AN IMAGE LABELING APP USING ACTIVE LEARNING

Nicholas Kebbas, Manjun Lu
College of Arts and Sciences, University of San Francisco

ACTIVE LEARNING

The active learning process involves choosing to only manually label the data that will be most informative to the model. This way one can begin with only a relatively small amount of labeled data, and selectively add to it. The goal is to achieve greater accuracy with fewer labeled training instances and therefore, reduce the total labeling cost. Active Learning is well-suited when unlabeled data is abundant, but labeling is expensive.[1]

In this project we are using the Ridge Regression model for the Active Learning analysis. At each training step, the model returns a visualization that the researcher can use to identify the lowest confidence cases. They can then manually label the images, and retrain the model.

The researcher can choose to follow this process until the model has achieved the desired accuracy.

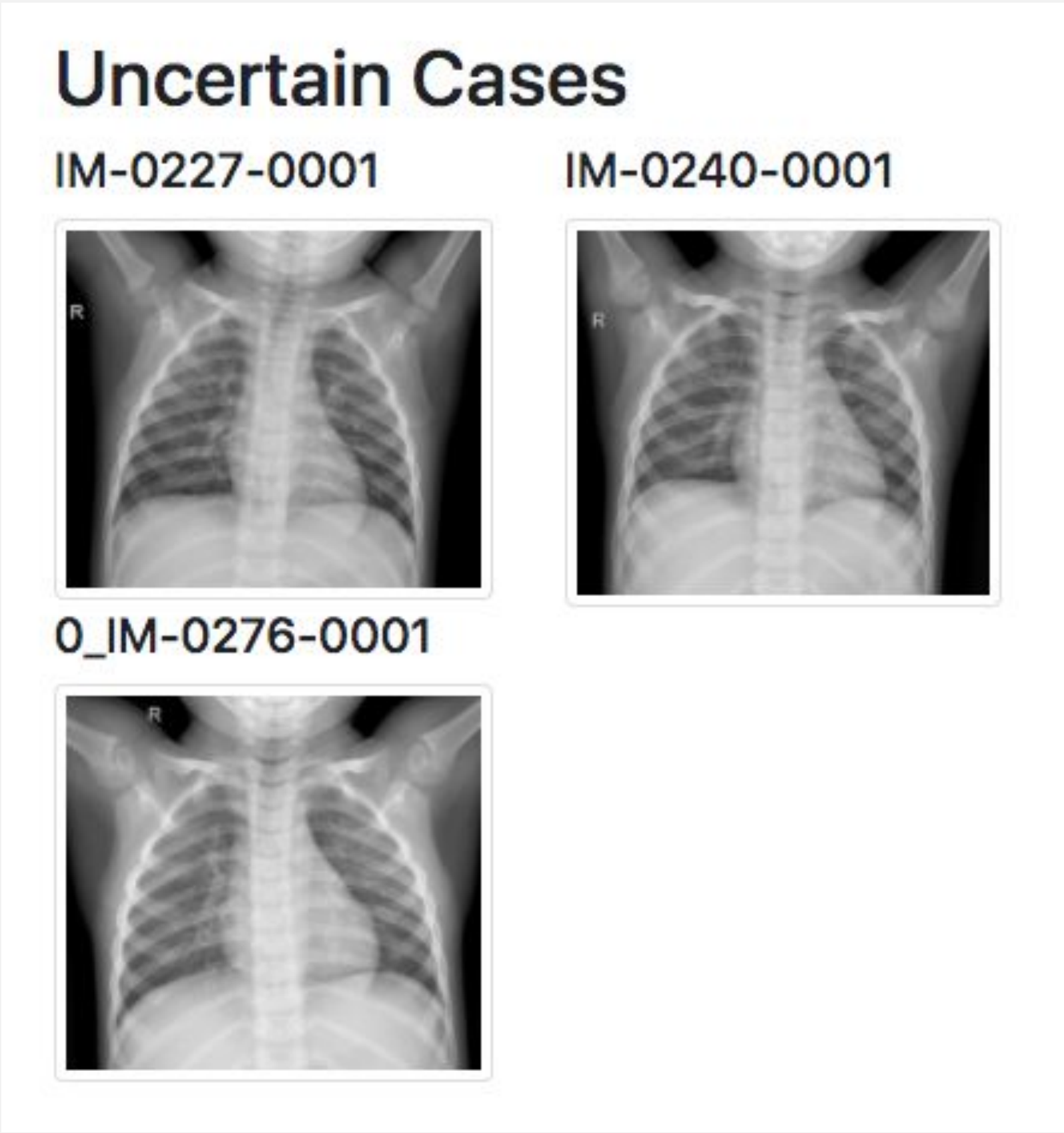

Figure 1. After originally training the model, the researcher will be shown which images the model was most uncertain of.

They can then manually label these images and choose to retrain the model with this new data.

DATASETS AND MODELING

Datasets: We used the following imaging datasets in this application:

- Chest X-Ray Images (Pneumonia) - kaggle [2]
  - 5,863 labeled images of Chest X-Rays
  - 2 categories: Pneumonia/Normal
- MNIST Handwritten Digit Classification Dataset [3]

Models: We used the following pre-trained keras models:

- DenseNet121

ABSTRACT

Unlabeled image data is widely available, but it is both time-consuming and expensive to manually label this image data, so the amount of labeling should be minimized. Therefore, when creating a mathematical model that attempts to perform image classification tasks on large amounts of mostly unlabeled data, it is critical to choose images that provide the maximum amount of information to the model about the classification task. Active learning is a popular technique in the field of machine learning in which a subset of labeled image data is used to train a model, and that model is used to generate the probabilities of each image in an unlabeled subset belonging to a particular class. Researchers can then prioritize labeling of images with the lowest probabilities in order to make the most significant impact.

Our team has extended the work of a previous group, and we have developed a web application that identifies and visualizes the lowest confidence cases, so that researchers can selectively label the images that will have the greatest impact on the model. Thus, we can minimize the amount of labeling a researcher would need to do to improve their model.

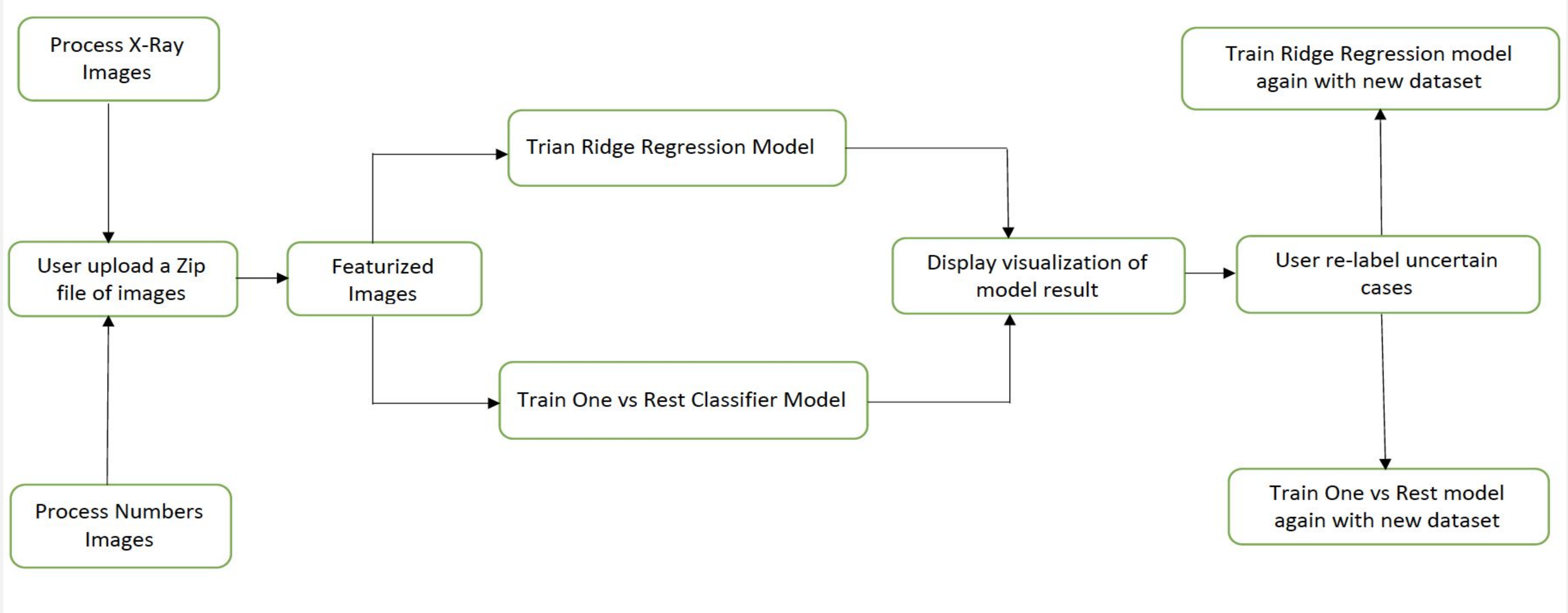
Process X-Ray Images
User upload a Zip file of images
Process Numbers Images
Featurized Images
Trian Ridge Regression Model
Train One vs Rest Classifier Model
Display visualization of model result
User re-label uncertain cases
Train Ridge Regression model again with new dataset
Train One vs Rest model again with new dataset

Figure 2. Django app design flowchart

IMPLEMENTATION

Active Learning:

We converted the previous group's active learning mathematical models from R code to Python code and added multiclass classification.

Web Application Design:

We used the Python framework Django to develop our web application, which follows Object Oriented Design Patterns and includes interfaces for batch image upload, image labeling, and data visualizations

[1] Settles B. Active learning literature survey. University of Wisconsin-Madison Department of Computer Sciences; 2009.
[2] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", Mendeley Data, v2 http://dx.doi.org/10.17632/rscbjbr9sj.2
[3] Colianni, Stuart (2017), "MNIST as .jpg", https://www.kaggle.com/scolianni/mnistasjpg

Image Labeling

The purpose of creating an application for image labeling is to reduce the workload and cost of labeling data by professionals. Usually, a supervised Machine Learning model requires large amounts of training data. However, manually creating labels is inefficient and expensive.

Therefore, our group created a web application for image classification, which runs the Ridge Regression Model for Active-Learning analysis and retrains the model after the user manually re-labels the uncertain cases. Our application allows users to do binary classification and multiclass classification.

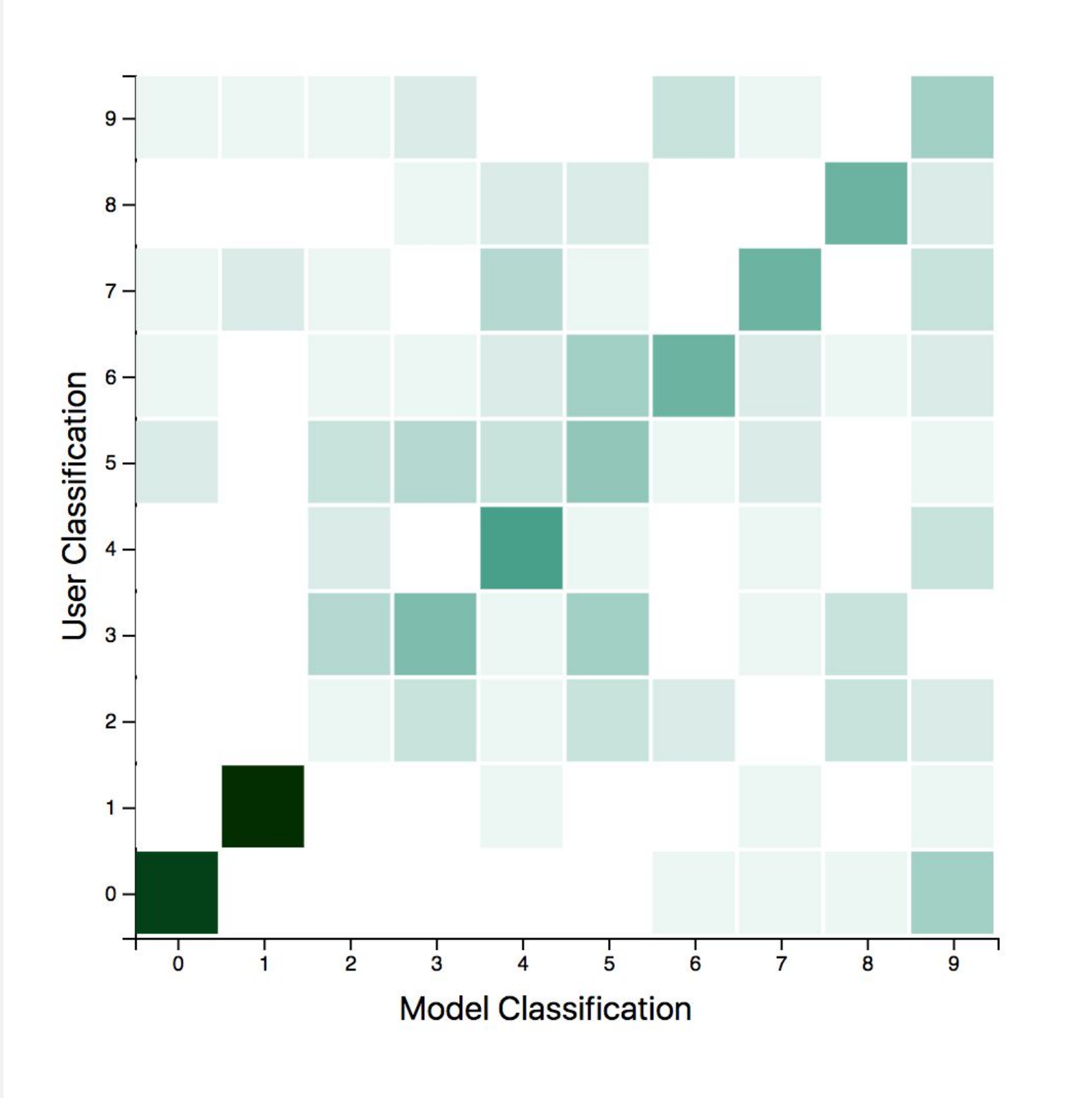

Figure 3. Visualization of User's label vs Model's label. Researchers will view and manually label images outside of the diagonal line through the origin.

FUTURE WORK

- ➢ Add flexibility to web application so that it works with any data set and any number of labels
- ➢ Research a strategy to find the best possible points for tuning the lambda values in Internal Cross-validation for Active Learning.
- ➢ Deploy the application to the web

Advisor:

Dr. Robert Horton, Senior Data scientist at Microsoft

Previous Group:

Tyler Iams, Omid Khazaie, Soodabeh Sarafrazi