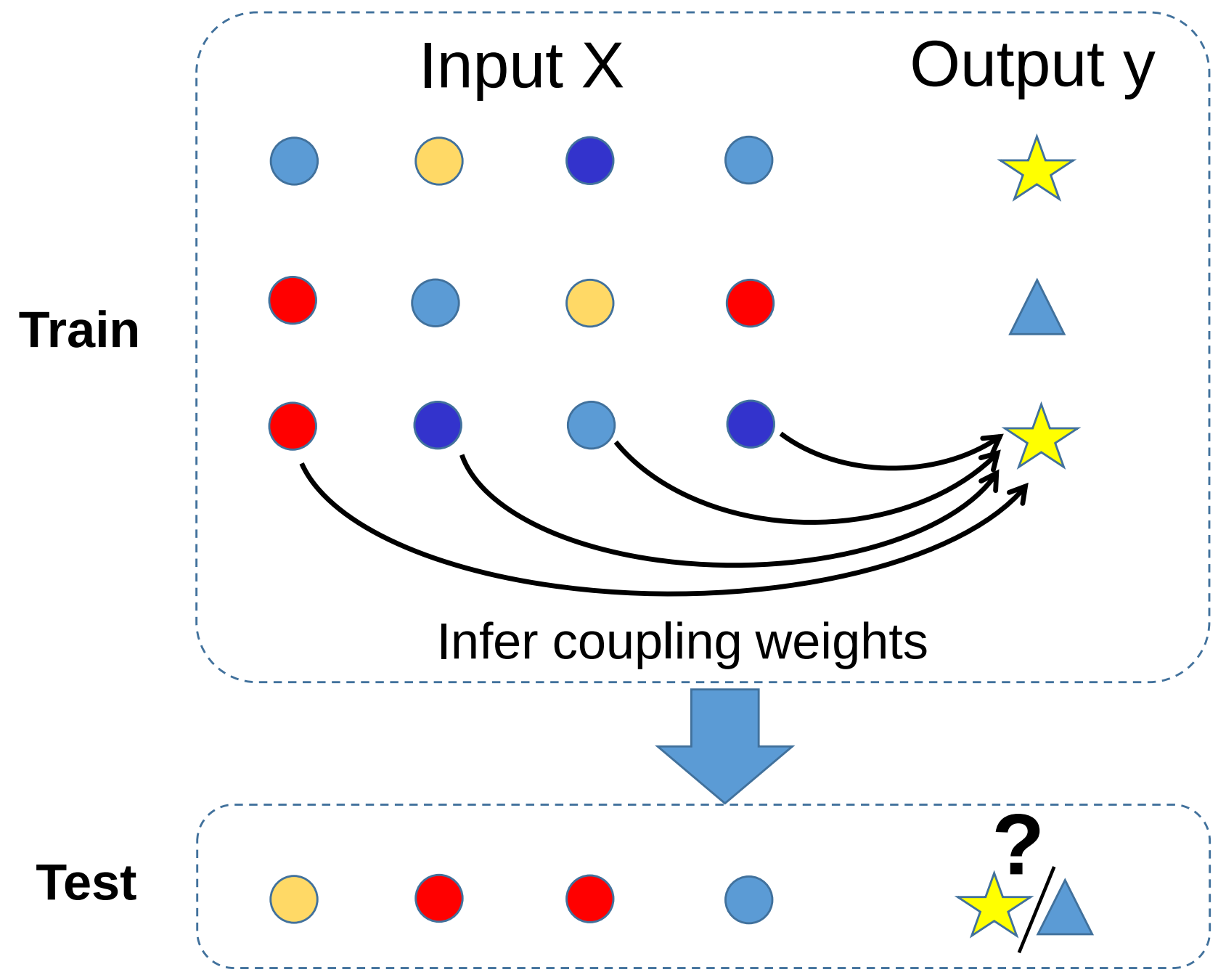
Classification using Expectation Reflection

Danh-Tai Hoang and Vipul Periwal

Laboratory of Biological Modeling, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD.

Introduction

- Classifying instances into one of two or more classes is a fundamental problem, not only in quantitative biology but more generally in data science.
- Mechanistic algorithms identify the influence (coupling weight) from observed variables to target based on a training set, then make a prediction for a new test set.



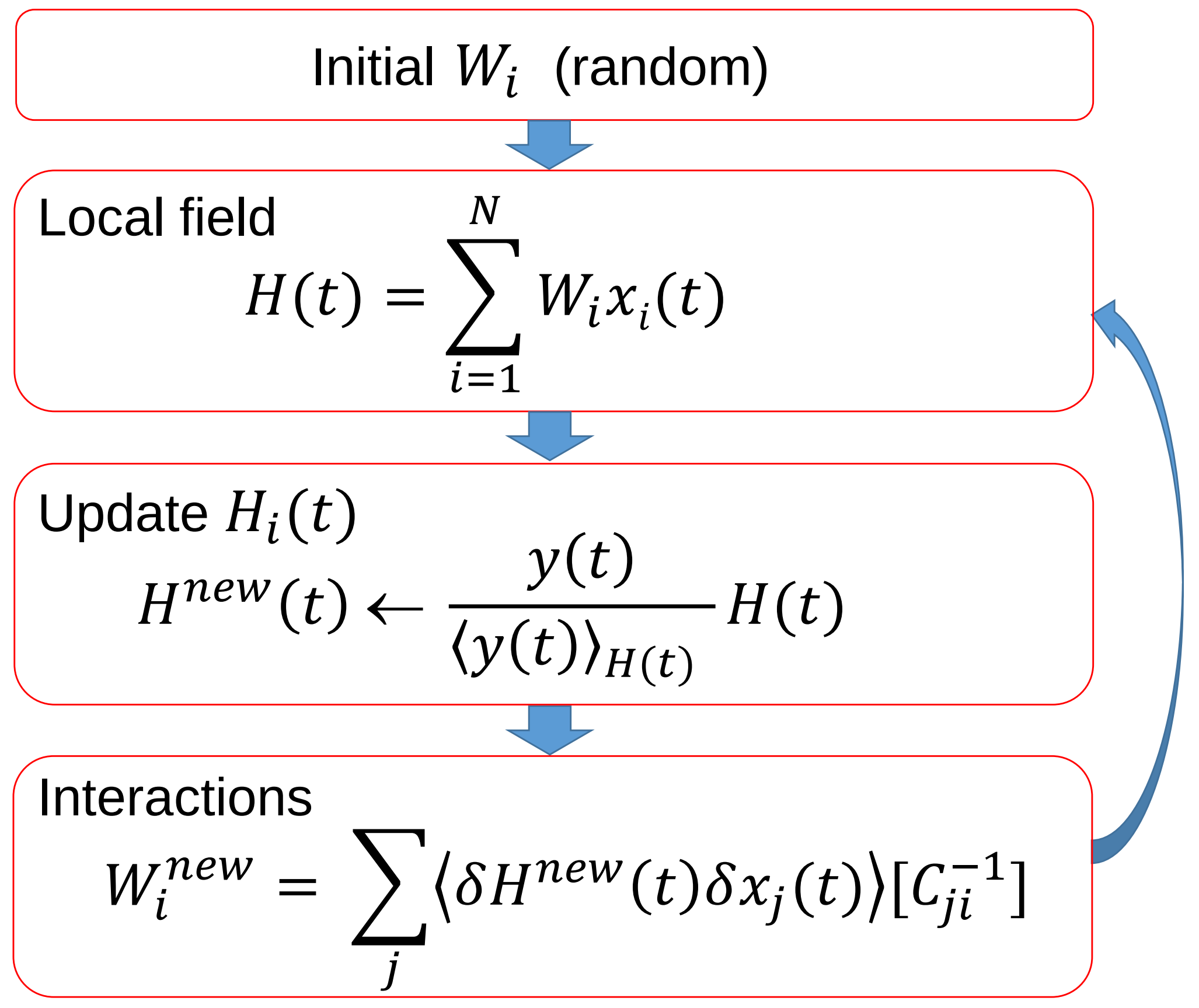
- We recently developed a new data-driven method, Expectation Reflection (ER), that outperforms the current state-of-the-art methods in inferring the network interactions between binary variables, especially in small sample sizes [1]. By introducing hidden variables, ER still works well in partially observed systems [2].
- In this work, we first extend ER to infer interactions from observed features to target in which the features are not restricted to binary but they can be continuous or categorical variables. We then apply to classify biomedical data.

Expectation Reflection Algorithm

❖ Model:

$$P[y(t)|\,X(t)] = \frac{e^{y(t)H(t)}}{\sum_{y(t)} e^{y(t)H(t)}} \qquad \text{where } H(t) = \sum_{i=1}^{N} W_i x_i(t)$$

❖ Method:



Initial W_i (random)

Local field

$$H(t) = \sum_{i=1}^{N} W_i x_i(t)$$

Update $H_i(t)$

$$H^{new}(t) \leftarrow \frac{y(t)}{\langle y(t)\rangle_{H(t)}} H(t)$$

Interactions

$$W_i^{new} = \sum_j \langle \delta H^{new}(t)\delta x_j(t)\rangle [C_{ji}^{-1}]$$

where $C_{ij} = \langle \delta x_i \delta x_j \rangle$ and $\delta f = f - \langle f \rangle$

Stopping criterion: $D(W) = \sum_{t=1}^{L} \left[y(t) - \langle y(t)\rangle_{H(t)}\right]^2$

Expectation Reflection for Classification

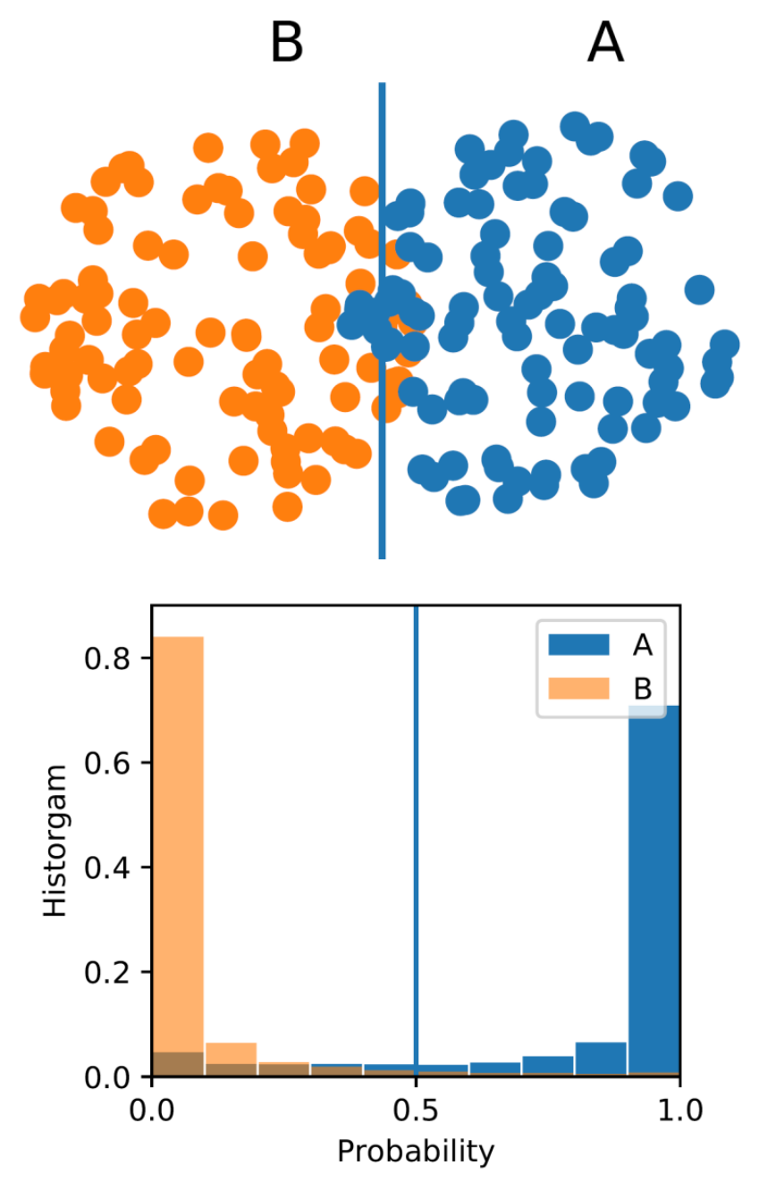
❖ Binary Classification

❑ **Training step:** Inferring interactions W from feature X to target y.
❑ **Testing step:** Predict target y from new feature X and inferred interactions W.

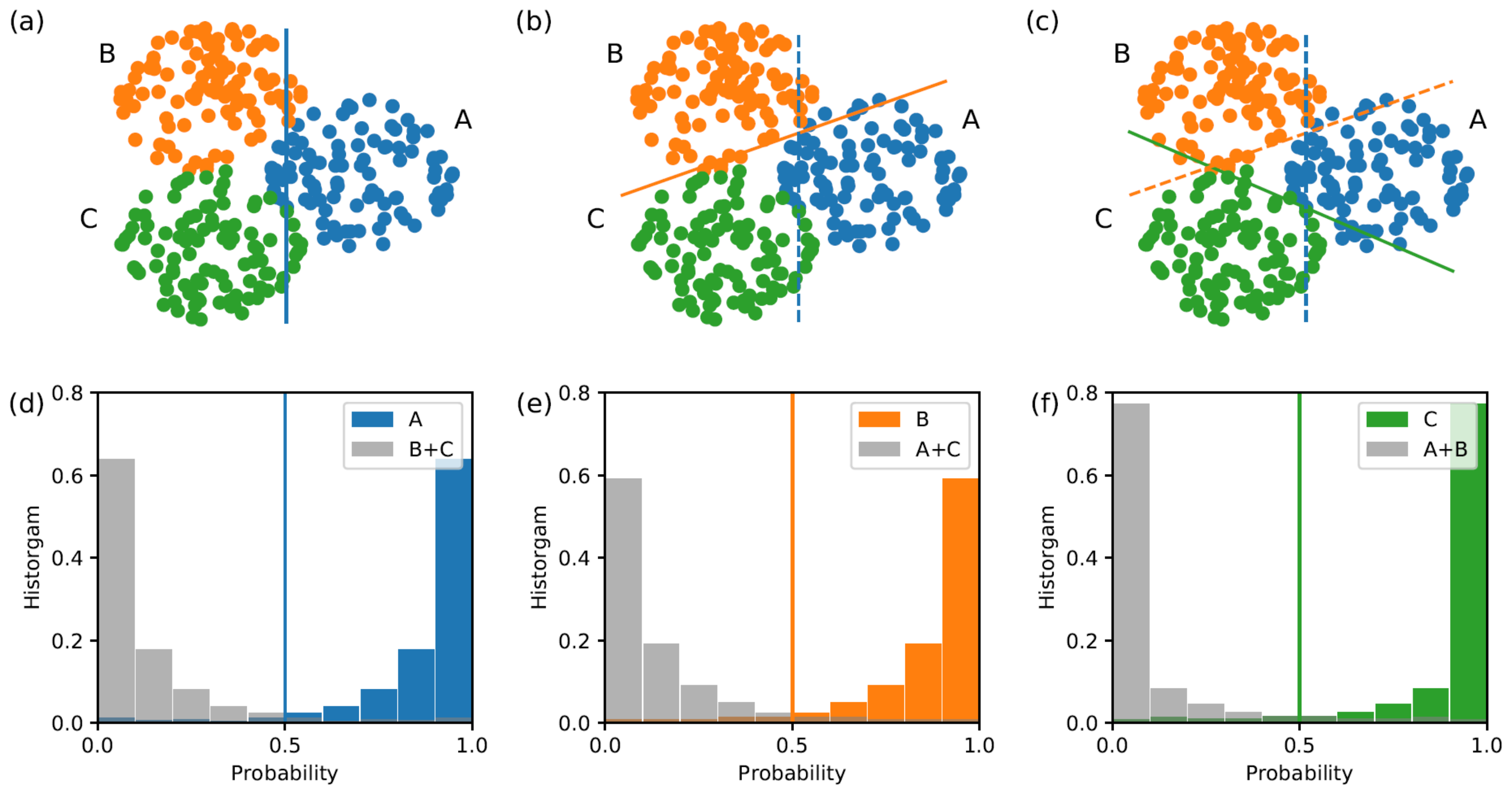
$$P[y(t) = A \mid X(t)] = \frac{e^{+H(t)}}{e^{+H(t)} + e^{-H(t)}}$$

$$\text{where } H(t) = \sum_{i=1}^{N} W_i x_i(t)$$

- ➢ $P \geq 0.5 \rightarrow$ class A
- ➢ $P < 0.5 \rightarrow$ class B



❖ Multi-class Classification

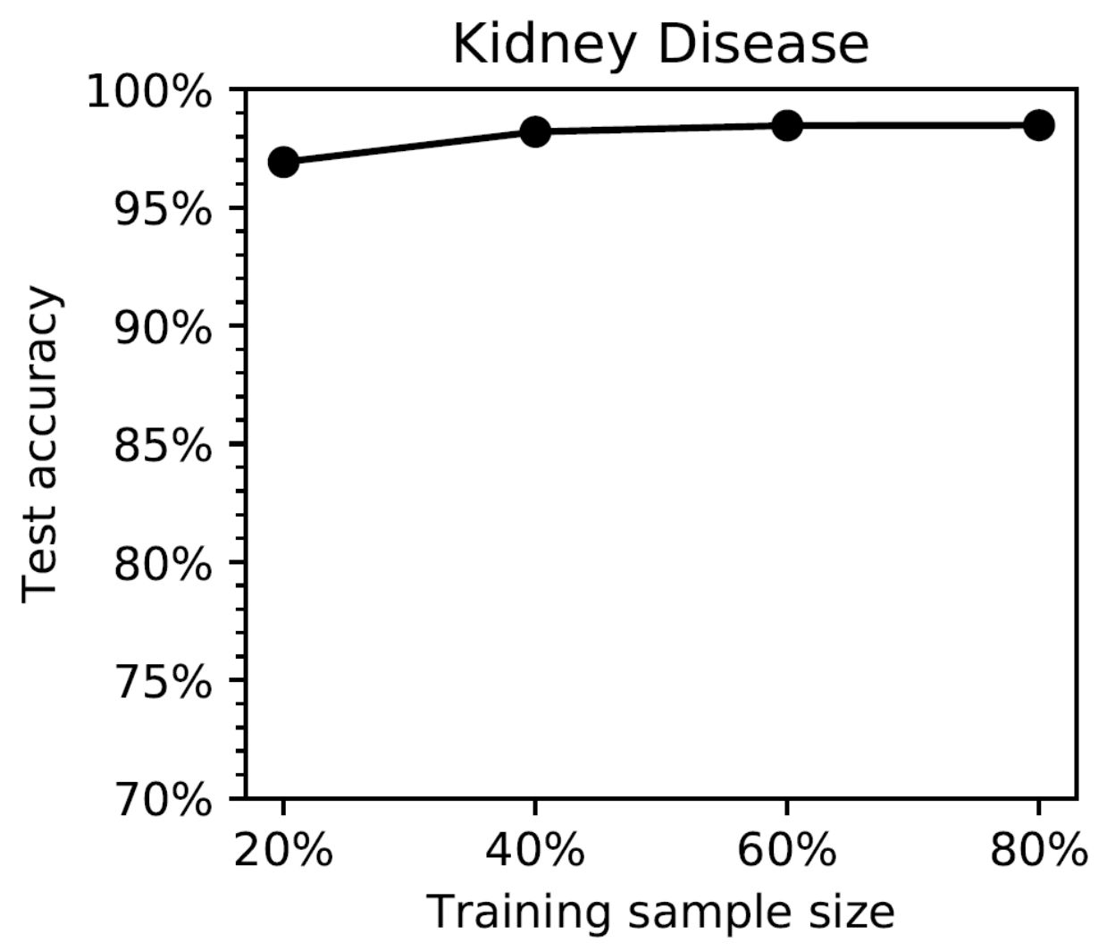


Results

The data sets were published on [3].

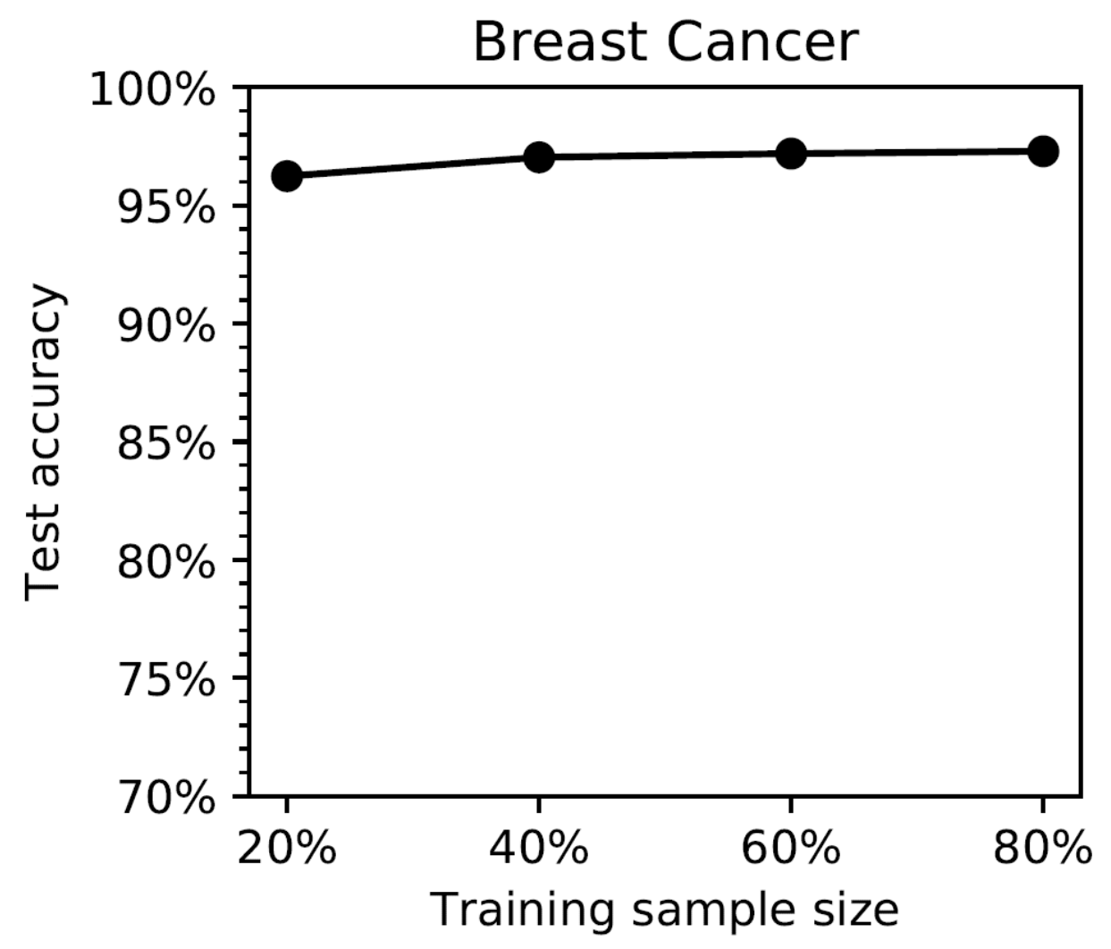
❖ Kidney disease

❑ **Features:** 25 features, including age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anemia.
❑ **Target:** Binary (chronic kidney disease, non chronic kidney disease).
❑ **Sample size:** 372



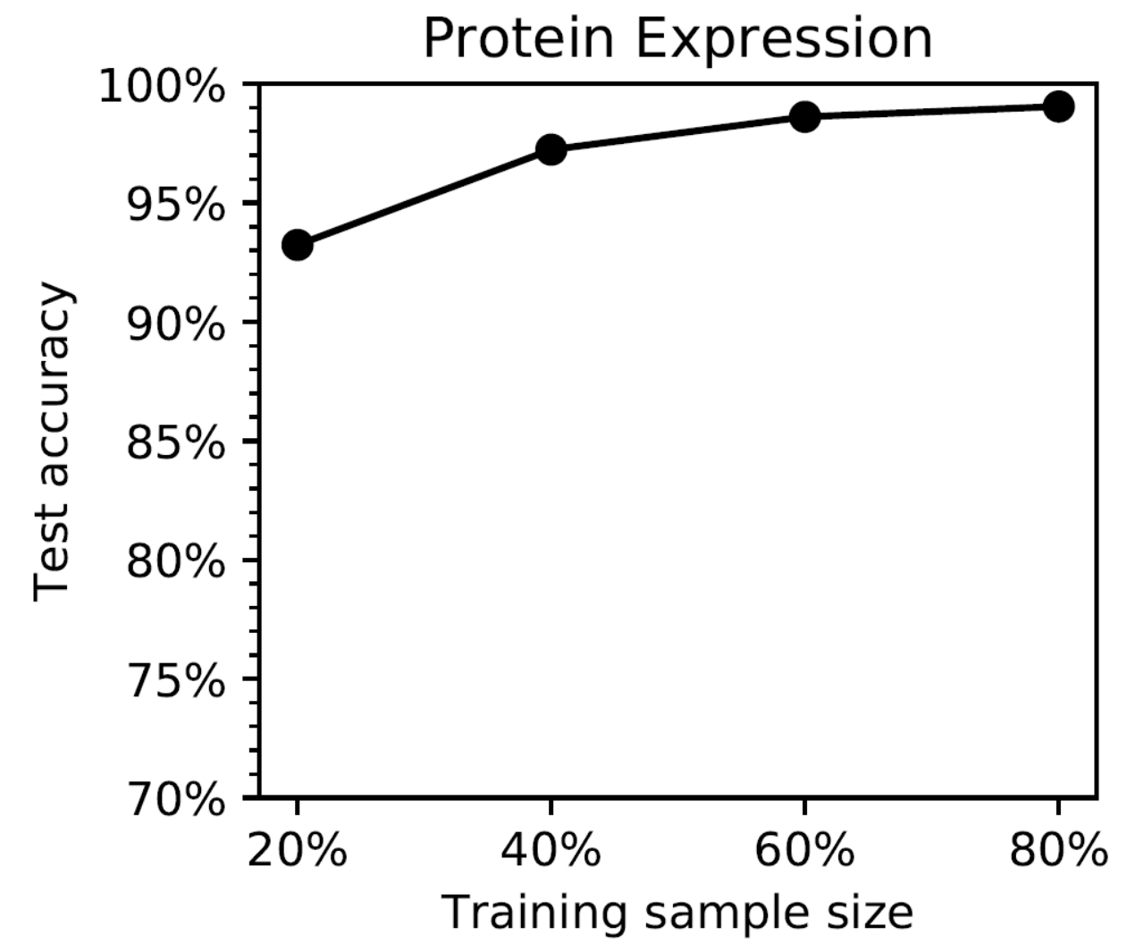
❖ Breast Cancer

❑ **Features:** 30 features describe characteristics of the snake-generated cell nuclei boundaries which were measured from a digitized image of a needle aspirate of a breast mass.
❑ **Target:** Binary (benign, malignant).
❑ **Sample size:** 569



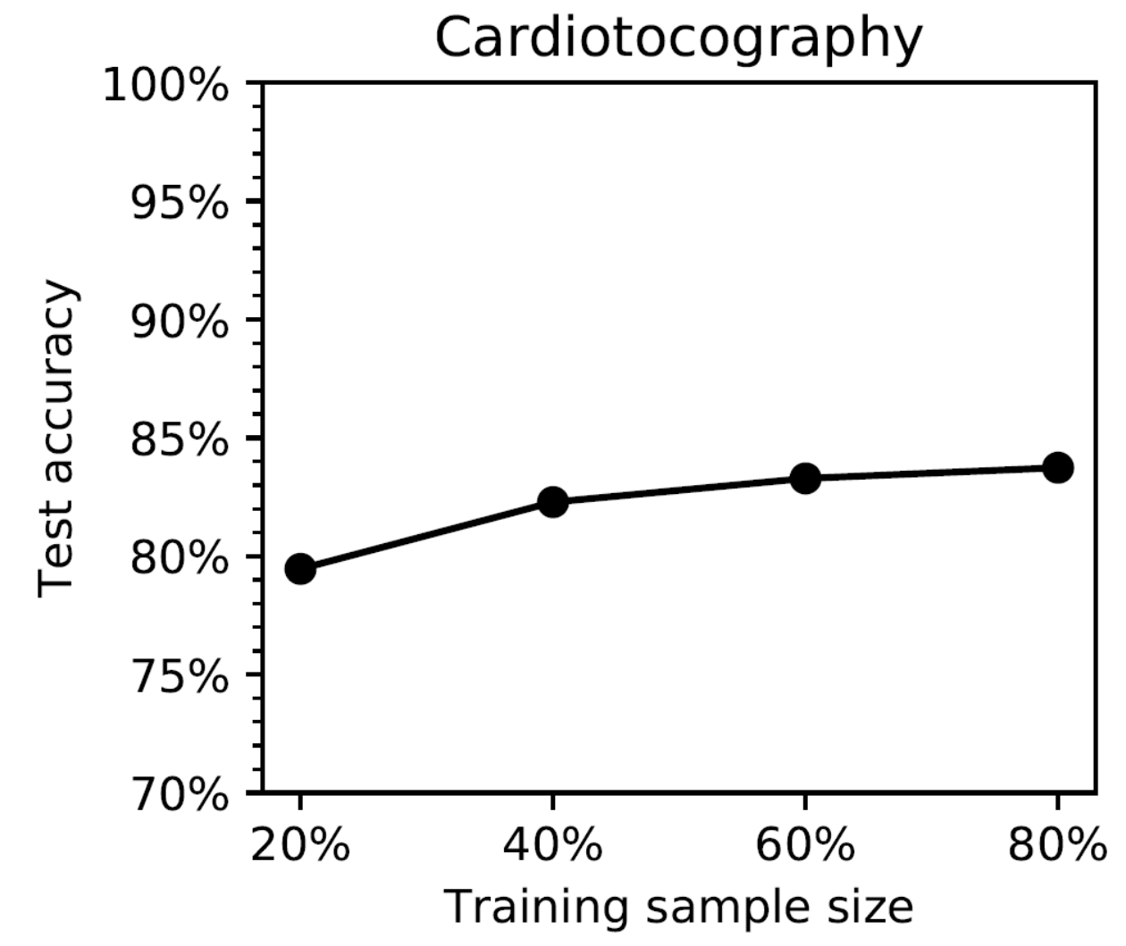
❖ Protein expression

❑ **Features:** 77 features, expression levels of 77 proteins that produced detectable signals in the nuclear fraction of cortex.
❑ **Target:** 8 classes, combination of genotype (control or trisomic), behavior (stimulated to learn or not), and treatment (injected with the drug or not).
❑ **Sample size:** 1080.



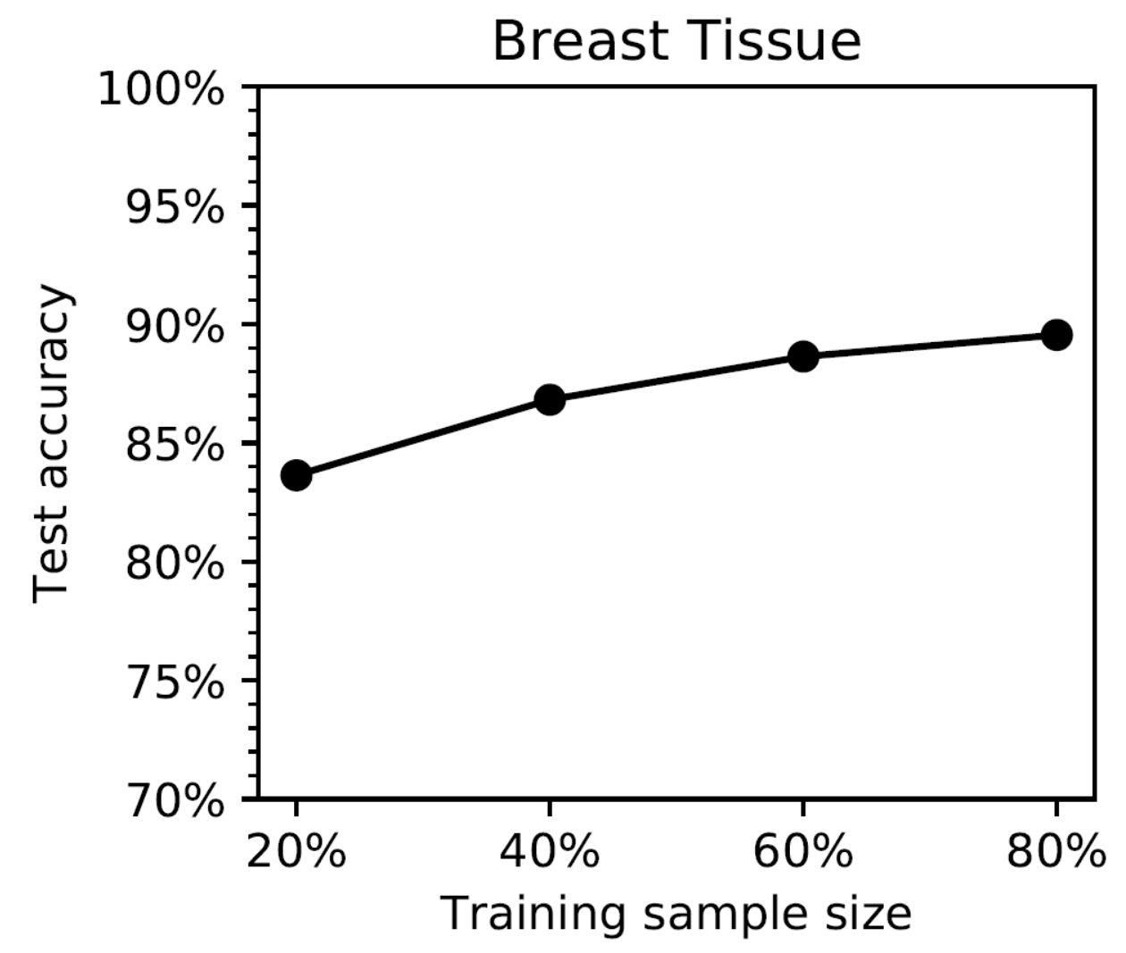
❖ Cardiotocography

❑ **Features:** 23 features measuring fetal heart rate and uterine contraction.
❑ **Target:** 3 classes, fetal state class code (normal, suspect, pathologic).
❑ **Sample size:** 528.



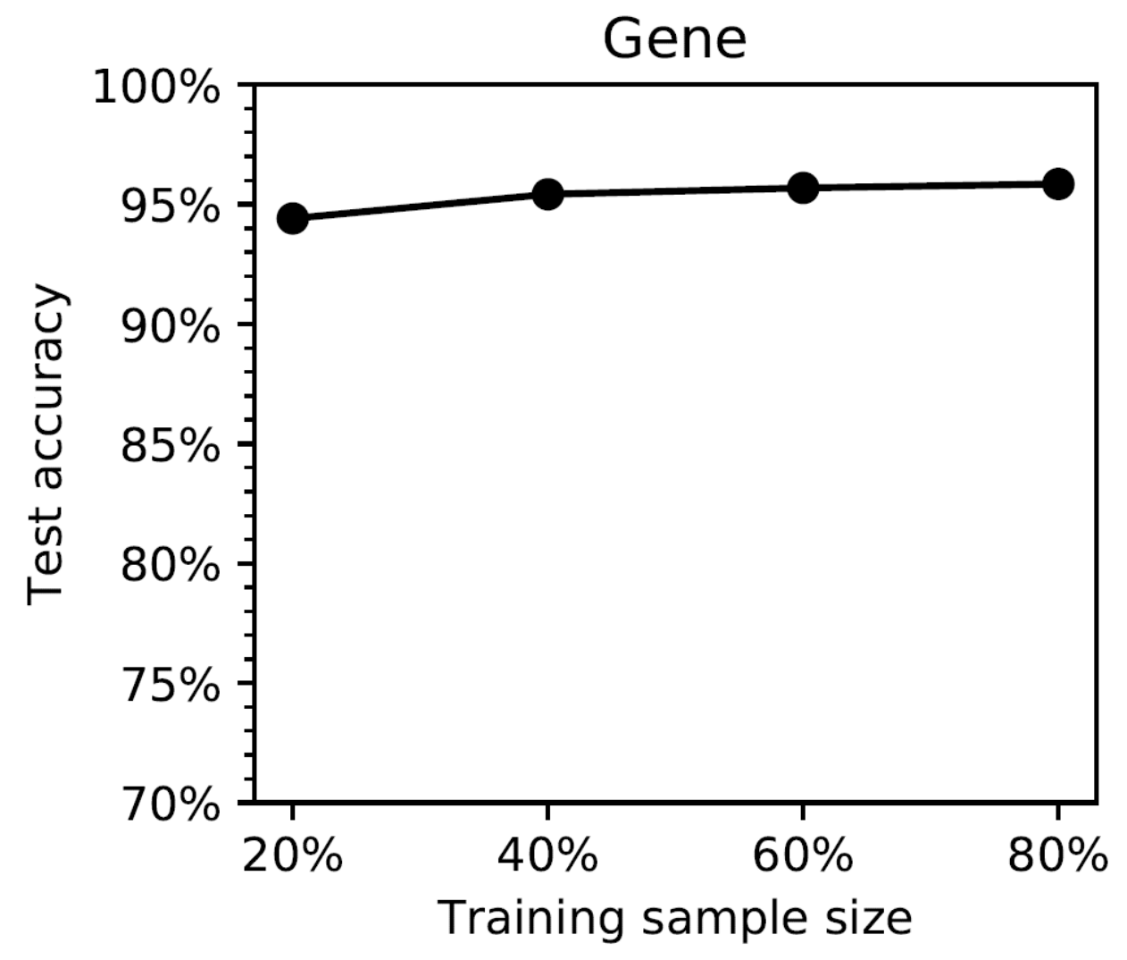
❖ Breast tissue:

❑ **Features:** 10 features, including electrical impedance measurements of freshly excised tissue samples from the breast.
❑ **Target:** 4 classes (carcinoma, connective, adipose, and merging class of bro-adenoma, mastopathy, glandular).
❑ **Sample size:** 106.



❖ Gene

❑ **Features:** 60 features of gene sequences.
❑ **Target:** 3 classes, exon-intron boundary (acceptor), intron-exon boundary (donor), and neither.
❑ **Sample size:** 3175



Conclusions

We extended ER to classification problems in which the features are not limited to binary variables but can be continuous or categorical variables, the target can be binary variable (binary classification) or categorical variable (multiclass classification). We demonstrated the performance of ER in 6 different biomedical data sets.

Acknowledgement: This work was supported by the Intramural Research Program of the National Institutes of Health, NIDDK.

References:

[1] D. T. Hoang, J. Song, V. Periwal, and J. Jo (2019), Network inference in stochastic systems from neurons to currencies: Improved performance at small sample size, Physical Review E, **99**, 023311.
[2] D. T. Hoang, J. Jo, and V. Periwal (2019), Data-driven inference of hidden nodes in networks, Physical Review E, **99**, 042114.
[3] https://archive.ics.uci.edu/ml/index.php