Comparison of Active Learning Classification Methods on High-Volume Cancer Gene Expression Data

Katrina Liu, Cameron Miller, Qingyi Peng

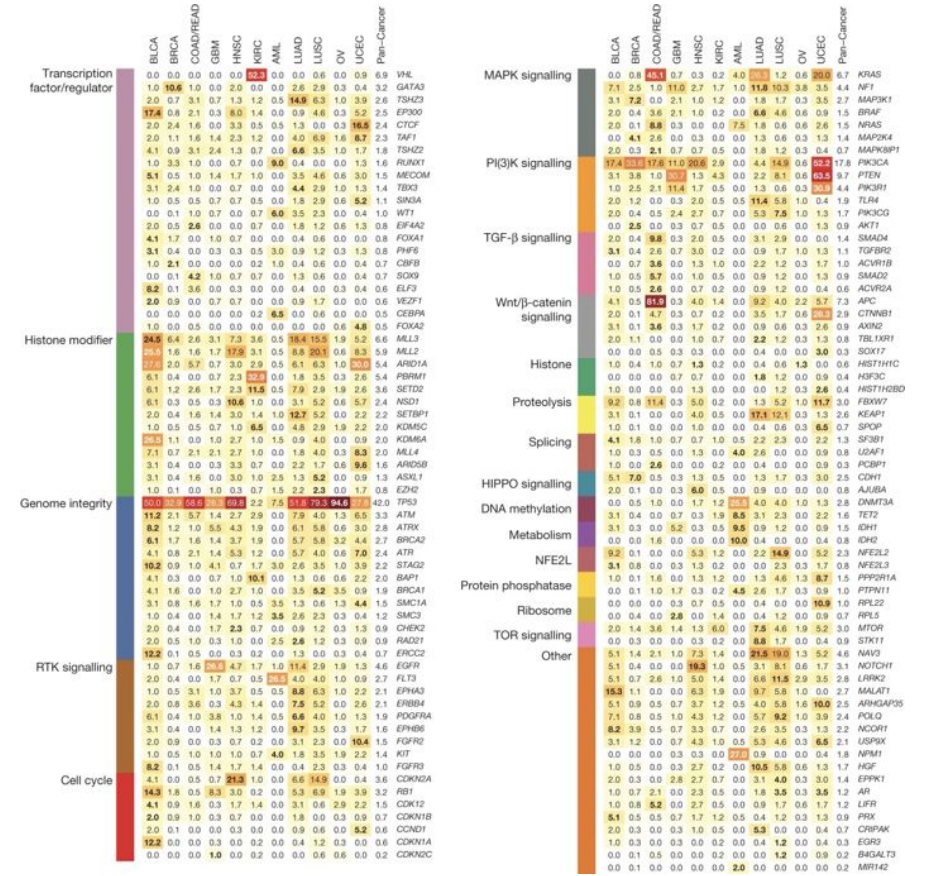
Background

- Surplus of biomedical data
  - Images
  - Genetic Sequences
  - **Gene Expression Profiles/RNA-seq data**
- Annotation has lagged behind information abundance
  - Normal vs. Cancerous Tissue
  - Can active learning ease the burden of human annotation?

    (work by a medical professional)

Background

- To provide a comprehensive basis for testing and comparing learning methods on cancer data, features were selected from a set of significantly mutated genes, and from a variety of mechanisms involved in cancer

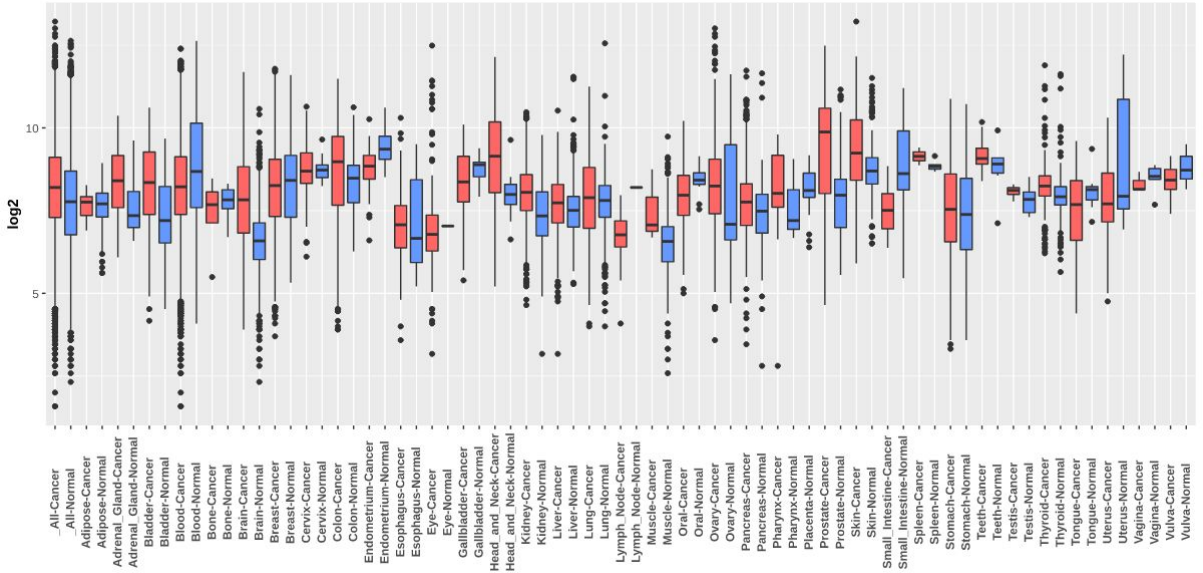


C Kandoth *et al*. *Nature* **502**, 333-339 (2013) doi:10.1038/nature12634

Dataset - GENT2

- A platform for exploring Gene Expression Patterns across Normal and Tumor Tissues
  - Contains over 68,000 samples
  - Gene expression profiles across 72 different tissue types

Gene expression profile across cancer experiments for TP53



Dataset

- Features: gene expression levels of 20 known cancer causing genes
- Labels: cancerous or normal, in our analysis tissue types were grouped and studied separately
- Our analysis was performed on tissue subtypes with computationally feasible amounts of data 100~300 samples: Adipose, Adrenal, Head and Neck, Endometrium

:A	NFE2L2	NAV3	MTOR	KRAS	IDH1	HIST1H1C	GATA3	FBXW7	EGFR	DNMT3A	CDKN2A	CDH1	ARID1A	APC	Label
551	9.672425	7.3750396	6.6724253	9.392318	11.014021	8.169925	7.8641863	7.948367	8.038919	8.939579	7.7879024	10.654636	9.424167	7.9425144	Adipose-Cancer
05	10.699573	7.4429436	6.9541965	8.84235	11.987264	9.769837	8.483816	8.326429	9.592457	7.0768156	5.83289	1.5849625	9.649256	8.209454	Adipose-Cancer
734	10.174926	8.0168085	6.8703647	8.820179	12.912328	8.661778	8.554589	8.3706875	9.61471	6.857981	5.857981	2.321928	9.544965	7.97728	Adipose-Cancer
77	10.188589	7.7879024	7.499846	8.689998	12.615629	10.062046	9.413628	8.228819	9.415742	7.0552826	6.357552	3	9.815383	8.076816	Adipose-Cancer
268	10.337622	7.876517	6.84549	8.97728	12.013672	9.668885	8.40088	7.9943533	9.447083	6.768184	6.044394	3	8.974415	8.442944	Adipose-Cancer
394	10.224002	7.857981	6.6582117	8.83289	12.635944	9.864186	9.515699	8.027906	9.541097	6.9188633	6.33985	4.169925	9.296916	7.7279205	Adipose-Cancer
352	10.440869	7.839204	6.9307375	9.057992	12.557224	10.06609	9.014021	8.519636	9.675957	6.857981	6.6293564	2.807355	9.543032	7.948367	Adipose-Cancer
288	10.127995	7.900867	7.1898246	8.848623	12.139232	10.485829	8.0223675	8.199673	9.139551	6.741467	6.6865005	2	9.324181	7.97728	Adipose-Cancer
944	10.451211	7.357552	6.491853	8.592457	12.531137	10.404078	9.649256	8.243174	9.214319	6.7548876	6.9188633	2.807355	9.505812	7.7279205	Adipose-Cancer
188	10.346514	8.038919	6.2094536	9.271463	11.361396	8.625709	7.139551	8.40088	9.071463	7.5999126	6.022368	3	10.002815	8.257388	Adipose-Cancer
5475	10.2620945	7.321928	6.6724253	8.2946205	12.353973	9.824959	8.308339	8.569856	8.9128895	7.228819	4.087463	1.5849625	8.330916	7.499846	Adipose-Normal
888	10.186114	7.4757333	7.1189413	8.144658	12.213408	10.28077	7.5849624	7.8703647	7.5468946	6.9068904	5.4594316	4.643856	8.495855	6.9541965	Adipose-Normal

Algorithm

- Baseline methods:
  - Random Forest
  - Support Vector Classification with Standard Scaler
- Active Learning Query Strategies:
  - Uncertainty Sampling with Least Confidence Level
  - Query By Committee with Committee Members for Each Gene
  - Expected Error Reduction

Query Strategies

- Uncertainty Sampling

$$x^*_{LC} = \underset{x}{\operatorname{argmax}}\ 1 - P_\theta(\hat{y}|x)$$

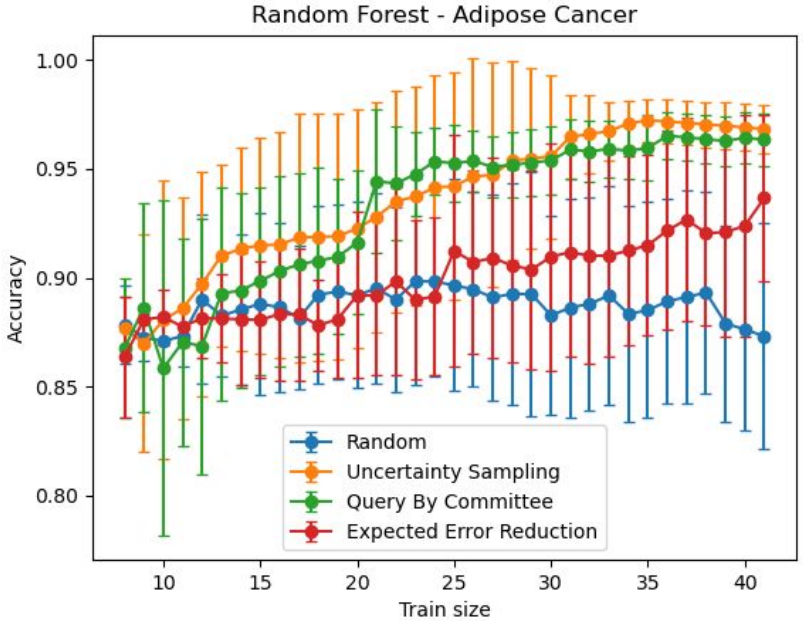
- Query by Committee

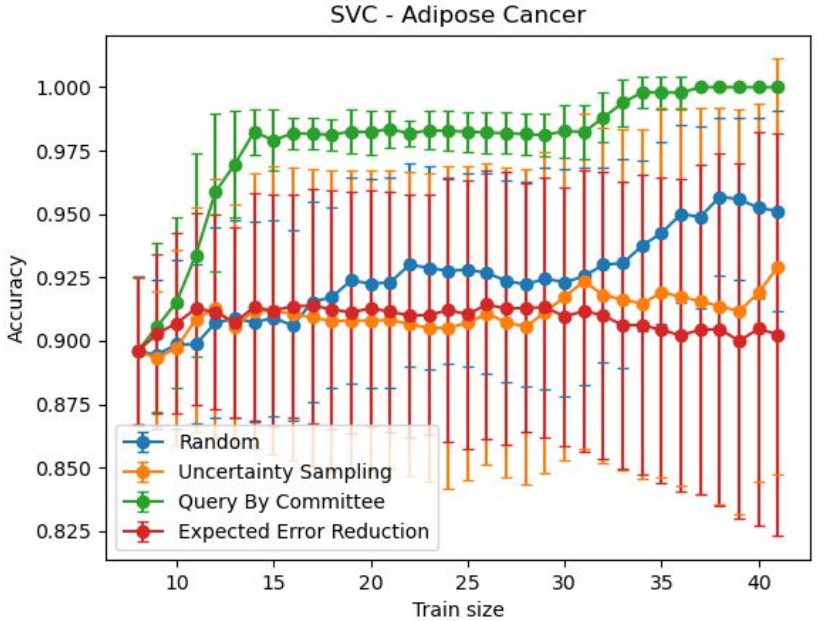
$$x^*_{VE} = \underset{x}{\operatorname{argmax}} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

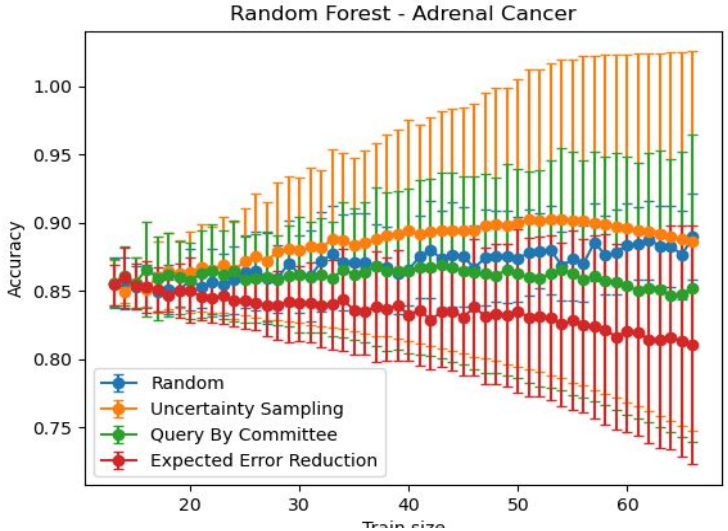
- Expected Error Reduction

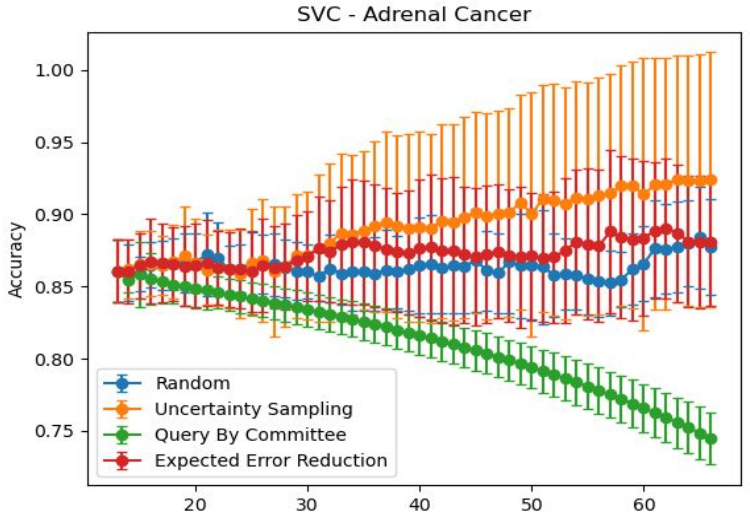
$$x^*_{0/1} = \underset{x}{\operatorname{argmin}} \sum_i P_\theta(y_i|x) \left(\sum_{u=1}^{U} 1 - P_{\theta^+\langle x, y_i \rangle}(\hat{y}|x^{(u)}) \right)$$

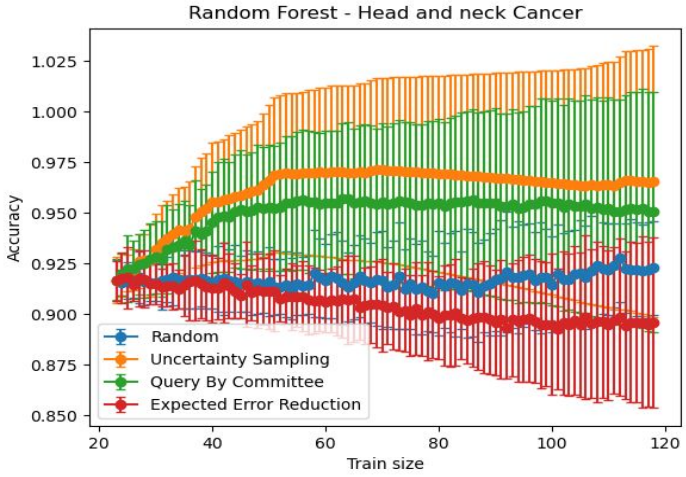
Results

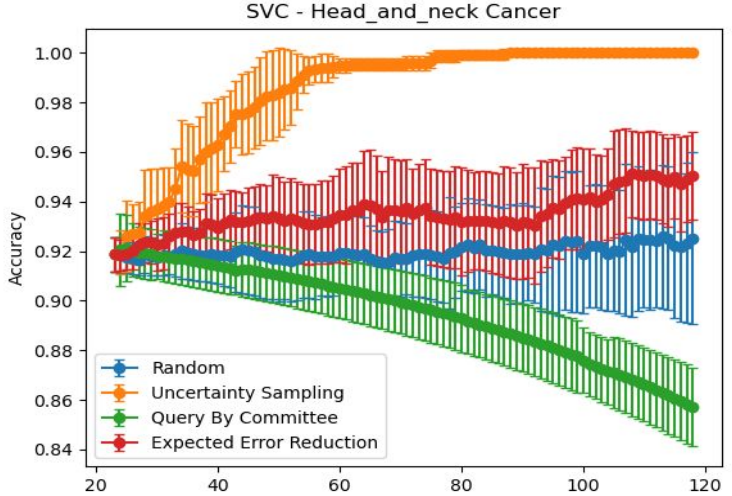

Random Forest - Adipose Cancer
Accuracy
Train size
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction

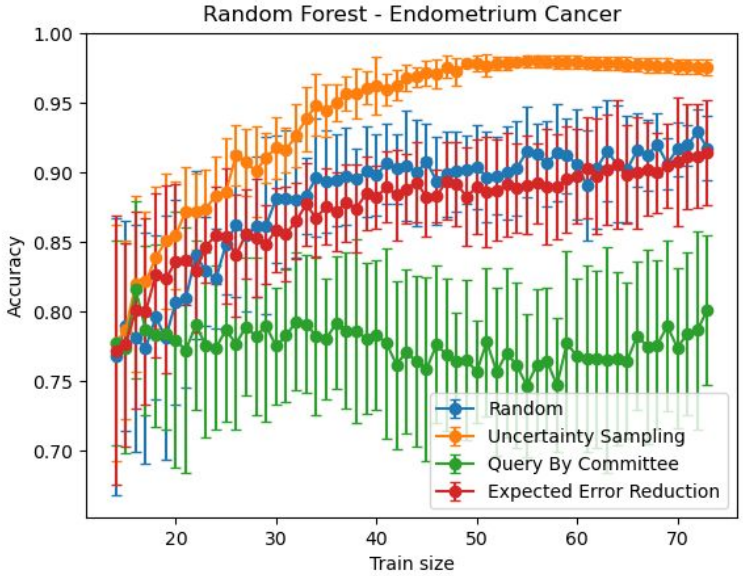

SVC - Adipose Cancer
Accuracy
Train size
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction

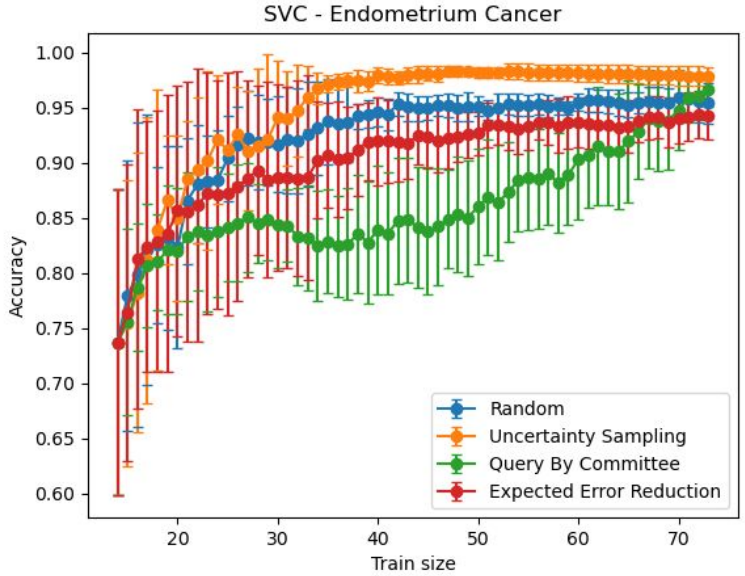

Random Forest - Adrenal Cancer
Accuracy
Train size
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction


SVC - Adrenal Cancer
Accuracy
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction


Random Forest - Head and neck Cancer
Accuracy
Train size
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction


SVC - Head_and_neck Cancer
Accuracy
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction


Random Forest - Endometrium Cancer
Accuracy
Train size
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction


SVC - Endometrium Cancer
Accuracy
Train size
Random
Uncertainty Sampling
Query By Committee
Expected Error Reduction

Analysis of Results

- Uncertainty sampling performs well on both accuracy and efficiency
- Minor drop in accuracy as training sample size increases.
- Query by committee and expected error reduction only performs better than random selection on certain type of cancers.
- Might due to the initial accuracies being too high and the number of samples being small. Therefore, they might be affected by the potential bias existed in the datasets.

Next Steps

- Datasets:
  - Explore datasets of cancer types with larger number of samples.
  - Incorporate gene expressions of more genes in the input data.
- Algorithm:
  - Experiment with different parameters for each strategies.
  - Explore mellow version of each strategies.
  - Explore more active learning query strategies.
  - Improve the efficiency of certain algorithms.

References

1. R. S. Bressan, G. Camargo, P. H. Bugatti and P. T. M. Saito, "Exploring Active Learning Based on Representativeness and Uncertainty for Biomedical Data Classification," in *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2238-2244, Nov. 2019, doi: 10.1109/JBHI.2018.2881155.

2. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* 20, 631–656 (2019). https://doi.org/10.1038/s41576-019-0150-2

3. Park SJ, Yoon BH, Kim SK*, Kim SY*. GENT2: an updated gene expression database for normal and tumor tissues. ***BMC Med Genomics.*** 2019 Jul 11;12(Suppl 5):101. doi: 10.1186/s12920-019-0514-7.

4. Kandoth, C., McLellan, M., Vandin, F. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339 (2013). https://doi.org/10.1038/nature12634