a feature’s set of values already. The choice to use mode over creating a new Unknown label when not already available was decided because of two primary reasons. The first was to avoid creating a value that very few rows have, which could have the side effect of incorrectly flagging these patients as disproportionately important during training time. The second was to avoid inflating feature dimensionality for features such as education level at injury because the introduction of an Unknown value would require a transition from an ordinal feature to a one-hot encoded one. Sex and age at injury were the only features that were fully populated whereas AIS score and level of injury at admission were the only features where a missing value resulted in a dropped row.

From analysis, the typical profile of a SCI patient was found to be a male, aged between 19 and 29 years old, white, and never married, although the dataset showed plenty of variation from this modal profile.

Of the dataset features, three variables were determined as the most suitable for measuring SCI recovery through the patient’s course of treatment: the AIS score, neurological disposition at discharge, and patient’s level of injury. These three features were suitable because they could capture patient physical improvements throughout the entire body. The AIS score at discharge was ultimately chosen due to its widespread use in literature (Roberts et al., 2017; Inoue et al., 2020; Buri et al., 2022; Chou et al., 2022; Okimatsu et al., 2022). Using this target variable meant that a 5-class classification modeling approach was to be designed, where each class is one of A, B, C, D, or E. With this definition, there was a possibility that some patient predictions could include worsening of AIS score as well. In the dataset, only 329 of these cases were found in
total, 275 in the eventual training set and 54 in the eventual testing set, for a combined ~1.6% of all samples. As a result, the vast majority of predictions were focused on recovery.

The finalized dataset after analysis that was used for model training is described in Table 2. 417 raw NSCISC features were reduced to 18, and later mapped to 53 model-ready features. The train/test split of ~90:10 was reached after trialling different ratios and evaluating test accuracy.

**Modeling and feature importance**

There were 8 different machine learning models tuned after the dataset was prepared and test set prediction accuracy was the decider in determining the best model. The process of data preparation through to model selection is outlined in Figure 1.

To extract feature importance from the top model, Shapely values were used (Parsa et al., 2020). These provide a quantitative measurement of how strong a certain feature is in predicting a specific class output. The magnitude of the Shapely value was used so feature strength for or against a specific class is captured. These values were computed per sample and taking the mean over all samples can provide an average importance. The strength of each of the 53 features with respect to a given class is thus given by Equation 1:

\[
\forall f, c: S_{f,c} = \frac{\sum_{i=0}^{N} |\text{shap}_{f,c,i}|}{N}
\]  

(1)

Where \( f, c, \) and \( N \) correspond to feature, class, and number of samples respectively.

Feature strength can be summed across all classes to give an overall importance metric with respect to the model as a whole, as seen in Equation 2:

\[
\forall f: S_{f,m} = \sum_{i=0}^{C} S_{f,i}
\]  

(2)
Where \( m \) and \( C \) are the model and the set of all classes respectively.

**Results**

**Machine Learning**

The results after training the data on all tuned models are described in Table 3. With a multi-class test accuracy of 73.6\%, it was found from the machine learning model metrics that the best performing model was Ridge Classifier over the NSCISC dataset for SCI recovery prognostication. SVM, Elastic Net, and Logistic Regression closely followed with 73.5\%, 73.2\%, and 73.2\% respectively. These results are very promising, given that they are only for information discovered or provided on an initial assessment.

Taking the Ridge Classifier and applying it to the dataset once again, but with removing the 329 patients who scored lower AIS scores at discharge than at admission, gave a higher multi-class test accuracy of 75.3\%.

To determine what inputs are most crucial to predicting recovery, Shapely values were applied as per Equation 1 and Equation 2 over the Ridge Classifier model. The top 20 most important features are visualized in Figure 2 and recorded in Table 4. It has been previously shown that the best indicator of AIS score at discharge is generally AIS score at admission and neurologic category at admission (Chay and Kirshblum, 2020) and the results confirm this.

In terms of demographic details playing a role, these were less impactful. From previous research, it was expected that age would be among the most important (Seel et al., 2001; Wilson et al., 2012, 2014). While it was, there were a number of other features, such as sex and race, which were also around the same level of absolute mean importance. Interestingly, age was disproportionately a higher importance for AIS C.
predictions. And, unexpectedly, marital status showed to be even more important than age.

**Discussion**

The results showed promising results in predicting AIS improvement. A 73.6% test accuracy can be considered a benchmark for improvement. Given the high amount of data samples, there is also the option of more complex models to trial, such as deep neural networks. The use of more engineered features can also be used as a tool to add more insight while reducing dimensionality. For example, height and weight features can instead be replaced with a body mass index measure. Finally, while the SCI recovery predictions is the result of one machine learning model in the pipeline, additional improvements in model performance could be gained by the creation of a model of models in which multiple sub-models optimize for predicting features of importance to SCI recovery, which then feed into an overall model for prognosticating the patient.

One attempt made to increase performance was by dropping the 1.6% of patients who deteriorated in AIS scores. The boost ended up being 1.7% to test accuracy to put it at 75.3%. This was likely because cases where patient scores deteriorated were difficult for the model to appropriately fit. There is opportunity to use this as a secondary model if patients have shown signs of AIS recovery prior to hospital discharge. Otherwise, the original model is more appropriate since it makes no assumptions about progression. Nonetheless, the amount of increase to test accuracy exceeded expectations, considering that this metric’s difference between Ridge Classifier and the next three best performant models on the original dataset was, at most, 0.4%.
Feature strength gave a better understanding of which areas most affected recovery, both when it came to validating the importance of admission AIS score, and with regard to understanding the role that demographics play. The surprising result of marital status exceeding the importance of age was an important outcome of the study. The presence of a support system for a patient may be a critical component of recovery success, though this would need to be examined more closely in future work. However, an important note for the Shapely analysis is to remember that the values computed from real data on SCI recovery, so there may be socio-economic factors at play which act as social determinants of health among disadvantaged and underserved groups. As a result, whether these demographic details are highlighted as strong or weak features may actually be partially or completely due to societal complexities.

Comparing the results with those found in (Okimatsu et al., 2022), the test accuracy here of 73.6% is an improvement over the MRI CNN accuracy of 71.4%. Furthermore, the Ridge Classifier is much easier to interpret and is a more time efficient model to train. The inclusion of MRI imagery as a set of features is a possible route of future research which could further bolster the Ridge Classifier as well. In contrast to the research performed by the authors in (Inoue et al., 2020; Fan et al., 2021; Buri et al., 2022; Chou et al., 2022; Okimatsu et al., 2022) as a whole, the study carried out here uses a patient base 1-2x larger, while including a comprehensive review of feature importance as well. To add on, the results are overall comparable or better while considering all AIS classes and using a very lightweight model which can be much more easily deployed.

To extend the machine learning research into SCI recovery outlined in this paper, the codebase found at https://github.com/kapoor1992/spinal_cord_injury_recovery can be augmented by
including more models or altering input features before metrics are recomputed.

**Limitations**

There are a few limitations with the dataset that can be a point of future research. Firstly, the features did not look at information past admission time. The inclusion of new variables during inpatient stay may be important to evaluate how much these initial features vary in importance. For example, the amount of weekly physical therapy, the amount of counselling patients receive, and others could drastically alter results. The current scope only captured a snapshot of prognosis at initial intake and evaluation of the newly injured patient. Also, in terms of model performance and tuning as more data is extracted from the NSCISC dataset, there is a growing opportunity to measure accuracy or other metrics in a longitudinal study. This can help identify the limitation that the underlying NSCISC data is a static data source, which does not have the capability for either batch or streaming updates to the data. Therefore, the quality of the machine learning model predictions may decay over time as the SCI patient population experiences underlying demographic shifts.

**References**


Chen Y (n.d.) National spinal cord injury model systems database.


Tables

Table 1. Feature mappings from those in the original dataset to transformed versions used at model training time, along with the imputation techniques used.

<table>
<thead>
<tr>
<th>Original Feature</th>
<th>Machine Learning Features</th>
<th>Missing Value Imputation Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occupation Status – Injury</td>
<td>Homemaker, In Training, In Workshop, Other, Retired, Student or Infant, Unemployed, Unknown, Working</td>
<td>Mapped to unknown</td>
</tr>
<tr>
<td>Diabetes – History</td>
<td>Same as original feature</td>
<td>Mode (no history)</td>
</tr>
<tr>
<td>Veteran</td>
<td>Same as original feature</td>
<td>Mode (not a veteran)</td>
</tr>
<tr>
<td>Race</td>
<td>Asian, Black, Multiracial, Native American, Unknown, White</td>
<td>Mapped to unknown</td>
</tr>
<tr>
<td>ASIA – Admission</td>
<td>A, B, C, D</td>
<td>None; rows dropped</td>
</tr>
<tr>
<td>Sex</td>
<td>Same as original feature</td>
<td>None; all values were populated</td>
</tr>
<tr>
<td>Education – Injury</td>
<td>Same as original feature</td>
<td>Mode (high school)</td>
</tr>
<tr>
<td>Depression – History</td>
<td>Same as original feature</td>
<td>Mode (no history)</td>
</tr>
<tr>
<td>TBI Likelihood – Injury</td>
<td>Same as original feature</td>
<td>Mode (improbable)</td>
</tr>
<tr>
<td>Level of Injury – Admission</td>
<td>Same as original feature</td>
<td>None; rows dropped</td>
</tr>
<tr>
<td>Daily Alcohol – History</td>
<td>Same as original feature</td>
<td>Mode (zero)</td>
</tr>
<tr>
<td>Anxiety – History</td>
<td>False, General Anxiety, Multiple, PTSD, Panic Disorder, Unknown</td>
<td>Mapped to unknown</td>
</tr>
<tr>
<td>Description</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>------------------------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td>Total Samples</td>
<td>20,790</td>
<td></td>
</tr>
<tr>
<td>Training Samples</td>
<td>18,737</td>
<td></td>
</tr>
<tr>
<td>Training Injury Dates</td>
<td>1972-2005</td>
<td></td>
</tr>
<tr>
<td>Testing Samples</td>
<td>2,053</td>
<td></td>
</tr>
<tr>
<td>Testing Injury Dates</td>
<td>2006-2016</td>
<td></td>
</tr>
<tr>
<td>Original Features</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Machine Learning Features</td>
<td>53</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Final dataset at a glance that was used for machine learning.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ridge Classifier</td>
<td>0.824</td>
<td>0.736</td>
</tr>
<tr>
<td>SVM</td>
<td>0.825</td>
<td>0.735</td>
</tr>
<tr>
<td>Elastic Net</td>
<td>0.823</td>
<td>0.732</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.824</td>
<td>0.732</td>
</tr>
<tr>
<td>Ensemble (Elastic Net, KNN, Random Forest)</td>
<td>0.860</td>
<td>0.717</td>
</tr>
<tr>
<td>KNN</td>
<td>0.840</td>
<td>0.711</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.927</td>
<td>0.693</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.591</td>
<td>0.429</td>
</tr>
</tbody>
</table>

Table 3. Accuracy results from machine learning model runs, ordered in descending order by test accuracy.
Table 4. Top 20 most important features for Ridge Classifier.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Machine Learning Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ASIA - Admission_A</td>
</tr>
<tr>
<td>2</td>
<td>ASIA - Admission_D</td>
</tr>
<tr>
<td>3</td>
<td>ASIA - Admission_B</td>
</tr>
<tr>
<td>4</td>
<td>Neurologic Category - Admission_Complete Paraplegic</td>
</tr>
<tr>
<td>5</td>
<td>Neurologic Category - Admission_Incomplete Tetraplegic</td>
</tr>
<tr>
<td>6</td>
<td>ASIA - Admission_C</td>
</tr>
<tr>
<td>7</td>
<td>Neurologic Category - Admission_Complete Tetraplegic</td>
</tr>
<tr>
<td>8</td>
<td>Neurologic Category - Admission_Incomplete Paraplegic</td>
</tr>
<tr>
<td>9</td>
<td>Level of Injury – Admission</td>
</tr>
<tr>
<td>10</td>
<td>Marital Status - Injury_Never Married</td>
</tr>
<tr>
<td>11</td>
<td>Marital Status - Injury_Married</td>
</tr>
<tr>
<td>12</td>
<td>Occupation Status - Injury_Working</td>
</tr>
<tr>
<td>13</td>
<td>Primary Insurance_Uknown</td>
</tr>
<tr>
<td>14</td>
<td>Occupation Status - Injury_Student or Infant</td>
</tr>
<tr>
<td>15</td>
<td>Age – Injury</td>
</tr>
<tr>
<td>16</td>
<td>Marital Status - Injury_Divorced</td>
</tr>
<tr>
<td>17</td>
<td>Race_White</td>
</tr>
<tr>
<td>18</td>
<td>Sex</td>
</tr>
<tr>
<td>19</td>
<td>Occupation Status - Injury_Retired</td>
</tr>
<tr>
<td>20</td>
<td>Education – Injury</td>
</tr>
</tbody>
</table>

Figures

Figure 1. Data preparation flowchart through to model selection (Full image at: https://github.com/kapoor1992/spinal_cord_injury_recovery/blob/release/submission/ml/modelling/plots/flowchart.png).
Figure 2. Top 20 most important features. (Full image at: https://github.com/kapoor1992/spinal_cord_injury_recovery/blob/release/submission/ml/modelling/plots/importance.png).

Extended Data 1: Code file.