

Construction of a Machine Learning Dataset through Collaboration

An unprecedented collaboration among two medical societies and over 60 volunteer neuroradiologists has resulted in the generation of the largest public collection of expert-annotated brain hemorrhage CT images [1]. Leaders of the project expect the dataset to help speed the development of machine learning (ML) algorithms to aid in the detection and characterization of this potentially life-threatening condition.

The creation of the dataset stems from the most recent edition of the Radiology Society of North America (RSNA) Artificial Intelligence (AI) Challenge. For the 2019 edition, participants were asked to create an ML algorithm that could assist in the detection and characterization of intracranial hemorrhage on brain CT. Accuracy in diagnosing the presence and type of intracranial hemorrhage is a vital part of effective treatment, as even a small hemorrhage can lead to death if it is in a critical location.

Rather than using an existing dataset, as had been done for the first two challenges, the competition's organizers set out to create one from scratch. They compiled the brain hemorrhage CT dataset from three institutions: Stanford University in Palo Alto, California, Universidade Federal de São Paulo in São Paulo, Brazil, and Thomas Jefferson University Hospital in Philadelphia, Pennsylvania.

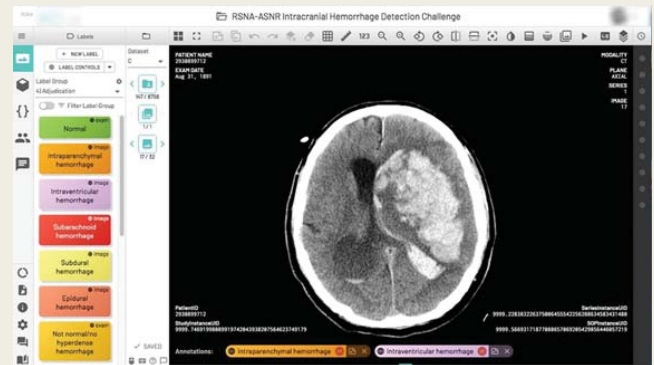
"The value of this challenge is to create a dataset that might lead to a generalizable solution, and the best way to do that is to train a model from data originating from multiple institutions that use a variety of CT scanners from various manufacturers, scanning protocols and a heterogeneous patient population," said the lead author, Dr. A E Flanders, neuroradiologist at Thomas Jefferson University Hospital. *"In this case, we had data from three institutions and international participation. The dataset is unique, not only in terms of the volume of abnormal images but also the heterogeneity of where they all came from."*

RSNA and the American Society of Neuroradiology (ASNR) collaborated to curate the dataset and organizers issued an open call for volunteers within the ASNR membership to annotate the images. A day-and-a-half later, they had 140 volunteers from which they selected 60 to annotate a vast trove of 874,035 brain hemorrhage CT images in 25,312 unique exams. The volunteers marked each image as normal or abnormal. For the abnormal images, they indicated the hemorrhage subtype.

"It was a nail-biter all the way along," Dr. Flanders said of the process. *"We were building the airplane while it was in flight. When you consider the number of images that we had to de-identify locally, consume, curate, label, cross-check*

"... the generation of the largest public collection of expert-annotated brain hemorrhage CT images..."

"...The dataset was released under a non-commercial license, meaning it is freely available to the AI research community ..."



A complex multicompartimental cerebral hemorrhage on a single axial CT image displayed using the annotation tool in a single portal window. Hemorrhage labels (left column) are shown that are relevant to the image display on the bottom of the image once selected. Image credit RSNA

and then organize into just the right datasets to release to the contestants, there was a lot of work involved by the volunteer workforce, the RSNA Machine Learning Subcommittee, data scientists, contractors and RSNA staff."

The dataset's release attracted interest from far and wide. Organizers received more than 22,200 submissions from 1,787 individual competitors in 1,345 teams from 75 countries. Dr. Flanders was particularly struck by the international reach of the project and the level of enthusiasm even from people outside of the medical realm.

The dataset was released under a non-commercial license, meaning it is freely available to the AI research community for non-commercial use and further enhancement.

Dr. Flanders said the objective of engaging with a subspecialty society to leverage their unique expertise in developing a high-quality dataset is an effective and useful pathway to follow for future collaborations. The model worked so well that organizers are using it again for this year's competition, a collaboration with the Society of Thoracic Radiology seeking improved detection and characterization of pulmonary embolism on chest CT.

"I was really impressed by the huge volunteer effort and the tremendous worldwide interest in this project," Dr. Flanders said. *"The dataset we created for this challenge will endure as a valuable ML research resource for years to come."*

REFERENCE

1. Flanders AE et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence* 2020; 2(3):e190211 . doi 10.1148/ryai.2020190211