Using AI to Label Laboratory Test Results

Phase 2

Iris Gao | Jackie Lam | Tae Yoon Lee

DSSG 2018

TEAM

Joy (Sizhe) Chen

Kenny Chiu

William Lu

Nelly (Nilgoon) Zarei

Motivation

- Laboratory test results are ***unstructured*** text.
- Reviewing and labeling lab results are ***manually*** done. 
- ***Time-intensive*** but **necessary** for population-level public health studies.
- The goal is to ***automate*** the labeling process using AI. 

Outline

Unstructured Data

Lab Test Result Description
no specimen received
bordetella parapertussis $ positive
no growth of salmonella
hide
positive for shiga toxin stx1 gene by pcr | although the genes isare present toxin expression may be variable clinical correlation is required | isolate identified as ecoli non o157 | specimen has been forwarded to a reference laboratory for further characterization | isolate serotyped as $ escherichia coli $ o117h7
respiratory syncytial virus detected by rt pcr using an assay capable of detecting influenza a b and rsv $ indeterminate for influenza b virus by rt pcr submit a repeat specimen if clinically indicated

Table 1: Examples of laboratory test results.

Existing Labeled Data (n=400k)

Lab Test Result Description	Test Performed	Test Outcome	Organism Name
no specimen received	No	NA	NA
bordetella parapertussis $ positive	Yes	Positive	bordetella parapertussis
no growth of salmonella	Yes	Negative	salmonella
hide	Yes	NA	NA
respiratory syncytial virus detected by rt pcr using an assay capable of detecting influenza a b and rsv $ indeterminate for influenza b virus by rt pcr submit a repeat specimen if clinically indicated	Yes	Positive	respiratory syncytial virus

Table 2: Examples of partially labeled lab test results

Initial Objective: Three Classification Problems

1. Test Performed

It has two classes: Yes and No.

1. Test Outcome

It has four classes: Positive, Negative, Missing, and Indeterminate.

3. Organism Name

There are 27 classes in the partially labeled data. While the first outputs have fixed numbers of classes, the number of classes for Organism Name may increase over time.

Next Objective

Find all the organism names in a test result and their corresponding test outcomes.

Lab Test Result Description	Test Performed	Test Outcome	Organism Name
no growth of salmonella	Yes	Negative	Salmonella
respiratory syncytial virus detected by rt pcr using an assay capable of detecting influenza a b and rsv $ indeterminate for influenza b virus by rt pcr submit a repeat specimen if clinically indicated	Yes, Yes, Yes	Indeterminate, Positive, Negative	Influenza B, Respiratory Syncytial virus, Influenza A

Table 3: Examples of fully labeled lab test results

Outline

Section 1 Problem Formulation

Section 2 Machine Learning Methods

Section 3 New Methods and Results

Section 4 Future Work

Machine Learning Workflow

1. Pre-processing
2. Feature Engineering
3. Train
4. Test

Feature Engineering: Bag of Words

Result Description 1: bordetella parapertussis | positive

Result Description 2: bordetella parapertussis | negative

vectorizer

[1 1 1 0]
[1 1 0 1]

	bordetella	parapertussis	positive	negative
Result description 1	1	1	1	0
Result description 2	1	1	0	1

MetaMap: An Annotation Tool

'Respiratory syncytial virus detected by RT PCR using an assay capable of detecting influenza A B and RSV $ indeterminate for influenza B virus by RT PCR'

MetaMap Annotation

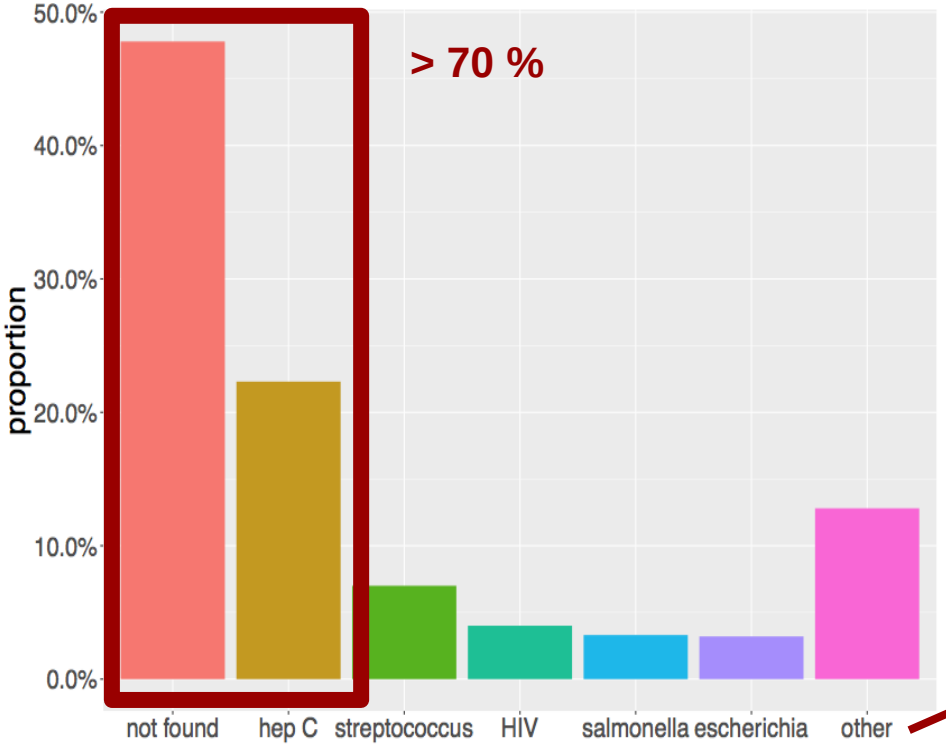
'Respiratory syncytial virus detected by RT PCR using an assay capable of

Virus | Finding | Molecular Biology Research Technique | Functional | Lab Procedure

detecting influenza A B and RSV. Inde r influenza B virus by RT PCR'

Finding | Idea | Virus | Molecular Biology Research Technique

Evaluating Classifier Performance: Accuracy



- Accuracy is a common measure to evaluate classifier performance, but not the best measure for data that has **unbalanced** classes.
- A model can still achieve high accuracy if it fails to correctly classify all small classes but predicts large classes well.

Figure 2: Distribution of Organism Names

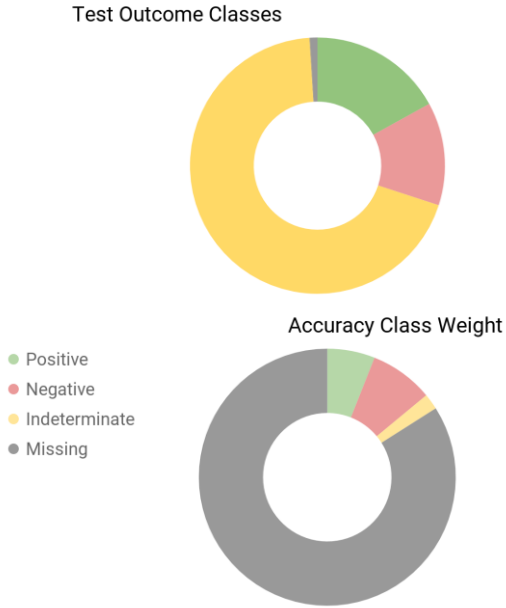
Inversely weighted Accuracy

Motivation: Since there are serious class imbalances for test outcome and organism name, we want to put more emphasis on correctly predicting the smaller classes.

Solution: Take a weighted average of accuracy, where the smaller class have larger weights associated with them.

Test Outcome	Proportion of total data	Class weight
Positive	0.17	0.06
Negative	0.13	0.08
Missing	0.69	0.02
Indeterminate	0.01	0.84

Table 4: Test Outcome class weights



F-score

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

Precision: True Positive / Predicted Positive

Recall: True Positive / Actual Positive

F-score: (Harmonic) average of Precision and Recall

For an output with more than two classes, we can calculate the F-score for each class and take the **inversely weighted average** to account for class imbalance.

Results - Test Performed

Test Performed	Bag of Words	Bag of Words + Number of Observations + Test Code + Metamap Candidates
Accuracy	0.996	0.999
F2-score	0.990	0.999

Table 5: Test Performed Results

Results - Test Outcome

Test Outcome	Bag of Words	Bag of Words + Number of Observations + Test Code + Metamap Candidates
Accuracy	0.992	0.994
Inversely-weighted accuracy	0.960	0.974
Inversely-weighted F1-score	0.944	0.961

Table 6: Test Outcome Results

Results - Organism Name

Organism Name	Bag of Words	Bag of Words + Number of Observations + Test Code + Metamap Candidates
Accuracy	0.947	0.956
Equally weighted F1 score	0.790	0.874
Inversely Weighted Accuracy	0.636	0.866
Inversely Weighted F1 score	0.664	0.848

Table 7: Organism Name Results

Outline

Section 1 — Problem Formulation

Section 2 — Machine Learning Methods

Section 3 — **New Methods and Results**

Section 4 — Future Work

New Objective

Find all the organism names in a test result and their corresponding test outcomes.

- The prior machine learning methods would not work well for this:
  - Multiclass, multilabel problem, can be used for find organism name
  - There would be no way of associating which label of test outcome goes with which organism name
  - Need a model that take into account the sequence of the text

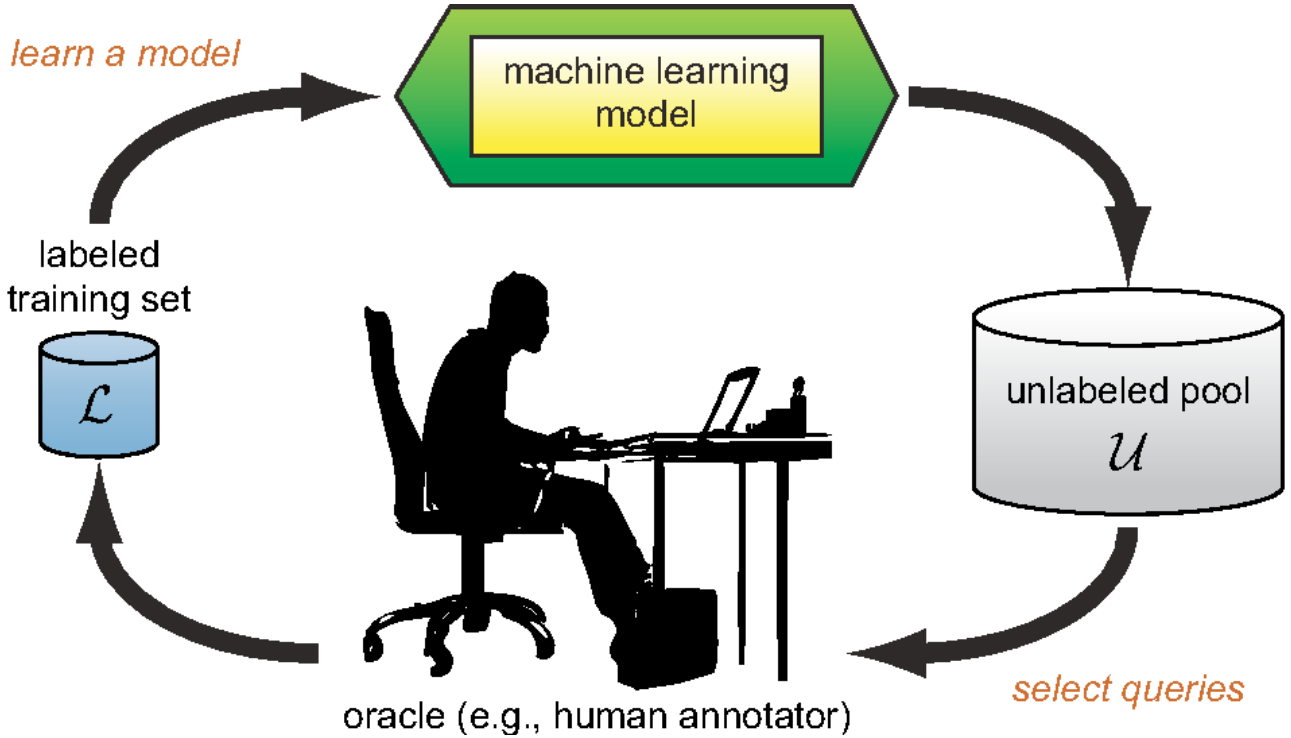
Problem - Lack of Labeled Data

Lab Test Result Description	Test Performed	Test Outcome	Organism Name
respiratory syncytial virus detected by rt pcr using an assay capable of detecting influenza a b and rsv $ indeterminate for influenza b virus by rt pcr submit a repeat specimen if clinically indicated	Yes	Positive	Respiratory Syncytial virus
respiratory syncytial virus detected by rt pcr using an assay capable of detecting influenza a b and rsv $ indeterminate for influenza b virus by rt pcr submit a repeat specimen if clinically indicated	Yes, Yes, Yes	Indeterminate, Positive, Negative	Influenza B, Respiratory Syncytial virus, Influenza A

What we have: Partially Labelled

What we need: Fully Labelled

Solution- Active Learning

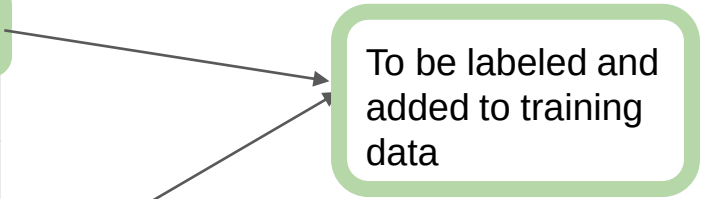


Active Learning- Query Mechanism

Uncertainty Sampling: Train model on existing labeled data, run model on all existing unlabeled data, and select the ones where the classifier is most uncertain about to be labeled.

Example: Querying 2 observation

Unlabeled Data	Prediction	Model Confidence
result_1	Positive	0.90
result_2	Negative	0.2
result_3	Negative	0.8
result_4	*Missing	0.7
result_5	Indeterminate	0.4



HCV Dataset

We will prototype the new objective by using the Hepatitis C Virus Dataset

- Guaranteed to only contain a single organism name (Hepatitis C Virus)
- Similar Class Imbalance:

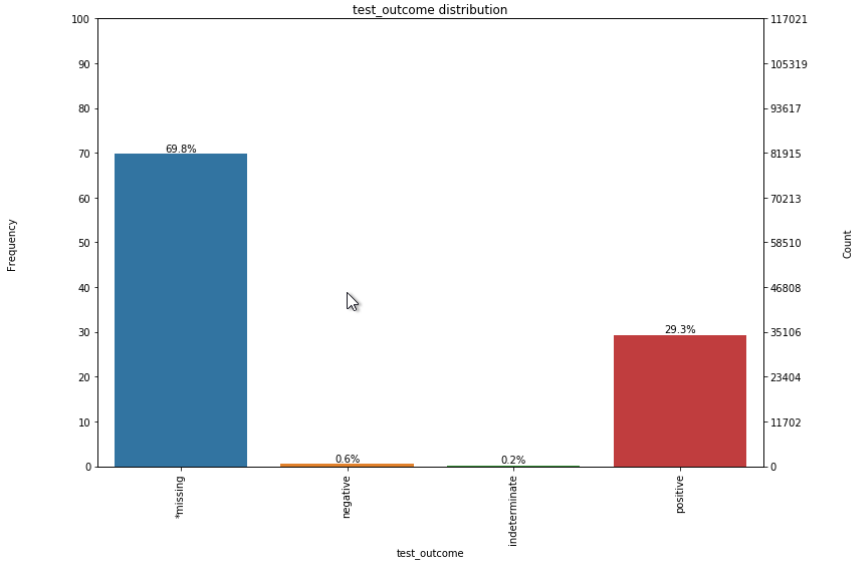


Figure 3: HCV Class Distribution

Active Learning vs. Random Learning

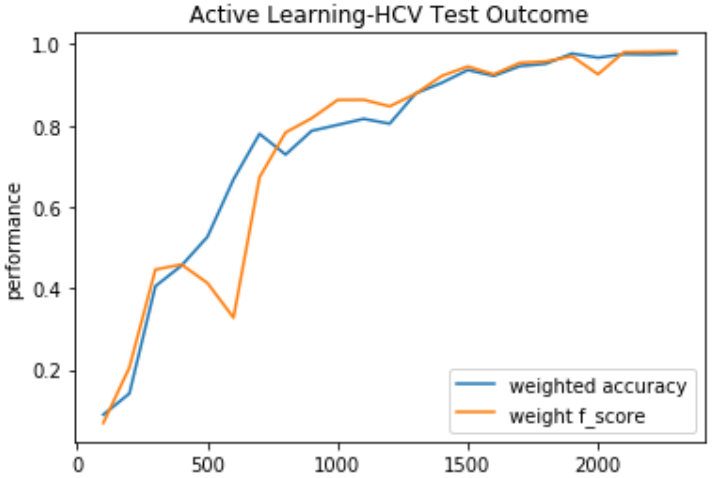


Figure 4: Active Learning Test Outcome

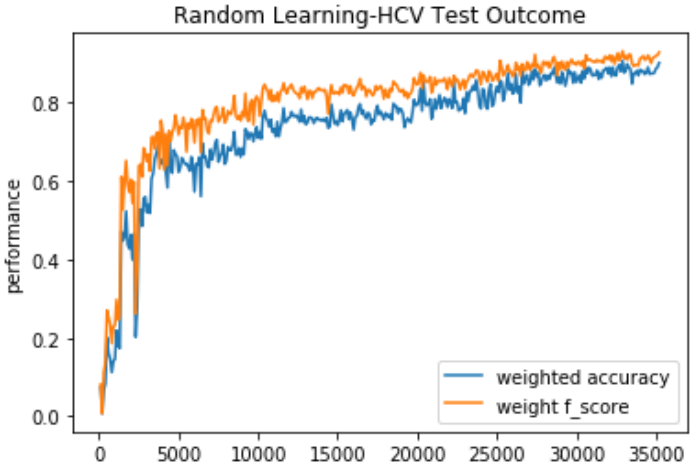


Figure 5: Random Learning Test Outcome

Using Active Learning, we can achieve the same model performance with much less labeled data.

Uncertainty Sampling vs. Random Sampling

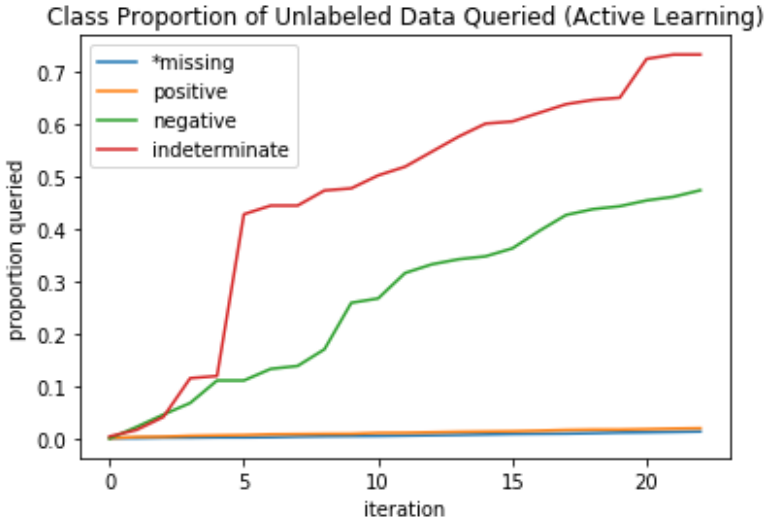


Figure 6: Uncertainty Sampling

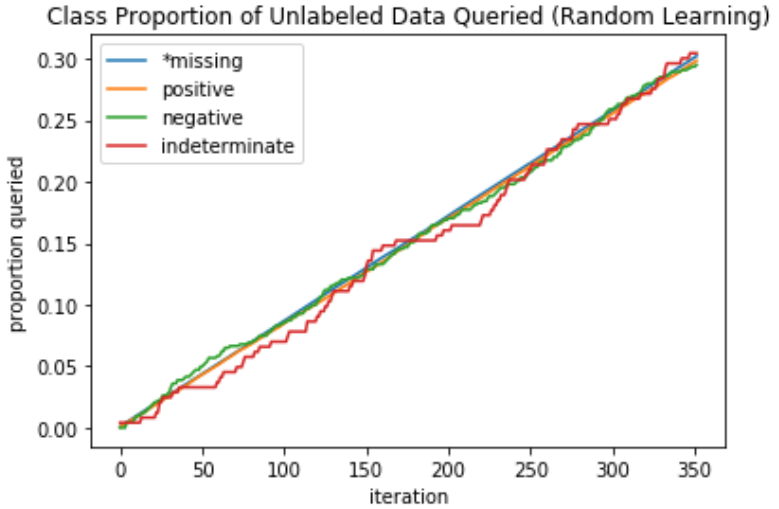


Figure 7: Random Sampling

With Active Learning Query Mechanism, sample that the model is most uncertain about is queried, resulting in querying more samples from the smaller classes.

Active Learning Performance by Class Accuracy

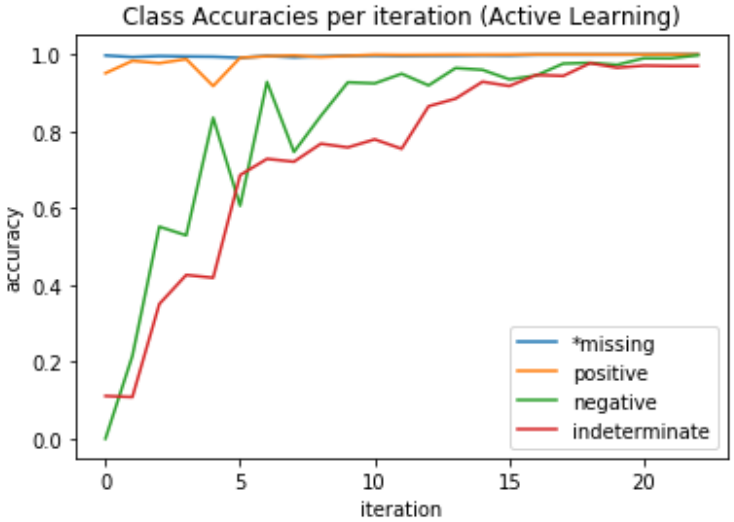


Figure 8: Active Learning Class Accuracies

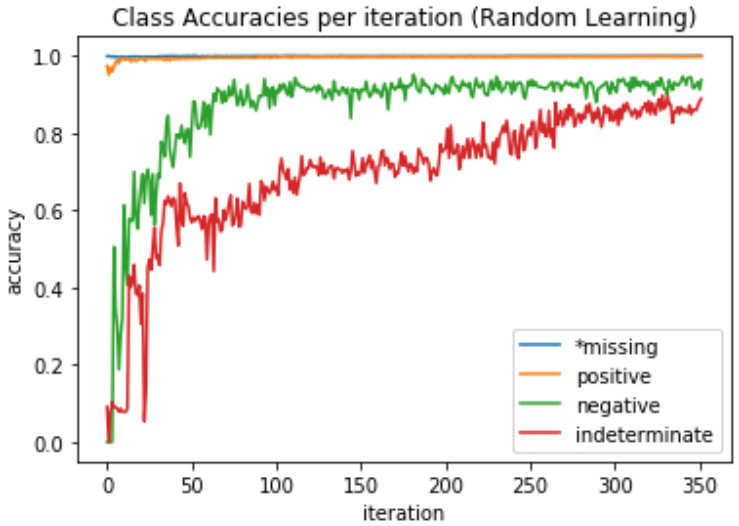
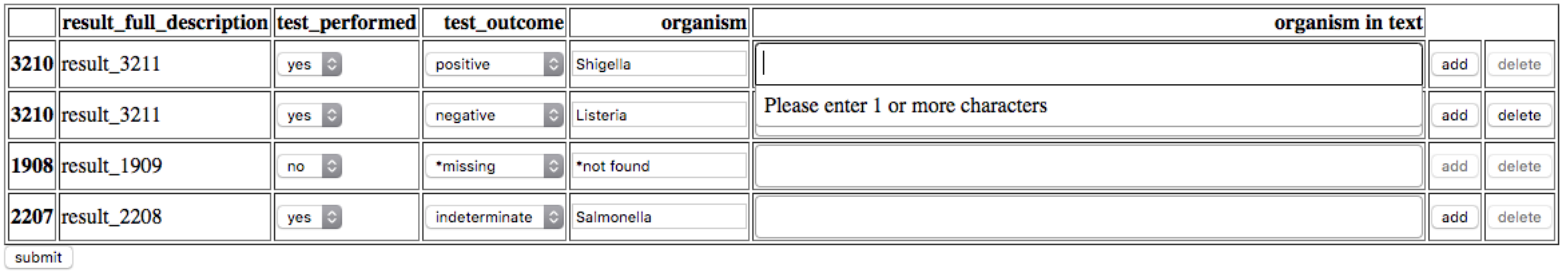


Figure 9: Random Learning Class Accuracies

Initially, the models do not predict the small classes well. Using Active Learning, the model quickly gets better at predicting the smaller classes by querying more samples from the small classes.

Finding Organism Name

- First, obtain some labeled data to build a baseline model that we will use moving forward.
  - **Classic Machine Learning Model**
  - **Deep Learning Model**
- Using active learning approach as previously outlined to obtain more labeled data (labeled by humans), further refine and develop the model to find all organism names in result description



	result_full_description	test_performed	test_outcome	organism	organism in text		
3210	result_3211	yes	positive	Shigella		add	delete
3210	result_3211	yes	negative	Listeria	Please enter 1 or more characters	add	delete
1908	result_1909	no	*missing	*not found		add	delete
2207	result_2208	yes	indeterminate	Salmonella		add	delete

submit

Need data to move forward with this objective

Deep Learning- Approach

- Assuming that we have a good model for obtaining all the organism names (assumes organism names model built)
  - For each organism found in a result description, duplicate the result description and assign one as “target” and the others as “other”

result_full_description
respiratory syncytial virus detected by rt pcr using an assay capable of detecting influenza a b and rsv $ indeterminate for influenza b virus by rt pcr submit a repeat specimen if clinically indicated

result_full_description	test_outcome	organism_name
<TARGET> detected by rt pcr using an assay capable of detecting <OTHER> and rsv $ indeterminate for <OTHER> virus by rt pcr submit a repeat specimen if clinically indicated	Indeterminate	Influenza B,
<OTHER>detected by rt pcr using an assay capable of detecting <TARGET>and rsv $ indeterminate for <OTHER> virus by rt pcr submit a repeat specimen if clinically indicated	Positive	Respiratory Syncytial virus
<OTHER>detected by rt pcr using an assay capable of detecting <OTHER>and rsv $ indeterminate for <TARGET> virus by rt pcr submit a repeat specimen if clinically indicated	Negative	Influenza A

Word Embeddings

- Need to vectorize our result descriptions, however cannot use “bag of words” as we want to maintain the order of the text
- Word Embeddings:
  - N-dimensional vector representation of a word
    - Word used in similar context have similar word embeddings
    - Word Analogies:
      - Eg. King - Man + Woman ≈ Queen
    - Models to train embeddings:
      - Word2vec, GloVe, fastText
  - Using BioWordVec: Pre-trained embedding for biomedical words trained on corpus of biomedical text (uses fastText)
  - Out of Vocabulary: generate random word embeddings for the tokens we created and other unknown words: (<TARGET>, <DATES>,etc.)
  - Further refine embeddings in our Neural Network

Deep Learning- HCV Data

- We do not have enough data to get all organism names, use HCV data (single organism) to test the deep learning approach
  - Since dataset only has single organism, we can use rule based approach to replace organism name with <TARGET> token
- We used a **Recurrent Neural Network** to account for the sequences of text

Results:

Test Outcome (HCV)	Deep Learning (RNN)	Machine Learning (Random Forest)
Accuracy	0.994	0.993
Equally Weighted F-Score	0.929	0.924
Inversely Weighted Accuracy	0.890	0.789
Inversely Weighted F-score	0.890	0.858

Table 8: RNN vs ML Results

Outline

Section 1 — Problem Formulation

Section 2 — Machine Learning Methods

Section 3 — New Methods and Results

Section 4 — Future Work

Future Work and Recommendations

- Use Active Learning for labeling new data
  - Develop the Organism Name Model when more data is there as people begin to label the data
- Further develop the Deep Learning Models
  - Refine hyperparameter choices and model architecture.
  - Custom loss function to account for class imbalances
- Our hope is that through the labeling process, the lab technician could see the value in having more structured result descriptions
- Use the classifiers on new lab reports to speed up the turnaround time from laboratory to patient, allowing patient/physicians to take necessary actions faster

Questions