

(Online Appendix)

An Unbiased Comparison of Support Vector Machines and Random Forests for Microarray-Based Cancer Classification

Alexander Statnikov, MS, MS^{1,2}, Constantin F. Aliferis, MD, PhD^{1,2,3,4}

¹Discovery Systems Laboratory, ²Department of Biomedical Informatics, ³Department of Biostatistics, ⁴Department of Cancer Biology, Vanderbilt University, Nashville, TN, USA

1. Complete information about microarray datasets used in the study.

| Task | Dataset name | Number of classes | Number of variables (genes) | Number of samples | Diagnostic or outcome prediction task | Reference |
|-----------|----------------------|-------------------|-----------------------------|-------------------|--|-----------|
| Diagnosis | <i>Su</i> | 11 | 12533 | 174 | 11 various human tumor types | 1 |
| | <i>Ramaswamy</i> | 26 | 15009 | 308 | 14 various human tumor types and 12 normal tissue types | 2 |
| | <i>Staunton</i> | 9 | 5726 | 60 | 9 various human tumor types | 3 |
| | <i>Pomeroy</i> | 5 | 5920 | 90 | 5 human brain tumor types | 4 |
| | <i>Nutt</i> | 4 | 10367 | 50 | 4 malignant glioma types | 5 |
| | <i>Golub</i> | 3 | 5327 | 72 | Acute myelogenous leukemia (AML), acute lymphoblastic leukemia (ALL) B-cell and ALL T-cell | 6 |
| | <i>Armstrong</i> | 3 | 11225 | 72 | AML, ALL and mixed-lineage leukemia (MLL) | 7 |
| | <i>Bhattacharjee</i> | 5 | 12600 | 203 | 4 lung cancer types and normal tissues | 8 |
| | <i>Khan</i> | 4 | 2308 | 83 | Small, round blue cell tumors (SRBCT) of childhood | 9 |
| | <i>Shipp</i> | 2 | 5469 | 77 | Diffuse large B-cell lymphomas (DLBCL) and follicular lymphomas | 10 |
| | <i>Singh</i> | 2 | 10509 | 102 | Prostate tumor and normal tissues | 11 |
| Prognosis | <i>Iizuka</i> | 2 | 7070 | 60 | Hepatocellular carcinoma 1-year recurrence-free survival | 12 |
| | <i>Beer</i> | 2 | 7129 | 86 | Lung adenocarcinoma survival | 13 |
| | <i>Veer</i> | 2 | 24188 | 97 | Breast cancer 5-year metastasis-free survival | 14 |
| | <i>Rosenwald</i> | 2 | 7399 | 240 | Non-Hodgkin lymphoma survival | 15 |
| | <i>Yeoh</i> | 2 | 12240 | 233 | Acute lymphocytic leukaemia relapse-free survival | 16 |
| | <i>Pomeroy</i> | 2 | 7129 | 60 | Medulloblastoma survival | 4 |
| | <i>Bhattacharjee</i> | 2 | 12600 | 62 | Lung adenocarcinoma 4-year survival | 8 |

All diagnostic microarray datasets were downloaded from <http://www.gems-system.org> and all prognostic datasets were obtained from the links given in ¹⁷.

2. Detailed classification performance results *without gene selection*. The reported measure of performance is area under ROC curve (AUC)^{18,19} for binary datasets and relative classifier information (RCI)²⁰ for multicategory datasets. All performance estimates were obtained by 10-fold cross-validation²¹. See text for details.

| Task | Dataset | SVM | RF (ntree = 500 mtryFactor = 1) | RF (ntree = 1000 mtryFactor = 1) | RF (ntree = 2000 mtryFactor = 1) |
|-----------|----------------------|--------|------------------------------------|-------------------------------------|-------------------------------------|
| Diagnosis | <i>Su</i> | 0.9580 | 0.9092 | 0.8899 | 0.9002 |
| | <i>Ramaswany</i> | 0.9053 | 0.8480 | 0.8600 | 0.8614 |
| | <i>Staunton</i> | 0.7700 | 0.7853 | 0.7837 | 0.8053 |
| | <i>Pomeroy</i> | 0.8231 | 0.5379 | 0.6112 | 0.5630 |
| | <i>Nutt</i> | 0.7749 | 0.7329 | 0.7504 | 0.7580 |
| | <i>Golub</i> | 0.9390 | 0.8589 | 0.8675 | 0.9049 |
| | <i>Armstrong</i> | 0.9442 | 0.9197 | 0.9197 | 0.9197 |
| | <i>Bhattacharjee</i> | 0.8945 | 0.7543 | 0.7543 | 0.7543 |
| | <i>Khan</i> | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | <i>Shipp</i> | 0.9917 | 0.9650 | 0.9733 | 0.9733 |
| | <i>Singh</i> | 0.9640 | 0.9480 | 0.9400 | 0.9440 |
| Prognosis | <i>Iizuka</i> | 0.6625 | 0.6875 | 0.7750 | 0.7750 |
| | <i>Beer</i> | 0.7980 | 0.6135 | 0.6639 | 0.6476 |
| | <i>Veer</i> | 0.7470 | 0.7672 | 0.7703 | 0.7733 |
| | <i>Rosenwald</i> | 0.6891 | 0.6523 | 0.6201 | 0.6391 |
| | <i>Yeoh</i> | 0.7766 | 0.6724 | 0.6513 | 0.6688 |
| | <i>Pomeroy</i> | 0.6917 | 0.6250 | 0.5167 | 0.5917 |
| | <i>Bhattacharjee</i> | 0.5194 | 0.5611 | 0.6139 | 0.5583 |

| Task | Dataset | RF (ntree = 500 mtryFactor = 0.5) | RF (ntree = 1000 mtryFactor = 0.5) | RF (ntree = 2000 mtryFactor = 0.5) | RF (ntree = 500 mtryFactor = 2) |
|-----------|----------------------|--------------------------------------|---------------------------------------|---------------------------------------|------------------------------------|
| Diagnosis | <i>Su</i> | 0.8878 | 0.9122 | 0.8986 | 0.9058 |
| | <i>Ramaswany</i> | 0.8421 | 0.8397 | 0.8408 | 0.8629 |
| | <i>Staunton</i> | 0.7163 | 0.7365 | 0.7838 | 0.7908 |
| | <i>Pomeroy</i> | 0.5290 | 0.5290 | 0.5290 | 0.6112 |
| | <i>Nutt</i> | 0.7580 | 0.7241 | 0.7253 | 0.7787 |
| | <i>Golub</i> | 0.8589 | 0.8272 | 0.8304 | 0.9160 |
| | <i>Armstrong</i> | 0.9197 | 0.8944 | 0.9197 | 0.9449 |
| | <i>Bhattacharjee</i> | 0.7176 | 0.7130 | 0.7311 | 0.7543 |
| | <i>Khan</i> | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | <i>Shipp</i> | 0.9650 | 0.9733 | 0.9650 | 0.9733 |
| | <i>Singh</i> | 0.9280 | 0.9440 | 0.9440 | 0.9440 |
| Prognosis | <i>Iizuka</i> | 0.7500 | 0.8000 | 0.8125 | 0.7625 |
| | <i>Beer</i> | 0.6397 | 0.6802 | 0.6679 | 0.6730 |
| | <i>Veer</i> | 0.7795 | 0.7623 | 0.7682 | 0.7672 |
| | <i>Rosenwald</i> | 0.6316 | 0.6406 | 0.6417 | 0.6418 |
| | <i>Yeoh</i> | 0.6605 | 0.6484 | 0.6470 | 0.6824 |
| | <i>Pomeroy</i> | 0.6167 | 0.5917 | 0.6083 | 0.5958 |
| | <i>Bhattacharjee</i> | 0.6056 | 0.5611 | 0.6028 | 0.5528 |

| <i>Task</i> | <i>Dataset</i> | RF (ntree = 1000 mtryFactor = 2) | RF (ntree = 2000 mtryFactor = 2) | RF (optimized by cross-val.) |
|-------------|----------------------|--|--|--|
| Diagnosis | <i>Su</i> | 0.9005 | 0.9001 | 0.9104 |
| | <i>Ramaswamy</i> | 0.8705 | 0.8705 | 0.8614 |
| | <i>Staunton</i> | 0.8516 | 0.8231 | 0.8187 |
| | <i>Pomeroy</i> | 0.6112 | 0.6112 | 0.6112 |
| | <i>Nutt</i> | 0.7497 | 0.7580 | 0.7329 |
| | <i>Golub</i> | 0.9049 | 0.9160 | 0.9337 |
| | <i>Armstrong</i> | 0.9197 | 0.9197 | 0.8944 |
| | <i>Bhattacharjee</i> | 0.7627 | 0.7627 | 0.7627 |
| | <i>Khan</i> | 1.0000 | 1.0000 | 1.0000 |
| | <i>Shipp</i> | 0.9833 | 0.9833 | 0.9733 |
| | <i>Singh</i> | 0.9520 | 0.9520 | 0.9440 |
| Prognosis | <i>Iizuka</i> | 0.7750 | 0.7875 | 0.7625 |
| | <i>Beer</i> | 0.6726 | 0.6790 | 0.6456 |
| | <i>Veer</i> | 0.7703 | 0.7783 | 0.7542 |
| | <i>Rosenwald</i> | 0.6399 | 0.6418 | 0.6293 |
| | <i>Yeoh</i> | 0.6437 | 0.6599 | 0.6604 |
| | <i>Pomeroy</i> | 0.6292 | 0.5958 | 0.6000 |
| | <i>Bhattacharjee</i> | 0.5972 | 0.5806 | 0.5611 |

References

- 1 Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* 2001 Oct 15;61(20):7388-93.
- 2 Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 2001 Dec 18;98(26):15149-54.
- 3 Staunton JE, Slonim DK, Collier HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. *Proc Natl Acad Sci U S A* 2001 Sep 11;98(19):10787-92.
- 4 Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 2002;415(6870):436-42.
- 5 Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* 2003 Apr 1;63(7):1602-7.
- 6 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999 Oct 15;286(5439):531-7.
- 7 Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* 2002;30(1):41-7.
- 8 Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* 2001 Nov 20;98(24):13790-5.
- 9 Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001 Jun;7(6):673-9.

- 10 Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 2002 Jan;8(1):68-74.
- 11 Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002 Mar;1(2):203-9.
- 12 Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, Mori N, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet* 2003;361(9361):923-9.
- 13 Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002 Aug;8(8):816-24.
- 14 van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002 Jan 31;415(6871):530-6.
- 15 Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N Engl J Med* 2002 Jun 20;346(25):1937-47.
- 16 Yeoh EJ, Ross ME, Shurtleff SA, Williams WK, Patel D, Mahfouz R, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 2002 Mar;1(2):133-43.
- 17 Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005;365(9458):488-92.
- 18 Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Technical Report, HPL-2003-4, HP Laboratories 2003.
- 19 Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996 Feb 28;15(4):361-87.
- 20 Sindhvani V, Bhattacharyya P, Rakshit S. Information Theoretic Feature Crediting in Multiclass Support Vector Machines. Proceedings of the First SIAM International Conference on Data Mining 2001.
- 21 Weiss SM, Kulikowski CA. Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems. San Mateo, Calif: M. Kaufmann Publishers; 1991.