Combining Multi-Omics Data to Predict Overall Survival of Breast Cancer Patients

Kevin Erazo (@kerazo) , Sameer Merchant (@smerchan) and Rudra S Bandhu (@rsbandhu)

Abstract

Precision medicine, an emerging field of medicine, has enabled prospects of customization of healthcare, medical decisions and treatments tailored to individual patients.The use of genomic and transcriptomic biomarkers as well as other multi-omics data has played a major role in precision oncology. Concurrent with the explosion of clinically relevant molecular data, the application of machine learning methods to multi-omics datasets has become more commonplace. In this paper we present a novel method of combining multi-omics datasets to predict breast cancer patients' overall survival.

1. Introduction

Cancer is a notoriously heterogeneous disease which makes tracking its progression and severity incredibly difficult. The most common cancer staging system is the TNM (tumor, nodes, metastasis) system which is based primarily on clinical information like tumor size, extent of spread, etc. [1]. Combining phenotypic and molecular data from cancer patients can lead to more detailed descriptions of disease progression and severity. To that end, we developed an algorithm that would take as its input molecular data (DNA methylation, RNASeq and miRNASeq data) to predict the overall survival of breast cancer patients. We experimented with linear models (SVM, KNN, PCA with SVM) on individual datasets. The performance of these methods didn't generalize to the validation and test sets well. As such, we opted to implement ensemble based learning methods as discussed in section 4.

2. Related Work

Sathipati and Ho [12] used an optimized SVM regression to identify miRNA signatures associated with survival time in patients with lung adenocarcinoma. They used a novel feature selection algorithm called IBCGA [13] and these features were then fed into traditional SVR. Although their custom SVR outperformed other regression methods, it did not generalize well to unseen validation data. Another issue with this paper was the size of datasets.

In another instance, Zhu et al. [6] incorporated patient profiles of somatic mutations, DNA copy number, DNA methylation, mRNA expression, miRNA expression, protein expression and their combinations to predict overall survival (3,382 samples across 14 cancer types). They created similarity/kernel matrices for all molecular datasets for all patients to understand the underlying biology across cancers that was leading to disease progression. They showed that predictive power depended on the particular cancer type and varied across molecular profiles; no single modality of data was superior across all cancers.

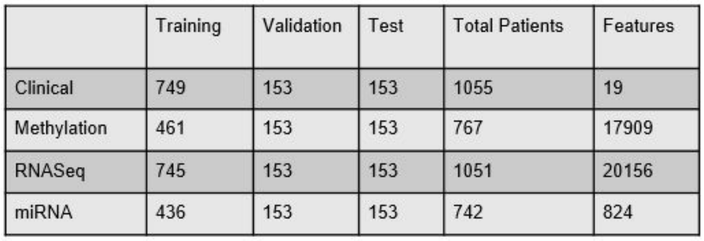
The work of Chaudhary et al. [9] is the most comparable to ours where they use RNASeq, miRNA, and methylation data from patients with hepatocellular carcinoma to identify survival subpopulations using an autoencoder, a Cox-PH feature selection strategy, and k-means clustering. The validity of the resulting clusters/subpopulations was confirmed using non-TCGA molecular datasets. In another example, Kwon et al. [8] combined RNASeq and miRNA data with an SVM classifier to identify prognostic biomarkers for pancreatic ductal adenocarcinoma. They similarly validated their findings by using external datasets and report 705 multi-markers for 27 miRNAs and 289 genes as promising potential biomarkers. Lastly, given the diversity of problems and machine learning approaches, the study by Lin and Lane [5] offered a detailed overview of how to approach multi-omics data integration. Based off their work, we opted to use model-based integration using RNASeq, miRNA, and methylation data for individual classifiers.

3. Dataset and Features

The 4 datasets came from the LinkedOmics database [2] which has organized and compiled data from The Cancer Genome Atlas (TCGA) project. In particular, we obtained datasets of TCGA patients with invasive breast carcinoma [3] (designated TCGA-BRCA, all datasets had patients as rows and features as columns): the clinical dataset, the gene-level miRNA dataset, the HiSeq gene-level RNASeq dataset, and the gene-level DNA Methylation dataset. We had clinical data for all patients in the TCGA-BRCA cohort, but not every patient was present in the available molecular datasets, so we had to split the data carefully. First, patients without an overall survival value were dropped. From the remaining 1055 patients, we found that only 612 patients were represented in all 3 molecular datasets (miRNA, RNASeq, and Methylation). We took half of those patients for validation and testing datasets (153 patients in each); the other 306 patients were used for training data in their respective originating datasets.

After splitting the datasets properly, we dropped features with missing values in each dataset since we had way more features than patients in all datasets. The final datasets for each modality of data are described below in Table 1. Regarding units, the RNASeq and miRNA datasets were provided in log-2 normalized RSEM and RPM, respectively; the methylation dataset was reported in centered beta values. Only the clinical dataset was completely without processing.

Table 1. Splitting available data into training/validation/test datasets after dropping missing values.



	Training	Validation	Test	Total Patients	Features
Clinical	749	153	153	1055	19
Methylation	461	153	153	767	17909
RNASeq	745	153	153	1051	20156
miRNA	436	153	153	742	824

3.1 Survival Data Analysis

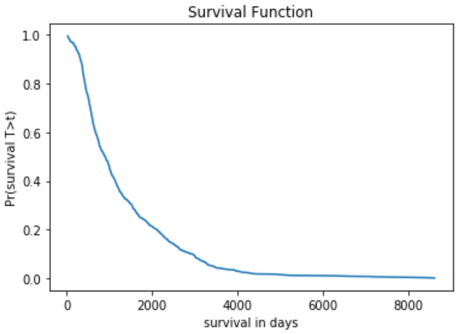


Figure 3.1.1

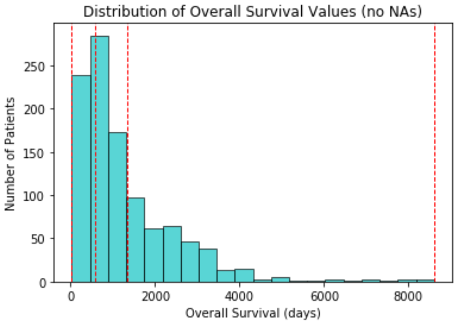


Figure 3.1.2

Figure 3.1.1 shows overall survival probability $P(T > t)$, in days. Our dataset had very few patients with high overall survival.Gene and molecular biomarkers cannot provide such precise survival time in days. Our initial attempts to predict survival time as regression failed due to (1) non-uniform distribution of samples (2) granular survival time. Most literature we surveyed, used classification to predict short versus long term survival. Hence we reframed our problem as a classification problem. We split overall survival time into three buckets (shown in Figure 3.1.2): short (<1.5yrs), medium term(1.5-3.5yrs), high (>3.5yrs) survival time. The samples were uniformly distributed across these categories.

3.2 Feature Selection:

Multi-omics dataset suffers from the curse of dimensionality $p >> n$. Based on prior studies [10-11], Supervised PCA has proven to be effective in selecting subset of features. Let X be a $(n \times p)$ feature data matrix and Y be an n dimensional overall survival target vector. For Supervised PCA, we (i) compute standard regression coefficients for each feature $s_j = \frac{X_j^T Y}{\|X_j\|}$ (ii) form a reduced X_θ matrix of size k consisting of features with highest s_k regression coefficients, (iii) compute principal components of this reduced matrix and use these as input features for classification. We tried different values of k for each dataset. Based on our experiments, we found 10 features from miRNA, 10 features from methylation and 25 features from RNA dataset worked best in reducing variance and improving generalization of our model.

4. Methods

In this section we discuss various methods used for selecting a framework for multi-omics based survival prediction. We did a baseline evaluation of performance on each dataset using standard classification methods (SVM, Random Forest, Gradient Boosted decision trees, K-NN). Based on prior work, we combined three datasets to train a model that generalized better as compared to models trained on individual datasets.

4.1 Ensemble Method

Prior work [5] has shown promising results in cancer survival prediction by combining a collection of weak models using SVM, kNN and Gradient Boosted Decision trees. We adopted two methods for ensemble learning.

4.1.1 Ensemble Model using Individual Datasets

In this method we trained three independent classifiers for each dataset. We took two parallel approaches. First, the features were selected using supervised PCA which were used to train an ensemble of learners (SVM, KNN and gradient boosted trees). The second approach was to use all features for training XGBoost classifier. This helped us determine the effectiveness of feature selection either through supervised PCA or automatically using boosted trees.

For ensemble based method, we used SVM, one-vs-rest classification method, with linear kernel and misclassification penalty parameter C=1000. We used 3 nearest neighbors, with uniform weights and brute force search. The predicted probability for each class from SVM and K-NN was combined using a soft vote voting classifier. The voting classifier picked highest probability for each class in each dataset.

In the XGBoost model, the boosting parameters for each model are tuned separately using only that dataset. For both methods, we use a weighted score of probability of each class from the 3 datasets to determine the final probability for each class which is then used to determine the class prediction (class with max probability). The weighting value for each dataset is tuned by optimizing the accuracy on the training set. Figure 4.1 shows the framework for both methods (ensemble and XGBoost).

4.1.2 Ensemble Method using Combined Datasets

We combined all three dataset as single feature vector. We used the same approaches as described in section 4.1.1 to determine the input features for the different models (ensemble based method used SPCA to pick top 200 features). We used an ensemble of SVM, K-NN and Gradient Boosted Decision Tree for classifying overall survival time. For SVM we used one-vs-rest method, with linear kernel and misclassification penalty parameter C=500. We used 7 nearest neighbors for k-NN method, with uniform probability and brute force search. For gradient boosted decision tree, we used 100 estimators, with min_sample_split=40, max_depth=2, learning_rate=0.01. Figure 4.2 shows the combined dataset ensemble method.

As described in section 4.1.1 the second approach was to use XGBoost using all features from all 3 datasets as a single input to the model (612 samples and 38889 features). Here we tune the parameters of the model through uniform sampling over an appropriately chosen range. Training was stopped when there is no improvement in the log-loss of the validation set after 40 rounds.

4.2 Neural Network

Recent literature suggested promising results in survival prediction with Deep Neural Networks and multi-omics data. We compared our ensemble method performance with a three hidden layer neural network.Input layer had 200 nodes, followed by hidden layers with 20 nodes, 200 nodes (similar to methods suggested by Chaudhary et. al [9].The output layer had 3 nodes, one for each class. We used sigmoid activation for output layer and ReLu for hidden layers. Figure 4.3 shows the Neural Network architecture. We applied Supervised PCA to reduce features used for training the network.

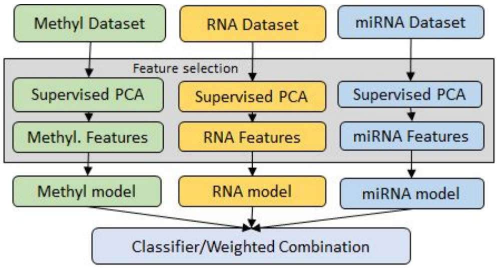


Figure 4.1

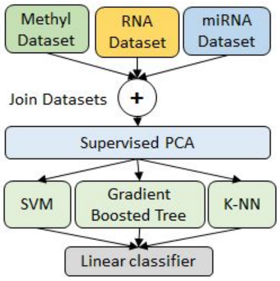


Figure 4.2

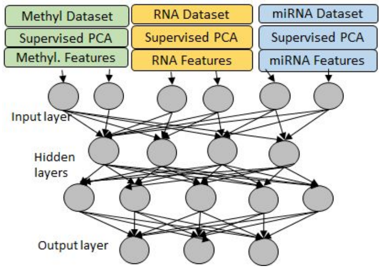


Figure 4.3

5. Experiments and Results

Our primary challenge with these datasets was high dimensionality which resulted in overfitting. All classification models trained on individual datasets using SVM, Random Forest, Gradient Boosted Decision Trees, suffered from overfitting with full feature set.

Combining weak predictors with ensemble methods has proven to give better results [7]. We used Supervised PCA (SPCA), for selecting subset of features. We used SPCA correlation coefficients to select top 10 features from methylation dataset, 25 features from RNA dataset and 10 features from miRNA dataset. Subsequently, we used 5-fold cross validation and GridSearch to tune hyperparameters for Gradient Boosted Decision Trees, SVM and k-NN. We tried SVM misclassification parameter C (10, 50, 100, 500, 1000), different max-depth (1, 2, 5, 10), min_sample_split (5, 10, 20, 50), learning rate (0.1, 0.05, 0.01) and number of estimators (50, 100, 200, 250) for Gradient Boosted Decision Trees. Based on GridSearch, we selected the final hyperparameters listed in section 4 to minimize variance across models.

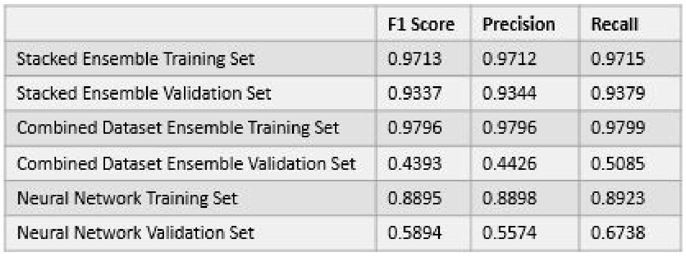
XGBoost method provided the best performance among all the methods we used. We used careful tuning of the boost parameters to achieve this result. Parameter “max_depth” was set to 2. Using a higher value for this parameter led to overfitting and poor generalization. The optimal learning rate was in the range (0.06, 0.18). Gamma parameter resulted in the range 6-9. Subsampling by rows (patients) was around 0.8 and sub-sampling by columns (features) was in the range 0.54-0.8. Parameter “min_child_weight” was kept in the range 5-9 to avoid avoid overfitting.

The highest accuracy, 43.8%, is achieved on the test set using ensemble of individual dataset features with XGBoost model (method1). The results are summarized for this method in Figure 6. We observed that the weights of the datasets are relatively balanced with RNAseq dataset appears to have the lowest weight (~ 0.23). The weights for the methylation, miRNA and RNAseq were around 0.42, 0.35, 0.23 respectively. This can be understood in terms of the number of features present in each of these datasets. Since the RNAseq data has the highest number of features the model has lower accuracy on this dataset and hence puts less weight on the prediction probability from this dataset. Using combined features from all 3 datasets, we were able to obtain similar accuracy on the test set (42.5%).

5.1 Performance and Error Analysis

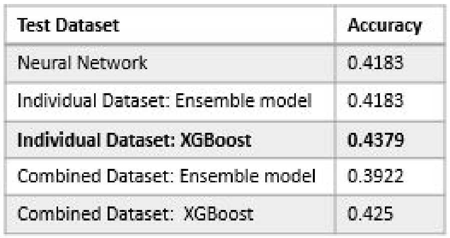
We evaluated our model performance using confusion matrix, F1 score and accuracy. Table 5.1 summarizes performance of each model. Table 5.2 summarizes accuracy on holdout test set for each model.

All models performed well on classifying short time survival (<1.6yrs), but had poor performance for survival (>3.5yrs). This is primarily due to imbalance in samples in our dataset that were representative of patients with higher survival time. The clinical data wasn’t sufficient to identify correlation between biomarkers from three dataset for higher survival.



	F1 Score	Precision	Recall
Stacked Ensemble Training Set	0.9713	0.9712	0.9715
Stacked Ensemble Validation Set	0.9337	0.9344	0.9379
Combined Dataset Ensemble Training Set	0.9796	0.9796	0.9799
Combined Dataset Ensemble Validation Set	0.4393	0.4426	0.5085
Neural Network Training Set	0.8895	0.8898	0.8923
Neural Network Validation Set	0.5894	0.5574	0.6738

Table 5.1



Test Dataset	Accuracy
Neural Network	0.4183
Individual Dataset: Ensemble model	0.4183
Individual Dataset: XGBoost	**0.4379**
Combined Dataset: Ensemble model	0.3922
Combined Dataset: XGBoost	0.425

Table 5.2

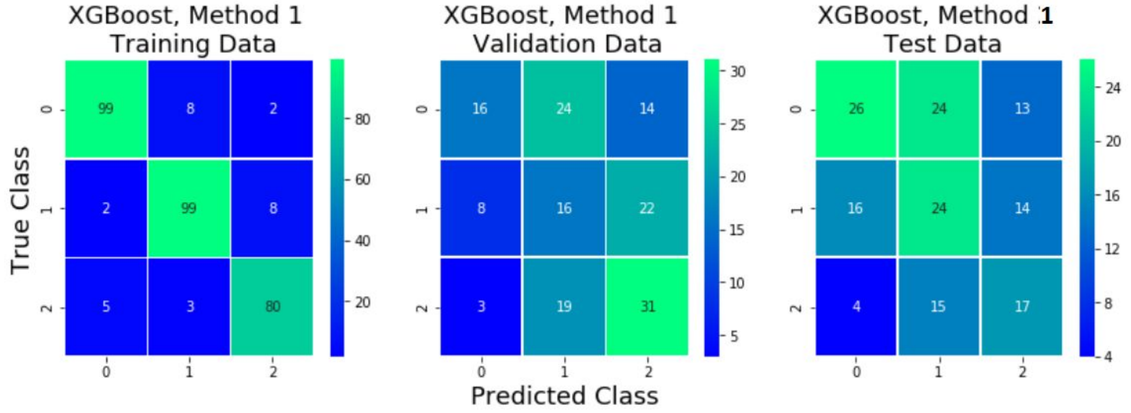


Figure 6: Performance of XGBoost (individual datasets) on the training, validation and test set.

6. Conclusion

After extensive experimentation, we found that an ensemble of 3 XGBoost classifiers, each trained on a different modality of data, performs best. Individual classifiers may overfit to their particular datasets, but the ensembling procedure, where each model's probabilities are weighed to yield a final classification attenuates overfitting and helps with generalization. While the performance of supervised PCA is slightly inferior to XGBoost, it provides explanatory power in a clinical setting by identifying key molecular features that are relevant markers for predicting overall survival. The results from SPCA picked out previously validated biomarkers from each dataset: from the methylation dataset, CDCA7 has been associated with progression of triple-negative breast cancer [15]; from the RNASeq dataset, AEBP1 signaling has been hypothesized to predispose mammary tissue to tumorigenesis [16]; and form the miRNA dataset, downregulation of miR-24-1 has been identified as a prostate cancer progression marker [17].

7. Future Work

We propose to explore the following approaches to improve performance: (1) Use class-balanced loss function to address imbalances in clinical dataset with natural split of overall survival, (2) incorporate phenotype data from clinical dataset as additional features, (3) incorporate clinically-relevant metrics like Inverse Probability Weighting into models, (4) deeper exploration using NN architecture, (5) augment clinical data from other sources to increase sample size and diversity, (6) use additional molecular data from the same clinical data source, (7) extend the method to predict overall for other cancer types.

8. Contributions

- Rudra S Bandhu: worked on the XGBoost method and initial data exploration phase (standard linear classifier methods) with RNAseq dataset.

- Sameer Merchant: Evaluating multiple regression and classification methods on miRNA dataset. Supervised PCA feature selection across all three datasets, Ensemble Learning with three data sets and stacking. Combined dataset Ensemble learning. Implemented Neural Network. Comparing ensemble learning with Deep Learning Neural Network with 3 layer network. Error analysis.
- Kevin Erazo: Analysis of the methylation dataset: multiple regression and classification strategies, various dimensionality reduction techniques, and hyperparameter searches for each regressor/classifier. Data preparation and dataset split for the final ensemble classifier.

Github link: https://github.com/smerchan/cs229project

Collabs Link:
https://colab.research.google.com/drive/17XP5tEvpOq7fAWv8A7LoqbE2iWiKPGvl?authuser=1#scrollTo=sXTQex2p0ReE

9. References

[1] https://www.cancer.gov/about-cancer/diagnosis-staging/staging
[2] Vasaikar, S., Straub, P., Wang, J. and Zhang, B. (2017). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research*, 46(D1), pp.D956-D963.
[3] http://linkedomics.org/data_download/TCGA-BRCA/
[4] A Framework for Implementing Machine Learning on Omics Data
[5] Lin, E. and Lane, H. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomarker Research*, 5(1).
[6] Zhu, B., Song, N., Shen, R., Arora, A., Machiela, M., Song, L., Landi, M., Ghosh, D., Chatterjee, N., Baladandayuthapani, V. and Zhao, H. (2017). Integrating Clinical and Multiple Omics Data for Prognostic Assessment across Human Cancers. *Scientific Reports*, 7(1).
[7] Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. and Ping, P. (2019). Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes*, 10(2), p.87.
[8] Kwon, M., Kim, Y., Lee, S., Namkung, J., Yun, T., Yi, S., Han, S., Kang, M., Kim, S., Jang, J. and Park, T. (2015). Integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. *BMC Genomics*, 16(Suppl 9), p.S4.
[9] Chaudhary, K., Poirion, O., Lu, L. and Garmire, L. (2017). Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical Cancer Research*, 24(6), pp.1248-1259.
[10] Bair, E., Hastie, T., Paul, D. and Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, 101(473), pp.119-137.
[11] The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Chap 18, Trevor Hastie, Robert Tibshirani, Jerome Friedman
[12] Yerukala Sathipati, S. and Ho, S. (2017). Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. *Scientific Reports*, 7(1).
[13] Ho, S., Chen, J. and Huang, M. (2004). Inheritable Genetic Algorithm for Biobjective 0/1 Combinatorial Optimization Problems and its Applications. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 34(1), pp.609-620.
[14] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, Serge Belongie, Class-Balanced Loss Based on Effective Number of Samples, CVPR 2019.

[15] Ye, L., Li, F., Song, Y., Yu, D., Xiong, Z., Li, Y., Shi, T., Yuan, Z., Lin, C., Wu, X., Ren, L., Li, X. and Song, L. (2018). Overexpression of CDCA7 predicts poor prognosis and induces EZH2-mediated progression of triple-negative breast cancer. *International Journal of Cancer*, 143(10), pp.2602-2613.
[16] Holloway, R., Bogachev, O., Bharadwaj, A., McCluskey, G., Majdalawieh, A., Zhang, L. and Ro, H. (2012). Stromal Adipocyte Enhancer-binding Protein (AEBP1) Promotes Mammary Epithelial Cell Hyperplasia via Proinflammatory and Hedgehog Signaling. *Journal of Biological Chemistry*, 287(46), pp.39171-39181.
[17] Goto, Y., Kojima, S., Nishikawa, R., Enokida, H., Chiyomaru, T., Kinoshita, T., Nakagawa, M., Naya, Y., Ichikawa, T. and Seki, N. (2014). The microRNA-23b/27b/24-1 cluster is a disease progression marker and tumor suppressor in prostate cancer. *Oncotarget*, 5(17).