# New Methods of Natural Language Processing using Machine Learning Methods to Identify Ischemic Stroke Presence, Acuity and Location from Clinical Radiology Reports

*Charlene Jennifer Ong MD, MPHS[1,2,3] Rebecca Zhang[3], Agni Orfanoudaki[3], Francois Pierre M Caprasse[3], Meghan Hutch[1], Liang Ma[1,] Darian Fard[1], Margaret Minnig[1], Stelios Smirnakis MD, PhD[2], Dimitris Bertsimas PhD[3]*

*[1] Boston University School of Medicine, [2] Harvard Medical School, [3] Massachusetts Institute of Technology Operations Research Center*

**Background.**

Unstructured text in the form of radiology reports or patient-notes contains some of the most useful real-time and patient-specific information to practicing clinicians, but can be difficult to access and organize in a retrospective and scaled fashion. This often results in studies that must either eschew the wealth of information contained in these reports for analyses, or institute a labor-intensive and manual "hand-labeling" of pertinent features that substantially reduces sample size. A substantial proportion of ICD-9/10 codes misclassify patients with ischemic stroke events. Moreover, they do not accurately distinguish acuity or location. The ability to extract this information quickly and accurately would provide a considerable improvement over traditional methods of identifying stroke retrospectively in large data-sets. An algorithm that correctly identifies diagnoses would also have substantial value in helping to triage critical reports in the clinical setting.

Previous efforts at using machine learning to classify radiologic text have included diagnosis of pneumonia, breast cancer, stroke and critical findings on Head CTs. However none of these methods went so far to specify acuity and location of ischemia. In this study, we developed a comprehensive framework that leverages the latest advanced Machine Learning (ML) methods to classify radiology reports to determine 1) presence 2) acuity and 3) location of ischemic stroke. We report the preliminary findings of our study for presence of ischemic stroke.

**Methods.**

We collected radiology reports from a cohort of patients with an ICD9/ICD10 labeled diagnosis code of ischemic stroke from 2003-2018. Patient data was collected from the Research Patient Data Registry, a clinical repository of patient information from Massachusetts General and Brigham and Women's Hospitals. Additional eligibility criteria for study inclusion consisted of full reports of Head Computed Tomography (CT) or CT Angiography studies, Brain Magnetic Resonance Imaging or Angiography studies of patients over 18 years of age. 1,357 original reports were collected and hand-labeled by two clinicians to determine presence, acuity and location of ischemic stroke.

We leveraged standard Natural Language Processing (NLP) techniques to create Bag-of-Words (BOW) and TF-IDF vector representations of unstructured text. In addition, we compiled a comprehensive set of general and content specific documents including online encyclopedias, neurologic textbooks, disease specific publications and radiographic reports to train context-specific GloVe embeddings. These word vectors will be made publicly available for use in other related research problems. We trained a wide range of both interpretable and uninterpretable ML methods for binary classification, including Logistic Regression, k-Nearest Neighbors, Optimal Classification Trees, Random Forest and Recurrent Neural Networks.

**Results.**

Out of 1357 hand-labeled reports, we determined that only 925 (68%) reported ischemic stroke. Our experiments demonstrated that BOW performs best when combined with interpretable classifiers such as Logistic Regression, being able to identify the presence of ischemic stroke with an Area Under the Curve (AUC) of 0.951. We found that less interpretable NLP techniques such as GloVe provide best results when combined with more opaque deep learning techniques (AUC=0.976). The cost of interpretability is not high in this setting, as the latter is associated with only a 2.1% improvement.

**Conclusion.**

Our study provides a comprehensive assessment of NLP methods for binary classification in the medical setting. By comparing a variety of interpretable and "black-box" methods, we demonstrate that extracting binary information from unstructured text can be adequately accomplished using more understandable techniques. Further study is needed to determine whether these techniques accurately classify more complex information including acuity, size or severity. The high sensitivity and specificity of our models in identifying stroke is consistent with other work in the literature and is a model for an improved approach of identifying stroke from large data cohorts for both clinical and research use.