Predicting Post-Stroke Cognitive Impairment (PSCI) Using Multiple Machine Learning Approaches

Presented at the International Stroke Conference (ISC), 2025

Yuzhang Xie
Emory University
Atlanta, GA, USA
yuzhang.xie@emory.edu

Fadi B. Nahab
Emory University
Atlanta, GA, USA
fadi.nahab@emory.edu

Yi Ge
University of California, Berkeley
Berkeley, CA, USA
billy.ge@berkeley.edu

Yuhua Wu
Emory University
Atlanta, GA, USA
gloria.wu@emory.edu

Jessica Saurman
Emory University
Atlanta, GA, USA
jessica.l.saurman@emory.edu

Carl Yang
Emory University
Atlanta, GA, USA
j.carlyang@emory.edu

Xiao Hu
Emory University
Atlanta, GA, USA
xiao.hu@emory.edu

I. BACKGROUND

Post-stroke cognitive impairment (PSCI) is a condition characterized by cognitive decline that occurs after a stroke. PSCI affects up to 60% of stroke survivors. Early detection of those at high risk for PSCI is essential for timely intervention and personalized care. Electronic health records (EHRs) contain valuable data that can be leveraged by machine learning to predict PSCI, potentially enhancing patient outcomes. This study focuses on developing and validating machine learning models to predict PSCI, aiming to enable earlier diagnosis and improve post-stroke care.

II. METHODS

7956 all-type stroke patients (including Ischemic & Hemorrhagic stroke) treated between 2012 and 2021 were extracted from Emory Healthcare system. We employed multiple methods to predict PSCI, using ICD codes and prescribed medications that were available up to the discharge of index strokes. First, we utilized traditional machine learning methods, including Logistic Regression, Support Vector Machine, and Random Forest to develop models. Then, we developed hypergraph models to enhance prediction performance. Unlike traditional graphs that only capture pair-wise relationships between pairs of entities, hypergraphs can model the more complex higher-order relationships among multiple entities, by allowing a hyperedge (encounter) to connect multiple vertices (ICD and medications) simultaneously among patient visits and EHR medical features. Finally, we compared the performance across different methods and selected the best one for the PSCI prediction task. We compared their performance on four metrics: ACC (Accuracy, the proportion of correct predictions), AUC (Area Under the ROC Curve, measuring the model's ability to distinguish between classes), AUPR (Area Under the Precision-Recall Curve, a comprehensive measure considering both precision and recall), and Macro-F1 (a balanced measure calculated by the harmonic mean of precision and recall).

III. RESULTS

We included 7956 all-type stroke patients (50% female, 56% non-white) in this analysis, where 1797 (23%) had diagnostic codes often used by clinicians at Emory to document PSCI (with demographic statistics displayed in Table I). According to the performance in Table II, the hypergraph model was associated with higher ACC, AUC, AUPR, and Macro-F1 than other models.

TABLE I
DEMOGRAPHIC STATISTICS OF THE DATASET

	PSCI Patients	Non-PSCI Patients	Total
Count	1797 (22.6%)	6159 (77.4%)	7956
Female	919 (51.1%)	3033 (49.2%)	3952 (49.7%)
Hispanic or Latino	36 (2.0%)	208 (3.4%)	244 (3.1%)
White	766 (42.6%)	2765 (44.9%)	3531 (44.4%)
Black or African American	931 (51.8%)	2638 (42.8%)	3569 (44.9%)
Native Hawaiian or Other Pacific Islander	1 (0.1%)	15 (0.2%)	16 (0.2%)
American Indian or Alaskan Native	7 (0.4%)	17 (0.3%)	24 (0.3%)
Asian	45 (2.5%)	227 (3.7%)	272 (3.4%)
More than One Race	9 (0.5%)	42 (0.7%)	51 (0.6%)

TABLE II

PERFORMANCE OF MACHINE LEARNING METHODS TO PREDICT POST-STROKE COGNITIVE IMPAIRMENT

Method	ACC	AUC	AUPR	MACRO-F1
Hypergraph (HypEHR)	0.7670	0.6558	0.3493	0.5171
Logistic Regression	0.7720	0.6298	0.3094	0.4553
Support Vector Machine	0.7771	0.5868	0.2933	0.4373
Multi-layer Perceptron	0.6544	0.5701	0.2625	0.5298
Random Forest	0.7771	0.6105	0.2984	0.4373
Naive Bayes	0.5173	0.6036	0.2931	0.5235
XGBoost	0.7675	0.6040	0.2973	0.4626

IV. CONCLUSION

By comparing the results of various machine learning methods, we found that hypergraph model approaches outperform traditional machine learning methods in utilizing EHRs for predicting PSCI after a stroke.