A Personalized Committee Classification Approach to Improving Prediction of Breast Cancer Metastasis – Supplementary Materials

Md Jamiul Jahid, Tim H. Huang, and Jianhua Ruan

February 13, 2014

I: Supplementary tables

Table S1: Cohort Information

Cohort	Size	# metastatic	# lymph node negative (pN-)	Adjuvant treatment received
NKI	295	78 (26.4%)	151 (51.2%)	10 (6.6%) pN- patients and 120 (83.3%) pN+ patients received adjuvant systemic therapy.
Wang	286	106 (37.1%)	286 (100%)	Patients had not received any adjuvant treatment.
UNC	116	75 (64.7%)	44 (37.9%)	Detailed treatment information was not available. However according to the reference, the cohort represents a heterogeneously treated group of patients compared to the NKI cohort.

Table S2: Relationship between σ values and the number of selected patients per base classifier.

Serial #	Cutoff (σ)	Median # of patients	Mean # of patients
1	0.001	68	68.3
2	0.0005	106	101.6
3	0.0002	168	156.7
4	0.0001	212	198.1
5	0.00005	242	231.8

Table S3: Performance of PC classifiers and other ensemble classifiers on NKI dataset.

See seperate excel file.

Table S4: Performance of PC classifiers and other ensemble classifiers on Wang dataset.

See seperate excel file.

Table S5: Performance of PC classifiers and other ensemble classifiers on UNC dataset.

See seperate excel file.

Table S6: Comparison of AUC scores of Dagging, AdaBoost, and PC classifiers

Classifiers	Datasets		
	NKI	Wang	UNC
PC-classifier	0.78	0.68	0.81
Dagging	0.72	0.61	0.75
AdaBoost	0.66	0.55	0.60

Table S7: Pathway enrichment scores

Group A genes		Group B genes		Group C genes	
Pathway	P-value	Pathway	P-value	Pathway	P-value
immune response	1.20E-04	regulation of cell proliferation	2.63E-07	response to estrogen stimulus	2.88E-10
response to nutrient	8.58E-04	regulation of cell migration	7.84E-06	response to wounding	2.9E-08
defense response	0.0019	enzyme linked receptor protein signaling pathway	1.01E-04	cell proliferation	4.39E-07
organic anion transport	0.0022	defense response	4.48E-04	response to hypoxia	1.24E-05
response to wounding	0.0051	cell communication	6.42E-04	cell adhesion	3.61E-05
epithelial cell differentiation	0.0051	response to wounding	9.93E-04	defense response	4.92E-05
cell death	0.0213	regulation of hormone secretion	9.81E-04	cell-cell signaling	3.1E-04
		cell adhesion	0.0025	immune response activity	4.9E-04
		cell cycle	0.0053	anti-apoptosis	5.14E-04
				cell cycle	0.0025

Table S8: Rank of genes in different classification models learned from NKI dataset.

See seperate excel file.

II: Supplementary figures

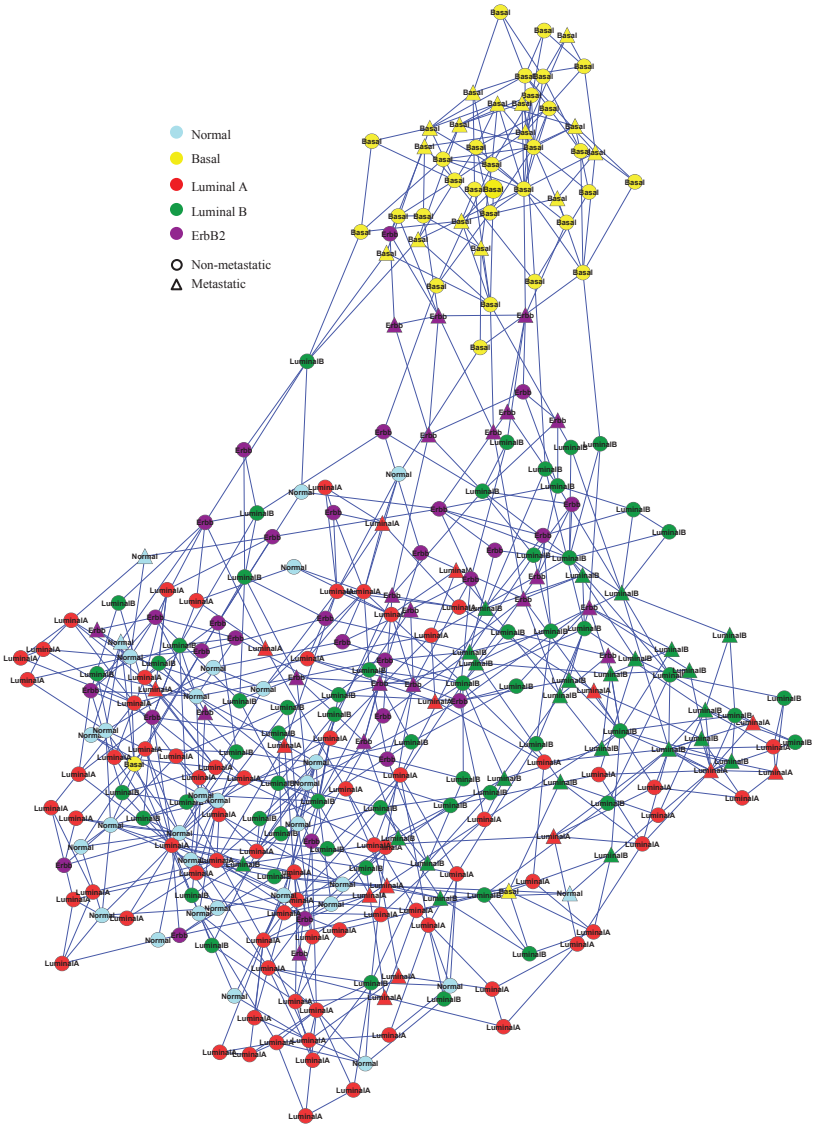

Fig. S1: Patient-patient network for NKI dataset. Triangle shaped nodes represent metastatic patients.

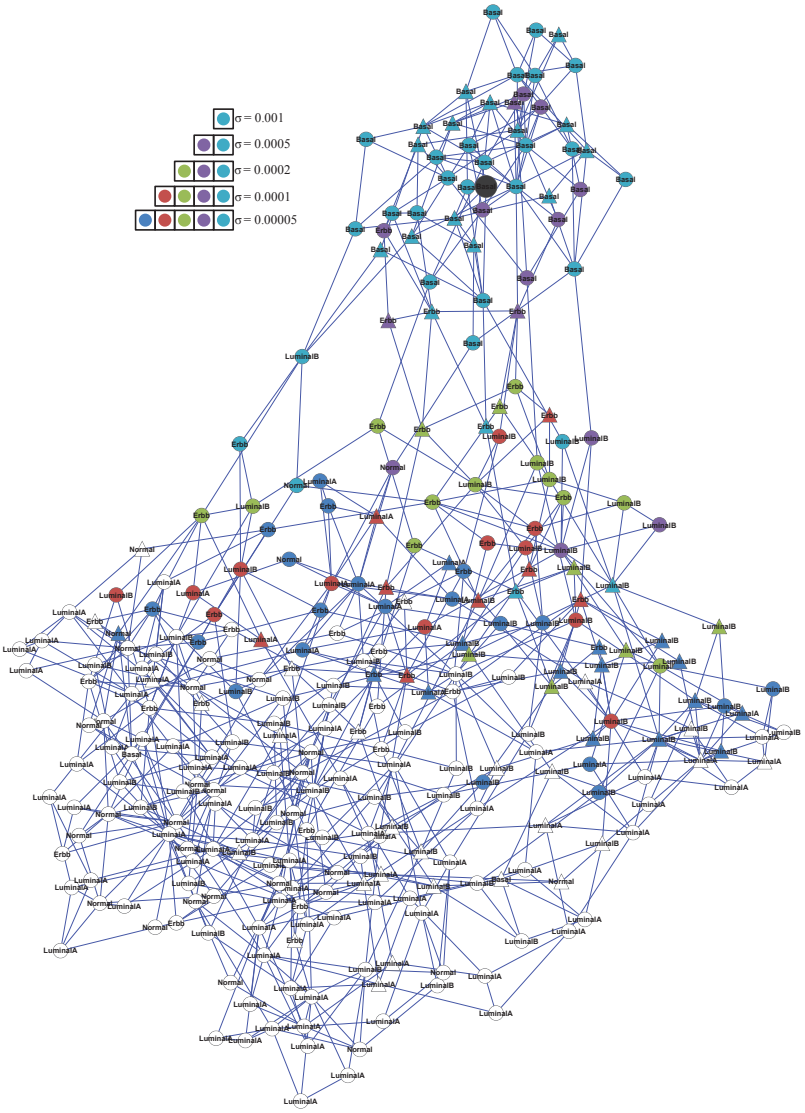

Fig. S2: Training patients selected for one basal patient (marked in black) on patient-patient network with different values of σ. Triangle shaped node represent metastatic patients.

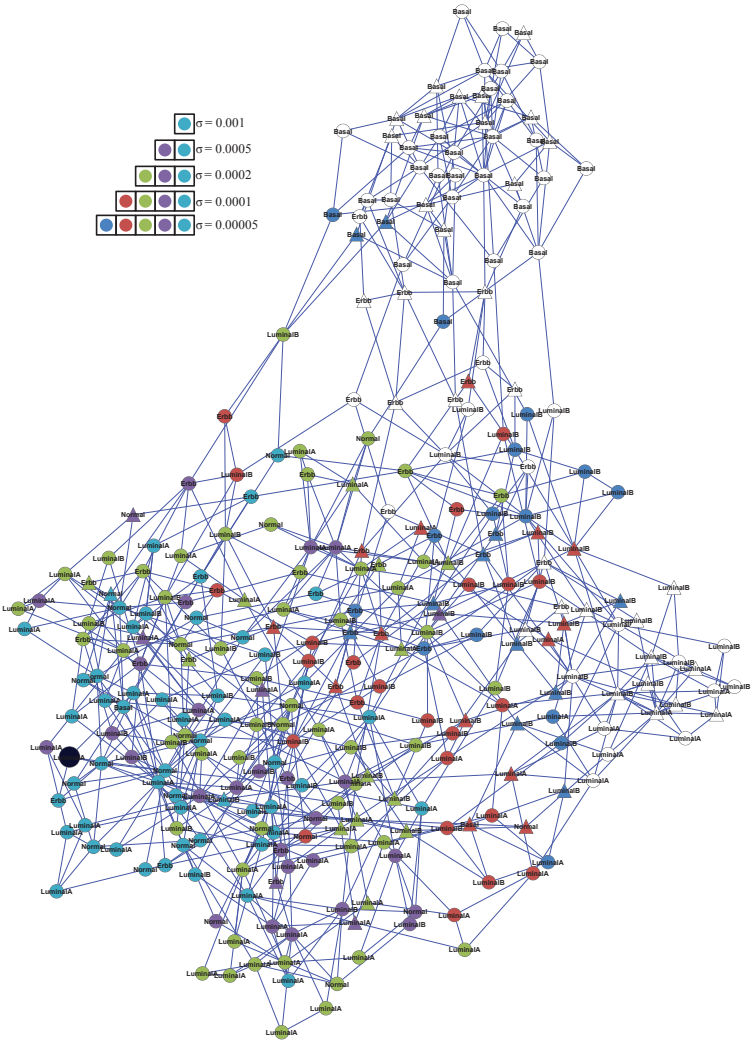

Fig. S3: Training patients selected for one luminalA patient (marked in black) on patient-patient network with different values of σ. Triangle shaped node represent metastatic patients.

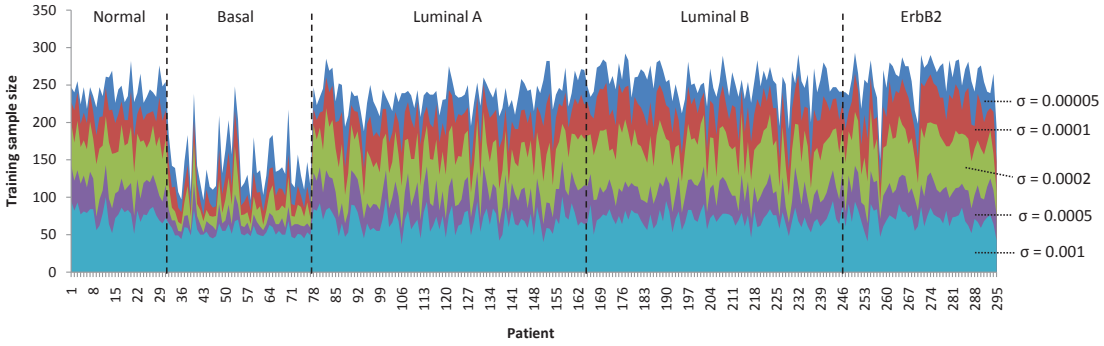

Fig. S4: Number of training sample patients used for each patient in NKI dataset with different cutoffs.

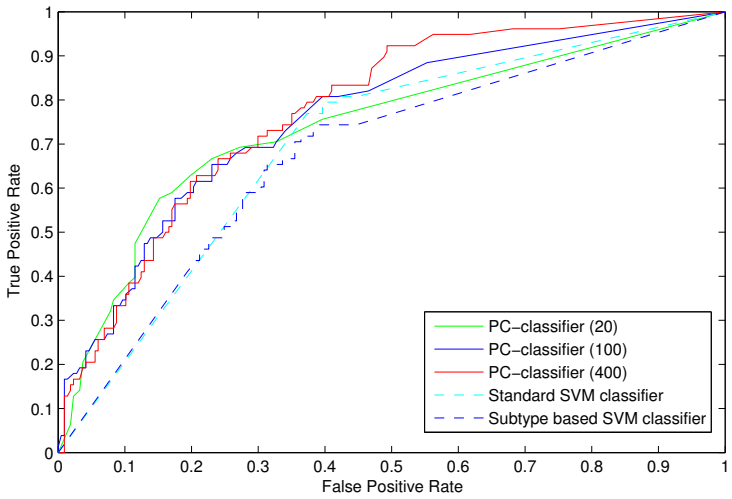

Fig. S5: ROC curve for NKI dataset. Number in parentheses represents number of base classifiers used.

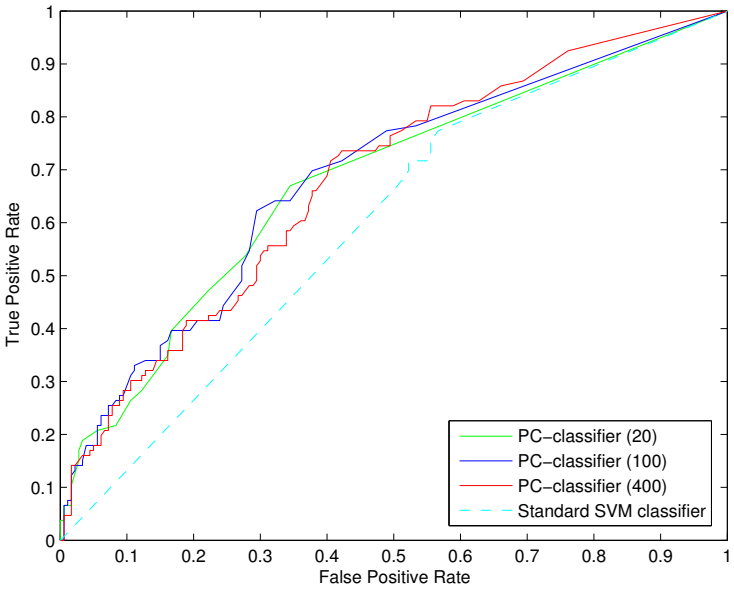

Fig. S6: ROC curve for Wang dataset with models from NKI dataset. Number in parentheses represents number of base classifiers used.

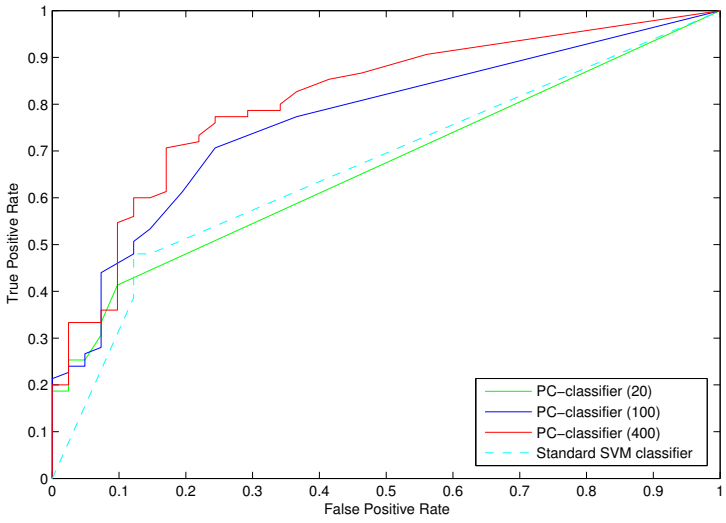

Fig. S7: ROC curve for UNC dataset with models from NKI dataset. Number in parentheses represents number of base classifiers used.

III: Rationale and supporting data for parameter selection

Relationship between c, σ, and the number of selected training samples

The two parameters c and σ are closely related. The random walk parameter c determines how much influence a node has on its neighboring nodes - a larger c gives a random walker higher probability to return to the starting node, which reduces the probability for the random walker to reach other nodes. On the other hand, the parameter σ is a cutoff used to determine which node should be included as a neighbor node of a starting node. Therefore, with a larger c, a smaller σ is needed in order to keep the same number of neighboring nodes, and vice versa.

In this work, we did not attempt to optimize the two parameters separately or extensively so as to maximize the performance of our algorithm, because we believe such practice may give our algorithm an undesirable advantage against the competing algorithms. We fixed the value for c to be 0.5, a typical value for the random walk algorithm seen in the literature in the machine learning field. We then chose a set of values for parameter σ based on the following rationales: (1) To build an accurate base classifier for a patient, we would ideally like to choose only neighboring patients belonging to the same subtype and eliminate patients not belonging to the same subtype, (2) We need to have a sufficient number of patients to construct a robust base classifier without overfitting the model, and (3) To make our method more general, we assume no knowledge about the frequency of each subtype in the cohort, or even the number of subtypes for the disease.

Based on the above rationales, we selected a set of σ values to construct base classifiers with a wide range of the number of training samples to cover possible subtype sizes. Supplementary Table S2 shows the mean / median number of training samples in each base classifier. From the table it can be seen that the numbers of training samples selected with different cutoffs are roughly evenly spaced, ranging from 1/4 to about 80% of the patients in the training cohort. The lower bound is chosen to ensure that each classifier was trained on a sufficient number of training samples and contains subtype-specific information, and the upper bound is chosen to ensure that different base classifiers were constructed using a different set of training samples by eliminating some of the least similar patients. The latter may become important when the disease studied do not have subtypes, but instead contain a small fraction of outliers that are not similar to the rest of

Table S9: Cutoff settings for three 3-cutoff PCC models.

Setting	Cutoffs	Comment
1	σ_1, σ_2 and σ_3	Contain most subtype-specific training samples where most of the training samples are within the local neighborhood of the representing patients in the patient -patient network
2	σ_1, σ_3 and σ_5	Contain less subtype-specific training samples than setup 1
3	σ_3, σ_4 and σ_5	Contains least subtype specific training samples

the patients.

Also note that even with the same cutoff, the number of neighboring patients is not a constant for different patients, and reflects the size of the corresponding subtype to some extent (Supplementary Figure S4). This demonstrates that the random walk procedure on the patient-patient network can automatically discover the different subtypes in the cohort to some extent.

Why 5 base classifiers for each training sample?

As mentioned in the previous subsection, we choose the five cutoffs of σ so that the numbers of training samples per classifiers are spaced roughly evenly in a reasonable range (Supplementary Table S2).

In general, we keep the total number of classifiers small, not only to increase the efficiency of the algorithm, but also to ensure that the classifiers are sufficiently different from one another. When two base classifiers were built with similar sets of training patients, it is expected that the two classifiers may behave similarly. For example, a classifier built using a patient's nearest 100 neighbors may share a high similarity with another classifier built using the same patient's nearest 120 neighbors, because the former was built using a subset of the training samples from the latter.

To see if fewer classifiers can improve the result, we tested several modified models with only three base classifiers per patient, using the cutoff settings shown in Supplementary Table S9. We first used cutoffs $\sigma_1 = 0.001$, $\sigma_3 = 0.0002$, and $\sigma_5 = 0.00005$, which still covers the whole spectrum of the dataset, but at a lower resolution. We found that this modification does not affect the classification performance significantly. As shown in Supple-

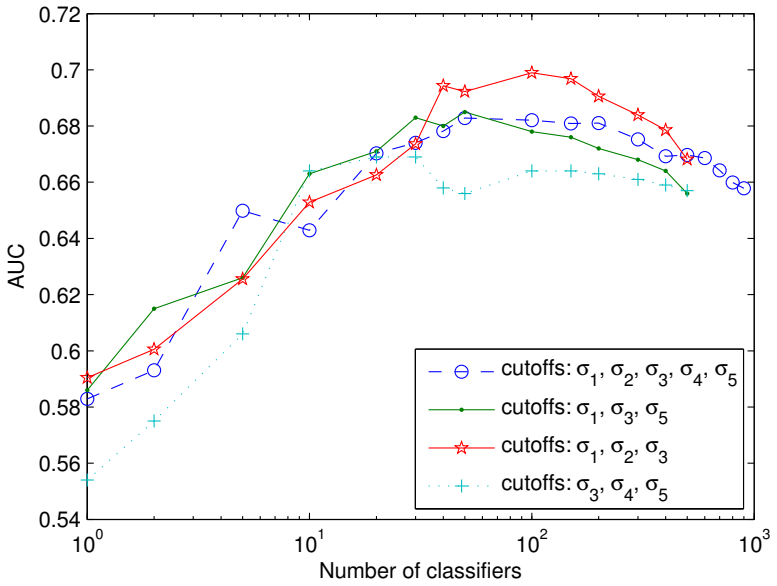

Fig. S8: Comparison between 5-cutoff PCC models and three 3-cutoff PCC models on Wang dataset.

mentary Figure S8, this modified model and the originally proposed model have very similar performance on the Wang dataset, both with the highest AUC at 0.68. Similar results were found on the other two datasets, with the 3-cutoff model performs slightly worse (data not shown). Note that compared to the 5-cutoff model, the 3-cutoff model actually achieves its best performance with a fewer number of classifiers, which is expected because of the smaller total number of base classifiers.

We then tested whether subtype-specific training samples are more important to predict metastatic than other samples. To this end, we used additionally two sets of cutoffs, setting #1 ($\sigma_1 = 0.001$, $\sigma_2 = 0.0005$, and $\sigma_3 = 0.0002$), or setting #3 ($\sigma_3 = 0.0002$, $\sigma_4 = 0.0001$, and $\sigma_3 = 0.00005$), where setting #1 includes the most subtype-specific training samples and setting #3 includes the least (Supplementary Table S2).

We found that on Wang dataset setting #1 has the most superior performance compared to the 5-cutoff models and the other 3-cutoff models, and setting #3 had the worst performance (Supplementary Figure S8). On the other hand, in UNC, the 3-cutoff PC-classifier with setting #1 has almost the same accuracy as the 5-cutoff PC-classifier, and in NKI the 3-cutoff PC-classifier performed slightly worse than the 5-cutoff PC-classifier (but still better than the other ensemble classifiers) (Supplementary Figure S9, S10). Therefore, while it is very likely that a more careful selection of σ values and the number of base classifiers per patient could potentially improve the results of PC-classifier further, such improvement might be dataset depen-

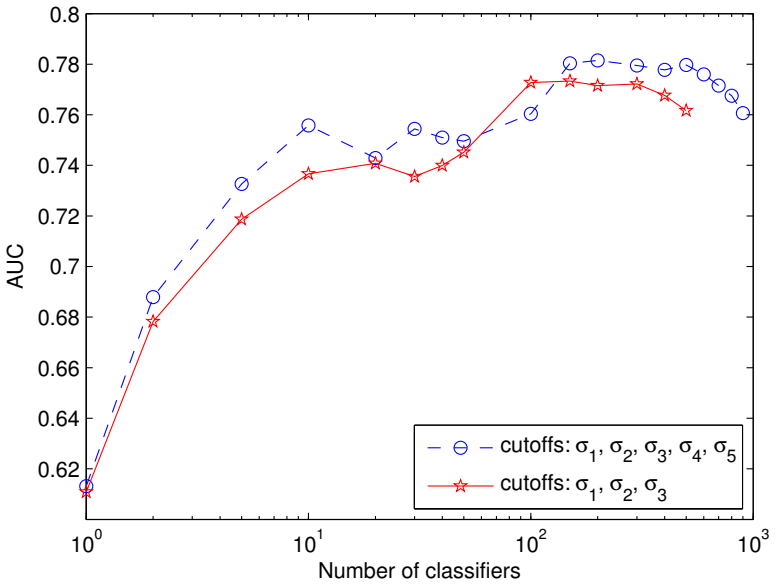

Fig. S9: Comparison between 3-cutoff and 5-cutoff PCC models on NKI dataset.

dent. In the manuscript we only report the classification results for the 5-cutoff models, which we believe are more robust and general, and leave disease/dataset-specific fine tuning of parameters to future studies.

It can be seen that in fact setting #1 has the most superior performance compared to the 5-cutoff models and the other 3-cutoff models (Supplementary Figure S8). On the other hand, setting #3 had the worst performance. While it is very likely that an even more careful selection of σ and the number of base classifiers per patient could improve the results of PC-classifiers further, such improvement might be dataset dependent. In fact, in UNC, the 3-cutoff PC-classifier with setting #1 has almost the same accuracy as the 5-cutoff PC-classifier, and in NKI the 3-cutoff PC-classifier performed slightly worse (but still better than the other ensemble classifiers) (Supplementary Figure S9, S10). Therefore, in the manuscript we only report the classification results for the 5-cutoff models, which we believe are more robust and general, and leave disease-specific fine tuning of parameters to future studies.

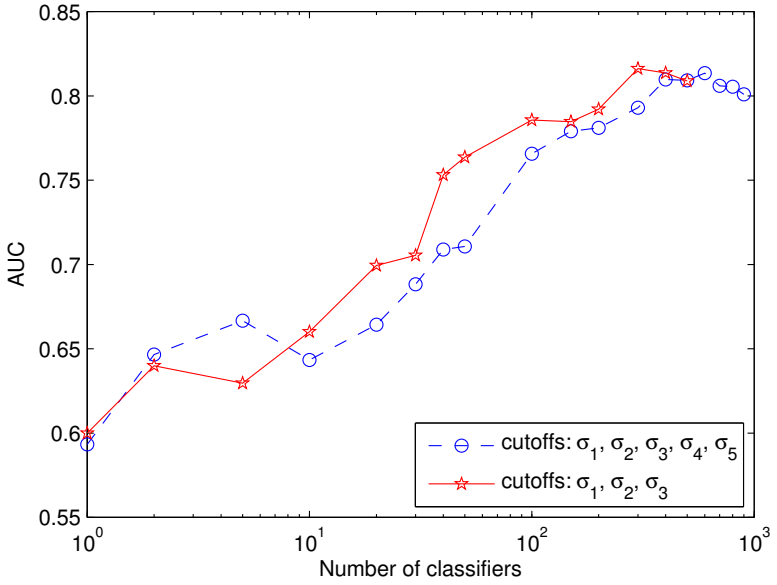

Fig. S10: Comparison between 3-cutoff and 5-cutoff PCC models on UNC dataset.