Published in final edited form as:

IEEE China Summit Int Conf Signal Inf Process. 2013 July ; 2013: 259–263. doi:10.1109/ChinaSIP.2013.6625340.

Exploration of Genomic, Proteomic, and Histopathological Image Data Integration Methods for Clinical Prediction

A. Poruthoor[2], **J.H. Phan**[2], **S. Kothari**[1], and **May D. Wang**[1,2]

[1]Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332

[2]Wallace H. Coulter department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, 30332

Abstract

The emergence of large multi-platform and multi-scale data repositories in biomedicine has enabled the exploration of data integration for holistic decision making. In this research, we investigate multi-modal genomic, proteomic, and histopathological image data integration for prediction of ovarian cancer clinical endpoints in The Cancer Genome Atlas (TCGA). Specifically, we study two data integration techniques, simple data concatenation and ensemble classification, to determine whether they can improve prediction of ovarian cancer grade or patient survival. Results indicate that integration via ensemble classification is more effective than simple data concatenation. We also highlight several key factors impacting data integration outcome such as predictability of endpoint, class prevalence, and unbalanced representation of features from different data modalities.

INTRODUCTION

Although ovarian cancer is the fifth leading cause of cancer death in American women, there is a lack of consistently successful treatment plans for this disease [1]. In order to gain a better understanding of the functionality of this cancer and the roots of variation in treatment response, researchers are increasingly turning to molecular-level analysis. Recent technological advances have resulted in large volumes of molecular-level data for many cancers, including ovarian cancer. Until now, however, researchers have mostly focused on homogeneous datasets containing information for just one kind of data (e.g., genomic or proteomic). Because multi-modal datasets can capture the holistic status of a patient, there is a growing need to study multi-modal data integration methods that can jointly use clinically relevant information for decision making [2].

We seek to explore the effects of multi-modal data integration on prediction of ovarian cancer endpoints in The Cancer Genome Atlas (TCGA). We first need to determine whether integration of genomic, proteomic, and histopathological imaging data can improve the prediction of ovarian cancer grade or patient survival. To investigate data integration

Correspondence, May D Wang, Ph.D., Biomedical Engineering Department, Georgia Institute of Technology & Emory University, 313 Ferst Drive, UA Whittakker Building Suite 4106, Atlanta, GA, 30332, Tel: 404-3852954, Fax: 404-894-4243, maywang@bme.gatech.edu.

methods, we compare the accuracy of survival and grade classification schemes based on (1) three homogenous datasets containing genomic, imaging, and proteomic data; (2) three bi-modal datasets containing genomic & imaging, genomic & proteomic, and imaging & proteomic data; and (3) one multi-modal dataset containing all three types of data, as shown in Figure 1. We also investigate an ensemble classification method that combines prediction models developed for each individual data modality into a single prediction result.

METHODS

2.1. Datasets

We used three datasets (genomic, proteomic, and imaging) containing ovarian cancer patient data obtained from The Cancer Genome Atlas (TCGA), as shown in Figure 1. Affymetrix gene expression data was processed using caCORRECT and log normalized [3]. All datasets were filtered by mutual patients and then by availability of survival and grade data. Genomic datasets were also filtered by mutual genes. Table 1 and Table 2 contain dataset feature and patient information.

2.2. Image Data Processing

We extracted features from histopathological whole-slide images (WSIs) after quality control steps including detection of tissue-fold and pen-mark artifacts from the lowest resolution WSIs [4]. To represent each patient using quantitative image features, we followed four steps. First, we cropped the highest-resolution WSI into 512 × 512 -pixel, non-overlapping tiles and selected tissue (excluding pen-mark and blank) tiles with less than 10% tissue-fold artifact. Second, from the tiles that passed quality control, we extracted 461 image features capturing various pixel- and object-level image features (Table 3) [4]. Third, we classified tiles into tumor and non-tumor tiles using a supervised classification model, trained using image features and manually annotated tumor and non-tumor tiles from 15 WSIs [4]. Finally, we combined image features from all tumor tiles of a patient. Since tile size is constant, we represented a WSI by simply averaging all pixel-level tile features except for Haralick and fractal features, where we summed co-occurrence matrices and histograms for tiles, respectively, to represent the WSI and then calculated the features (i.e., 13 Haralick features and eight fractal dimensions [5]). For combining object-level tile features, we assumed that objects in a tile were a subset of all objects in a WSI and combined features using group statistics accounting for the number of objects in each tile.

2.3. Data Integration

We considered four integrated datasets (Figure 1 and Table 1). Each integrated dataset is a concatenation of its component datasets. For example, the complete integrated dataset is a concatenation of the genomic, imaging, and proteomic datasets. The feature counts of each dataset after integration are summarized in Table 1.

2.4. Classification

We tested data integration methods using two binary classification endpoints, i.e., cancer grade and patient survival (Table 2). For the grade endpoint, we divided patients into grades 1 or 2 and grades 3 or 4 as the two classes. For the survival endpoint, we separated patients with survival known to be less than 5 years from those with survival known to be greater than or equal to 5 years.

Patients not known to be still living and with no follow-up after 5 years (i.e., patients that are normally censored) are grouped into the class with known surviving patients. Each endpoint was classified using eight different scenarios, outlined in Table 4. We used nested cross validation to estimate prediction performance for the first seven scenarios; and ensemble classification for the last scenario.

Nested Cross Validation—We estimated classification performance for each dataset using the SVM classifier and nested cross validation (Figure 2) [6]. Within the fifteen iterations of the outer loop, the entire dataset was randomly partitioned into training and testing sets. Thus, for each data combination scenario and endpoint, we (1) selected features from the training set using Minimum Redundancy Maximum Relevance (mRMR) [7], (2) trained the classifier using the selected features and training set samples, and (3) evaluated the classifier using testing set samples to produce 15 measures of prediction accuracy (i.e., one value for each iteration). The feature selection step requires one parameter, i.e., the number of features to be selected. We optimized this parameter within three iterations of the inner cross-validation loop. That is, we selected the feature size that produced the highest inner cross-validation accuracy. Within the inner cross-validation, we performed two-fold cross validation using mRMR feature selection and tested feature sizes ranging from 1 to 200. For each training and testing combination, we used rank normalization to normalize the testing data to the training data. The feature number that outputs the greatest accuracy on testing data is selected as optimal. Thus the inner loop outputs an optimal feature number.

Ensemble Classification—We used ensemble classification to estimate combined prediction performance using genomic, proteomic, and imaging data [8]. Specifically, we used nested cross validation as previously described, but simultaneously applied it to all three un-integrated datasets (i.e., the genomic, proteomic, and imaging datasets) to obtain three classification decisions for each sample. We then used voting to determine the final classification of each sample as follows. For each sample x_i, we obtained three decision values after training each classifier, $f(x)$: $f_{gene}(x_i)$, $f_{image}(x_i)$, and $f_{protein}(x_i)$. The output of each classifier function can be either 1 or −1, indicating grade or patient survival, depending on the endpoint in question. Thus, the final ensemble classification for sample x_i is the mode, or most frequently occurring class of the set $\{f_{gene}(x_i), f_{image}(x_i), f_{protein}(x_i)\}$.

2.5. Performance Evaluation

We used boxplots to compare the accuracies of the different classification scenarios. We first compared the prediction performance of the integrated datasets to that of their constituents to determine if data integration affects prediction performance. We used a one-tailed, two-sample t-test to determine if integrated datasets resulted in statistically significant improvements in prediction performance compared to individual datasets.

RESULTS

Boxplots of the prediction performances for each dataset are displayed in Figure 3, organized by data type and endpoint. Results from the t-tests are displayed in Table 5. As is evident in the figure, perhaps the most striking result is the drastic difference in accuracy between survival and grade classification. For survival, mean accuracies range from 0.51 to 0.55, while grade classification accuracies range from 0.83 to 0.88. Also interesting is that the standard deviation in accuracy values for survival (ranging from 2.5 to 4.3) are consistently higher than those for grade (ranging from 1.3 to 2.1), indicating greater variation in accuracy.

Comparison between different classification methods yields equally interesting results. For example, the imaging data produces the best results for one endpoint (grade) and the worst results for the other (survival). Other patterns of note include the consistent accuracy of proteomic and ensemble classification relative to the other methods and the superiority of ensemble accuracy over accuracy from the complete concatenated dataset. Also significant is that, in the integrated datasets produced through concatenation, classification accuracy is statistically similar to the accuracy of the genomic dataset, by far the largest dataset used.

DISCUSSION

Perhaps the most significant result of this research is that information integration via ensemble classification results in higher prediction performance compared to basic data concatenation. Ensemble classification removes the effect of dataset size (i.e., feature size) and reduces the effect of random errors via voting.

The difference in classification accuracy between the two endpoints can probably be explained by a combination of two factors. First, the low classification accuracy of patient survival can likely be attributed to the inherent difficulty of predicting survival; second, the relatively high classification accuracy of grade may be affected by the imbalance of classes in the grade data. These factors likely compounded to result in the large disparity in accuracy between the results for the two endpoints. In the future, evaluating results by AVC rather than accuracy would eliminate the effect of the large inter-class prevalence difference.

The consistent superiority of the proteomic data in classification, meanwhile, can likely be attributed to the smaller number of features present in the dataset relative to the number of samples. In the larger datasets, it is much more difficult to isolate informative features, which can decrease predictive accuracy, giving the proteomic dataset (the smallest dataset) an inherent advantage. It is likely that the superior performance of imaging data in grade classification can similarly be attributed to the size of the dataset. However, this may also be attributed to the fact that histopathological image data was originally used by pathologists to determine cancer grade. The feature size factor may also play a large role in the similarity in accuracy between the genomic and complete integrated datasets. That is, because there are so many genomic features, the genomic features represent nearly 95% of the total features in the complete datasets. This has implications on the proportion of genomic features used in final classification. To counteract some of these biasing effects of size an initial feature selection could be used to balance the proportion of features contributed by each data modality.

In the future, repeating this process with other novel and more biologically-relevant integration methods could yield more answers regarding the effect of integration on classification accuracy. Alternatively, using ensemble classification with different types of data might provide greater insight into the results presented in this paper. These results highlight some key factors to be considered when using integrated data for classification, i.e., predictability of endpoint, class prevalence, and unbalanced representation of features from different data modalities.

Acknowledgments

This research has been supported by grants from the National Institutes of Health (U54CA119338, 1RC2CA148265, and RO1CA163256), Georgia Cancer Coalition to Professor MDW, Hewlett Packard, and Microsoft Research.

REFERENCES

1. Bell D, Berchuck A, Birrer M, Chien J, Cramer D, Dao F, et al. Integrated genomic analyses of ovarian carcinoma. 2011
2. Phan J, Quo C, Cheng C, Wang M. Multi-scale integration of -omic, imaging, and clinical data in biomedical informatics. IEEE Rev Biomed Eng. 2012; 5:74–87. [PubMed: 23231990]
3. Moffitt R, Yin-Goen Q, Stokes T, Parry R, Torrance J, Phan J, et al. caCORRECT2: Improving the accuracy and reliability of microarray data in the presence of artifacts. BMC bioinformatics. 2011; 12:383. [PubMed: 21957981]
4. Kothari, S., Osunkoya, AO., Phan, JH., Wang, MD. Biological interpretation of morphological patterns in histopathological whole-slide images; Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine; 2012. p. 218-225.
5. Kothari S, Phan JH, Young AN, Wang MD. Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer. Proc IEEE Int Conf Bioinformatics Biomed. 2011:422–425.
6. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2:27.
7. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max- relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2005; 27:1226–1238.
8. Dietterich T. Ensemble methods in machine learning. Multiple classifier systems. 2000:1–15.

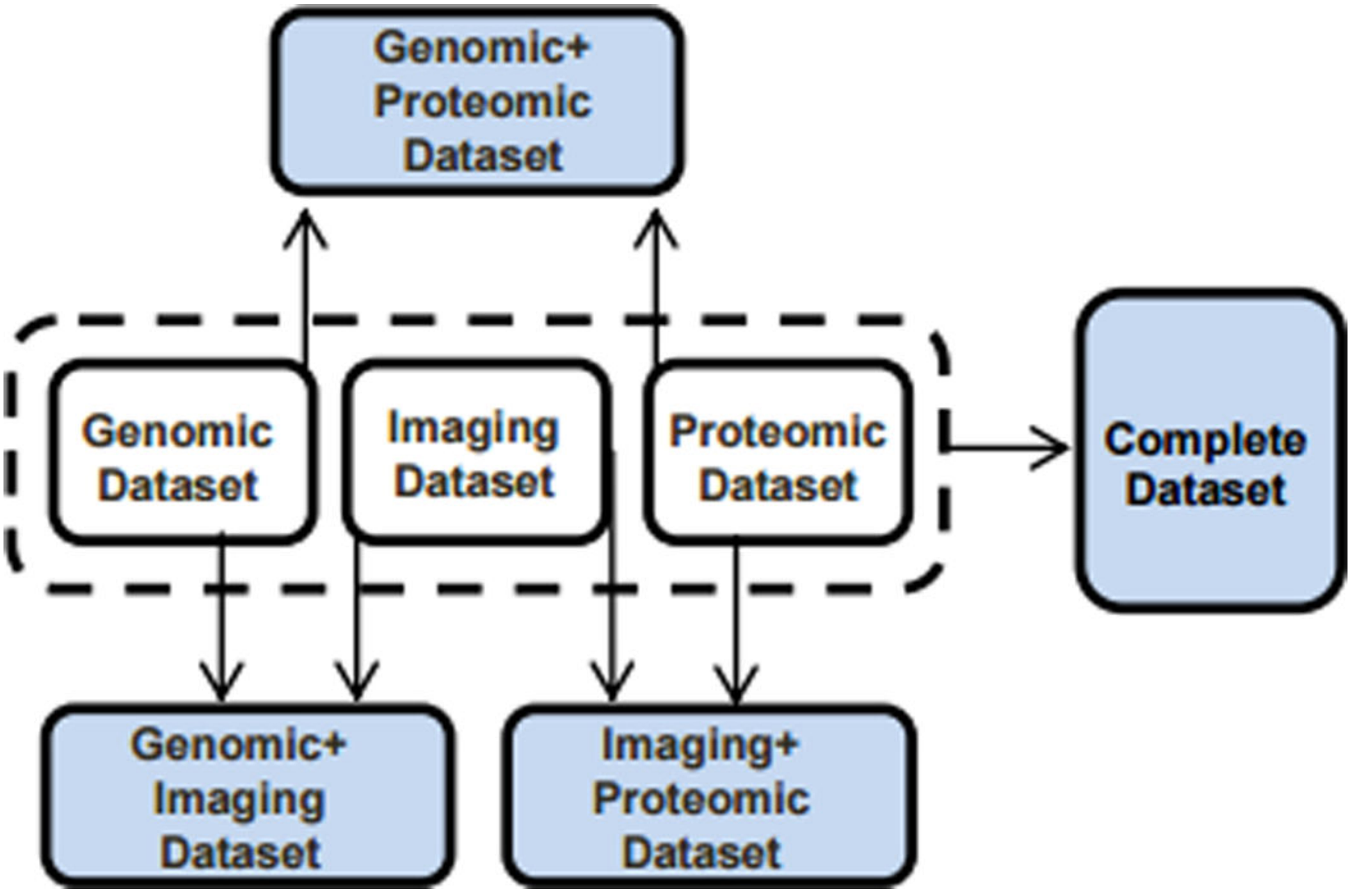

Figure 1.

Depiction of the flow of integration, indicating the component datasets (in the lighter boxes) of each integrated dataset (in the darker boxes).

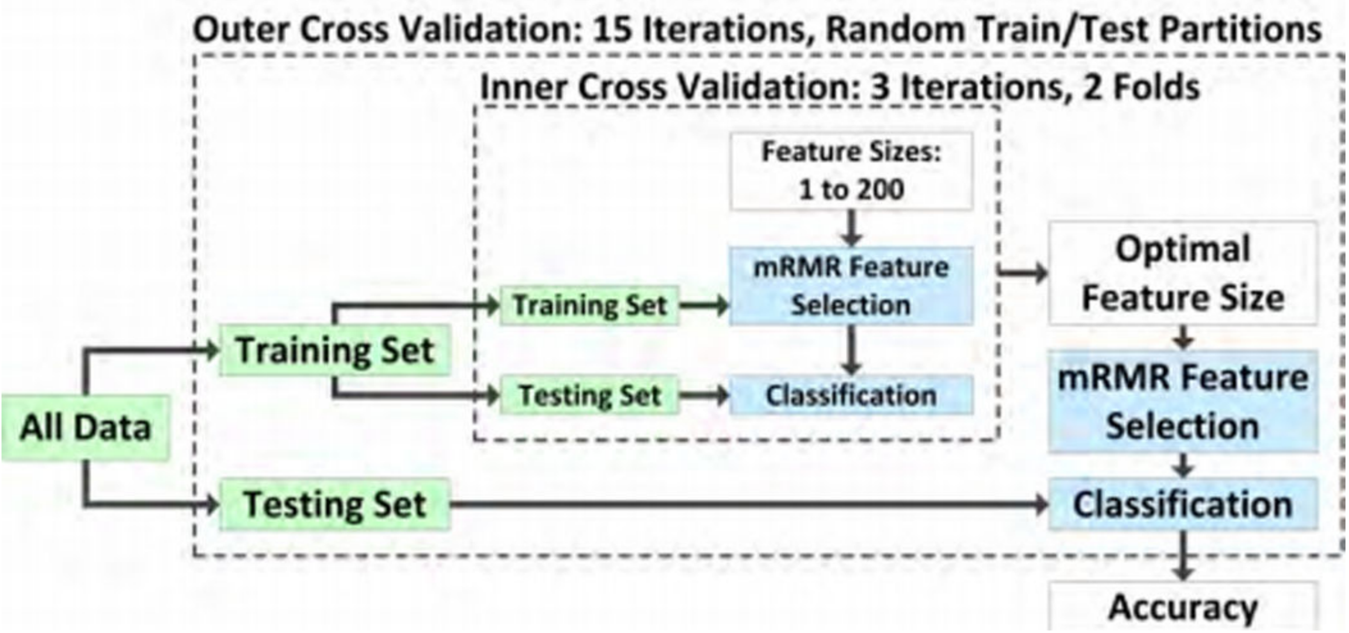

Figure 2.
Estimation of classification performance using nested cross validation.

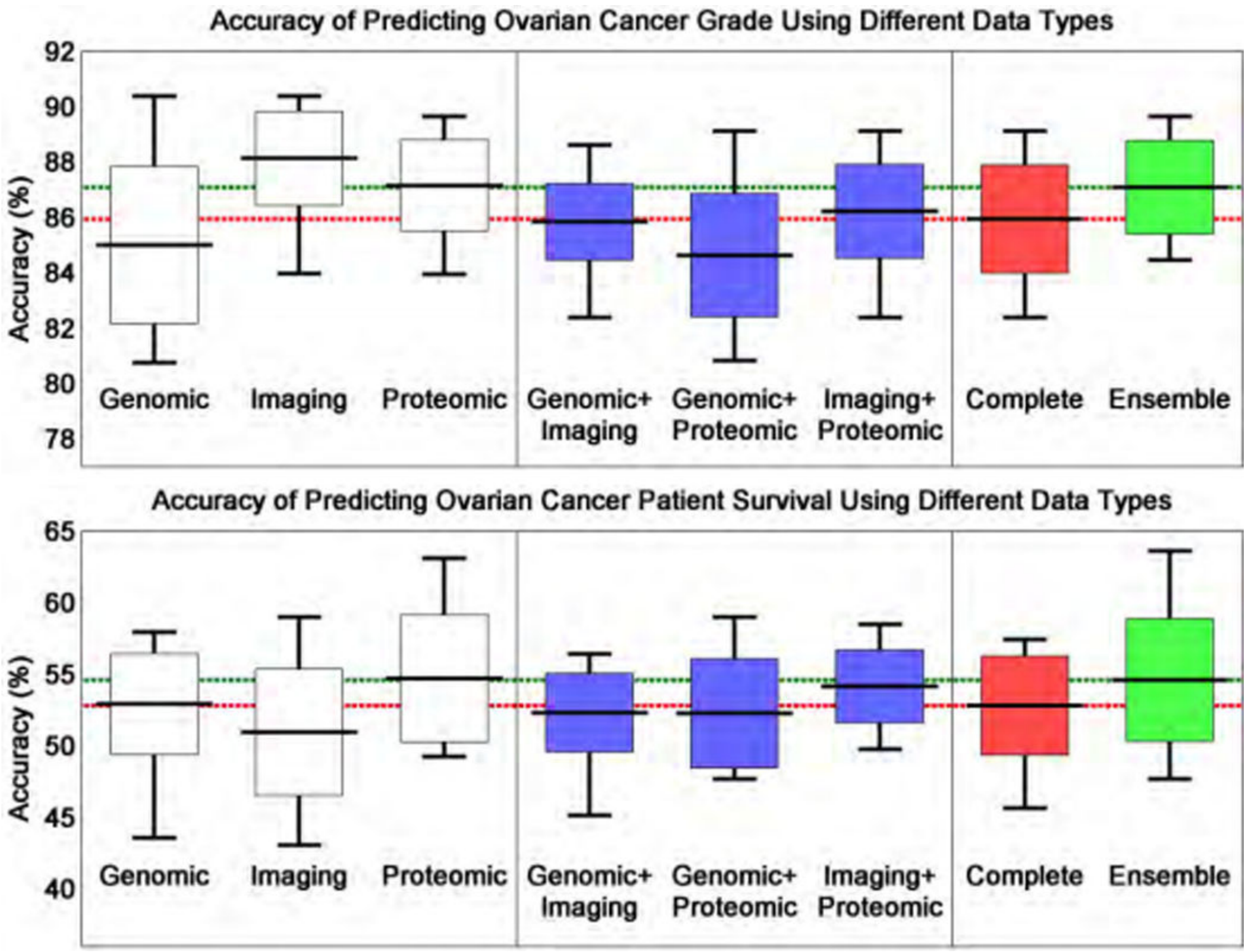

Figure 3.

Prediction of ovarian cancer grade (top) and patient survival (bottom) using individual and integrated datasets. Boxes represent mean ± one standard deviation. Whiskers indicate data range. Red and green dashed lines indicate mean prediction performance for complete and ensemble methods, respectively.

Table 1

Summary of un-integrated/integrated ovarian cancer datasets used for classification.

	Data Modality	**# of Features**
Un-Integrated	Genomic	11410
	Imaging	461
	Proteomic	165
Integrated	Genomic+Imaging	11871
	Genomic+Proteomic	11575
	Imaging+Proteomic	626
	Complete	12036

Table 2

Summary of patient classes for two ovarian cancer clinical prediction endpoints.

	Grade 1 or 2	Grade 3 or 4	Total
# of Patients	50	337	387
	Known Survival <5 Yrs	**Known Survival ≥5 Yrs OR No Follow-Up after 5 Yrs**	
# of Patients	173	209	382

Table 3

Summary of histopathological image features.

Features Subset	Count	Description
Color	73	RGB histograms, histogram statistics, and stain co-occurrence
Global Texture	138	Haralick, fractal, GHM multi-wavelet, gray-level histogram statistics, and Gabor filter
No-Stain-Object Shape	51	Pixel area, boundary fractal, bending energy, convex hull area, solidity, perimeter, elliptical area, major-minor axes lengths, eccentricity, and count
Eosinphilic-Object Shape	51	Pixel area, boundary fractal, bending energy, convex hull area, solidity, perimeter, elliptical area, major-minor axes lengths, eccentricity, and count
Eosinphilic-Region Texture	18	Haralick and gray-level histogram statistics
Basophilic-Object Shape	51	Pixel area, boundary fractal, bending energy, convex hull area, solodity perimeter, elliptical area, major-minor axes lengths, eccentricity, and count
Basophilic-Region Texture	18	Haralick and gray-level histogram statistics
Nuclear Shape	26	Count, elliptical area, major-minor axes lengths, eccentricity, and cluster size
Nuclear Topology	35	Delaunay triangle, minimum spanning, Voronoi diagram tree, and closeness

Table 4

Summary of data combination scenarios tested.

Combination	Description
Genomic	Uses genomic dataset
Imaging	Uses imaging dataset
Proteomic	Uses proteomic dataset
Genomic+Imaging	Uses genomic data concatenated with imaging data
Genomic+Proteomic	Uses genomic data concatenated with protein data
Imaging+Proteomic	Uses imaging data concatenated with protein data
Complete	Uses a concatenation of genomic, imaging, and proteomic data
Ensemble	Uses voting from genomic, imaging, and proteomic prediction models.

Table 5

P-values for two-sample t-tests comparing each integrated dataset's prediction results with those of its constituent datasets. Each row corresponds to an integrated dataset and each column for which it has a value is one of its constituent datasets; the p-value at the intersection of two datasets reflects the probability that their accuracy is the same. Bold p-values indicate statistically significant improvements in performance as a result of data combination.

GRADE	**Genomic**	**Imaging**	**Proteomic**
Gene+Img.	0.1540	0.9998	n/a
Gene+Prot.	0.6500	n/a	0.9993
Img.+Prot.	n/a	0.9978	0.9309
Complete	0.1476	0.9986	0.9609
Ensemble	**0.0104**	0.9494	0.5447

SURVIVAL	**Genomic**	**Imaging**	**Proteomic**
Gene+Img.	0.7011	0.1581	n/a
Gene+Prot.	0.6842	n/a	0.9404
Img.+Prot.	n/a	**0.0104**	0.6592
Complete	0.5318	0.1019	0.8971
Ensemble	0.1259	**0.0143**	0.5255