Integration of Multi-Modal-Omics Data for Breast Cancer Patients Survival Analysis

Shreenu Sivakumar
Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA USA
ssivakumar36@gatech.edu

Michael-Alexander Rivera
Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA USA
maxrivera@gatech.edu

BACKGROUND AND MOTIVATION

Breast cancer is the most prevalent type of cancer affecting women both in the US and the rest of the world. In 2019 alone, more than 250,000 new cases were recorded among women, resulting in 41,760 deaths (DeSantis et al). With a 13% mortality rate among women, it is the second deadliest among cancer types (Mitchel et al). Unfortunately, breast cancer is very complex and highly variable in many different aspects, such as molecular, clinical, etc. Because of its high diversity among patients, there is a high demand for personalized, precision medicine and identification of important biomarkers (Li et al).

Unfortunately, many biomarkers have been identified by doctors with questionable accuracy and consistency. For example, the concurrence of tumor evaluations by several separate pathologists is less than 50% (Gyorffy et al). Incorrect tumor subtyping can lead to a dramatically wrong direction for treatment, with irreversible consequences. Because of inconsistent doctor analysis, biomarker identification has been increasingly focused on molecular modalities (gene expression, protein expression, etc.).

In addition, the explosion in technology used to collect biomolecular data has contributed to the movement toward molecular biomarker discovery. Advances in high-throughput sequencing have enabled cheap and frequent data collection of various types of genomic data (Liu et al). For example, The Cancer Genome Atlas (TCGA) has hundreds of patient samples with data from mRNA sequencing, miRNA, DNA methylation, and copy number variation (CNV) for many types of cancer. With the abundance of data available only increasing, the need to make sense of it using informatics techniques has also risen. Analyzing different genomic data modalities using bioinformatics is paramount to personalized and precision medicine (Phan et al).

1 Technical Objective

By integrating data from multiple sources and modalities, we can make better use of scientific insights, enabling earlier diagnosis and improved treatment plans for patients (Phan et al). Multi-modal data makes use of the interactions and associations that may occur between different types of data. In biomedical research, multi-modal integration of data has shown improved prediction performance over single modal integration (Mitchel et al).

Because of the high dimensional and complex nature of genomic and biomolecular data, there is a clear limitation in doctors' expertise and ability to analyze or integrate it. Machine learning is a technique that can be used to find patterns in data that are otherwise too convoluted to see. Using machine learning, high dimensional data containing important biomarkers can be integrated to improve cancer prognosis and clinical outcomes (Cheerla et al).

The Cox Proportional Hazard Model is a machine learning algorithm that is commonly used in analysis where how long it takes for an event to occur is important, not just if an event occurred or not. Using this model can tell us the simultaneous effect of one of more predictor variables, features of covariates, on the time it takes for an event to occur. This effect is known as the hazard or risk and is the probability that an individual at any given time will experience an event, known as the instantaneous event rate. The Cox Model is useful for a study involving time-to-event data where the aim is to predict outcomes based on covariates. It is commonly used in clinical drug trials, time to relapse in cancer studies, future medical expenses, and disease prognosis. This model will be utilized in this study to analyze breast cancer gene expression data and predict survival.

2 Problem Statement

Our aim for this project is to improve upon previous classifiers by producing a model that can handle censored data and survival time information in order to predict breast cancer survival based on the integration of four data modalities: miRNA expression, RNA sequencing, copy number variation, and DNA methylation. More specifically, our prediction model will produce a risk score per patient based on gene expression data and estimate the likelihood of survival in relation to time after detection.

LITERATURE CRITIQUE

Prior to developing a proposal for the integration of multi-modal omics breast cancer data, research was done into relevant works spanning from prior survival models to studies investigating a specific method of data processing within modeling. These prior works were helpful in guiding the process of developing and implementing our proposal.

First, the aim of Gyorffy et al, a Hungarian preclinical study published by Springer Science+Business Media in 2009, is similar to ours: to produce an online tool that analyzes gene expression levels from thousands of different genes and output survival plots with the prognostic value of certain genes of interest (Gyorffy et al). The methods that this research team used included normalizing the GEO microarray data using the MAS5 function in R, performing quality control by removing repeat samples, dividing the remaining data into two groups based on the median expression of the gene, conducting hypothesis testing using an alpha of .05, and producing a Kaplan-Meier plot to visualize the separation of high and low gene expression groups (Gyorffy et al). The main strength of this paper was that it was a preliminary exploration into the creation of a tool to predict the likelihood of prognosis for a specific gene given a large input of patient data. This study was conducted over a decade ago making it an important reference for the state of the art earlier on, even though it is not representative of the complexity of gene expression studies performed today. More specifically, this paper produced Kaplan-Meier plots as a survival model based on median expression and in their discussion, they explained that the study could be improved by using the Cox proportional hazards model to calculate specific hazard ratios each data point (Gyorffy et al). Because Kaplan-Meier plots can clearly display the separation of two groups, we decided to combine this visualization with the Cox proportional hazards model in our study.

Mitchel et al, a Georgia Tech undergraduate research team, applied decision-level binary classifiers to the integration of multi-omic data in order to predict survival outcome of patients, achieving model accuracy of 85% and AUC of 87% (Mitchel et al). Because this decision-level integration of multi-omic data avoids increasing dimensionality of the feature matrices, a problem that occurs in feature-level integration methods, we seek to use it as well in our survival analysis model. Binary classification used in this study required the division of patient groups and additional steps of removing right-censored data. These additional steps can be seen as a limitation, opening up the possibility of improvement by incorporating models that do not require the process, such as Cox regressions.

Goli et al conducted a study to compare survival models on censored data, including support vector machines (SVMs) and the standard Cox proportional hazard model (Goli et al). The study found that SVM actually outperformed the standard Cox model without any feature selection at all (Goli et al). However, once dimensionality was reduced, the difference between the two methods was not significant. Model evaluation included C-index and log-rank test. Because of this, we choose to focus on Cox proportional hazard models and their variations. A limitation of this study is the use of many different modalities such as age of diagnosis, tumor size, histological type, etc. that may not translate to our model design.

Many studies have also used Cox proportional hazard models for survival analysis with high accuracy. The current state of art model was produced by Liu et al. who used a cox-regression model for survival prediction from breast cancer data (Liu et al). Interestingly, they used only one mode of data, mRNA expression, to train the model (Liu et al). This was due to research that found mRNA expression to have the most predictive power compared to other modalities. The advantage of this is the ability to create a predictive model from only one form of data and later apply that model on a group of patients with different modalities. Because of this, mRNA data will be an important set of features in our model. However, this model does not have the ability to integrate multiple data modalities for prediction, which is important due to the results of previous studies indicating that multi-omic integration increases performance (Mitchel et al). Because of this limitation, this model is not an ideal framework for our survival analysis.

A more complex application of the Cox Model was created at the University of Cambridge by Ching et al and a novel combination of artificial neural networks (ANNs) with the Cox-regression model to integrate multiple modalities of genomic data (Ching et al). It was found that Cox regressions that use a single hidden layer outperformed standard Cox models and random forest trees (Ching et al). However, the model also outperformed ANNs with more than 2 hidden layers, indicating that the use of deep learning with several hidden layers may be unreliable (Ching et al). A problem often run into with the use of neural networks and deep learning is overfitting because of the large amount of parameters and small amount of samples. This new model, called “Cox-nnet”, combatted this issue by implementing a drop-out method of feature selection (Ching et al). Drop-out methods prevent overfitting by randomly “dropping out” or removing neurons from a neural network, in order to force remaining neurons to compensate, resulting in a more generalized model. To evaluate the Cox-nnet performance, c-index was calculated, as well as log rank p-value in order to visualize high and low risk groups (Ching et al).

Huang et al used Cox regression combined with a deep neural network to predict survival outcomes of breast cancer patients, published in *Frontiers in Genetics* (Huang et al). A common problem with deep learning models is optimizing many parameters with a limited sample size. A strength of this study was its method to reduce dimensionality of genomic data through the use of co-expression matrices containing associations between RNAseq and miRNAseq data. This reduced features by 99.46%, making the neural networks more efficient. They also chose to include a Cox proportional hazards model over previously used binary classification because of its ability to incorporate survival time. To evaluate their model, concordance index was used due to its popularity in survival analysis models. Classification methods typically use ROC curves for evaluation, but C-index is equivalent to ROC for survival models. In addition, the use of the log rank test for separating high and low risk groups was important to evaluate how well the model could differentiate risk. Though this model yielded a high C-index of 0.7285, its performance was not significantly better than Cox-nnet (Ching et al).

In our analysis of related works, we noticed a trend in many studies skimming over normalization and quality control steps, simply stating the function used for normalization and highlighting the number of samples removed after quality control without giving adequate reasoning why. Therefore, we also dove deeper into studies that focused specifically on the methods of normalization and quality control. For example, Rapaport et al examined RNA sequencing data in the context of how to evaluate differences in expression levels through normalization techniques (Rapaport et al). This study was published in 2013 on *Genome Biology,* a journal with an excellent impact factor making it a reliable and important source (Rapaport et al). They specifically used log2 normalization as the basis before comparing other normalization methods so that the magnitude of data points was reduced (Rapaport et al). We observed that the use of log2 as a basis for normalization was fairly common in gene expression studies. For example, Chow et al also used the log2 transformation as the first step before the data was normalized using robust spline normalization, simple scaling normalization, and variance stabilizing normalization (Chow et al). A strength of the log2 function is that it is an efficient and common way to scale down data, however another normalization method generally must be used afterwards to actually “clean up” the data points.

A common normalization technique that we encountered in our reading of related works was the min-max normalization method. Sun et al performed a study on breast cancer prognosis prediction using deep learning, specifically multimodal neural networks to perform data integration (Sun et al). This research team published their findings in 2019 making it a recent snapshot of survival modeling in bioinformatics and an incredibly useful tool for us moving forward (Sun et al). Although their survival modeling was done using deep learning, the preprocessing of clinical data was done through simple min-max normalization in which the range of data is converted from their normal range to values between 0 and 1, while maintaining the magnitude of variances between points (Sun et al). The benefit of using this method is that it allowed researchers to standardize the range of data between different data modalities, yielding a more accurate comparison between them. A weakness inherent within this method is that thorough quality control must be performed prior because outliers will affect the distribution of values after the min-max function is used. This is because values are reset to a range between 0 and 1 so if there is an extremely high data point then that value will take on the new value of 1, when in reality it is not reflective of the actual distribution of data.

When researching how to evaluate the efficacy of the Cox proportional hazards model, we came across studies mentioning the use of a "C-index" (Mayr et al). Mayr et al described the concordance index as a "non-parametric measure to quantify the discriminatory power of a prediction tool" (Mayr et al). This article was published in 2014 on PLOS ONE, a journal with a relatively average impact factor of 2 (Mayr et al). This is a fairly specific study into a singular metric for evaluating survival models and therefore has a niche audience, which makes it helpful for our research into how the C-index works. Mayr et al described the favorable use of the C-index in studies where "the aim is to subdivide patients into groups with good or poor prognosis" (Mayr et al). The aim of using the Cox proportional hazards model is to produce risk scores for each patient based on data integration and separate the patients into high and low risk groups making the C-index a valid metric for evaluation (Mayr et al). A weakness of the C-index highlighted by this paper is that it does not provide a visual representation of the significance of high or low risk groups and instead just yields a numerical value. Therefore, it is important to supplement this metric with other evaluation methods to more thoroughly analyze the results of the Cox proportional hazards model, such as a Kaplan-Meier plot.

METHODOLOGY AND SYSTEM DESIGN

3 Data Methodology

All data was downloaded from The Cancer Genome Atlas (TCGA) and organized by USC Xena. These modalities include: miRNA expression, RNA sequencing, DNA methylation, and Copy Number Variation. Copy Number Variation data describes how many times different genes or sequences have been repeated. DNA methylation describes the levels of methylation occurring at various sites on the genome. RNA sequencing data show the level of expression of various genes (how often a gene has been transcribed). Additionally, miRNA expression data is another modality corresponding to levels of gene expression. There have been several studies showing the individual predictive power of miRNA expression, DNA methylation, RNA sequencing, and Copy Number Variation for cancer patient survival analysis (Liu et al). Because of this, we intended to incorporate all 4 modalities into our predictive model. However, due to time constraints, the pipeline was only completed for miRNA and RNA Sequencing data with relation to the survival data.

Modality	Unit	Feature	Type	# Features	# Samples	Dynamic Range
miRNA	RPM	miRNA identifier	Continuous	1,882	1,202	0-20
RNA seq	FPKM	Gene ID	Continuous	60,484	1,217	0-20
DNA Methylation	Beta value	Cg probe identifier	Continuous	485,578	890	0-1
Copy Number Variation	[unitless]	Gene ID	Discrete	19,730	1,104	0, 1, -1

Table 1. Characteristics of each modality of data downloaded from The Cancer Genome Atlas TCGA database

Preprocessing of the data began with normalization in which we scaled down each data modality so that it was easier to visualize and integrate the four different datasets. Our preprocessing consisted of three steps, specifically reducing the magnitude of gene expression in each data modality, removing outliers, and adjusting the range so that all modalities could be compared. Prior to our preprocessing, each data modality from the UCSC Xena Explorer had already been log2 transformed to reduce the magnitude of gene expression levels. The resulting data had a range from 0 to several thousand.

The second step in normalization is the removal of outlier data using the Z-transformer. In this method, we removed outliers outside of three standard deviations from the mean. The reasoning behind this was that it is imperative to remove outlier data prior to standardizing the range because the min-max function is especially sensitive to outliers because it adjusts the range of values to a distribution of [0,1]. The min-max function, rescale() in MATLAB, is the final step in normalization and is used to reduce the range of gene expression values so that they can be accurately compared to copy number values, all -1,0, or 1, in the copy number variation dataset. Without normalization, it would be impossible to accurately integrate these datasets because the ranges would be so different, making changes in copy number variation seem negligible in comparison to the magnitude of change in gene expression levels. The min-max function was chosen because it is a simple normalization function that is used as a standard in many gene expression studies (Sun et al, Mitchel et al)

Preprocessing also consists of quality control, a method to clean up the dataset so that only complete samples are considered in analysis. We performed quality control by removing repeat samples, samples with missing entries, and samples with all gene expression levels of 0 for a feature. It was necessary to remove repeat samples so that the validity of each Sample ID was preserved. Samples with missing entries were also removed as well because including Sample IDs for missing gene expression values alters the median of the sample. Features with gene expression levels of all zeros were also removed because the goal of this project is to analyze the impact of gene expression levels on survival. If all samples do not express that gene then that feature is not relevant to the aim of this study.

4 Informatics System Workflow

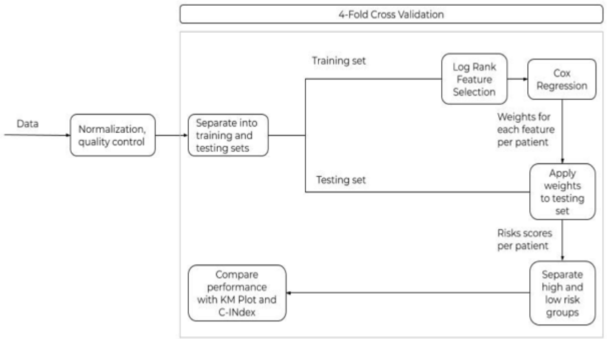

4-Fold Cross Validation
Data
Normalization, quality control
Separate into training and testing sets
Training set
Log Rank Feature Selection
Cox Regression
Weights for each feature per patient
Testing set
Apply weights to testing set
Risks scores per patient
Separate high and low risk groups
Compare performance with KM Plot and C-INdex

Figure 1. Machine learning pipeline for the integration of modalities and survival data to yield patient risk scores and survival probabilities. Data was preprocessed through normalization and quality control before being split into training and testing sets. Log rank feature selection was applied separately on each modality of data to reduce dimensions before being fed into separate Cox regression models and finally integrated at the decision level. Model performance was assessed using the C-index and Kaplan-Meier plot.

5 Classification with Prediction Methodologies

5.1 Cross Validation

After preprocessing, the next step in the pipeline was cross-validation. When dealing with a small sample size, it is important to minimize the influence it has on the model during training (Mitchel et al). Cross validation is used to split the data into balanced training and testing sets. The training is then further split into training and validation sets. This method is often used to compare how well models work before they are used on actual test data.

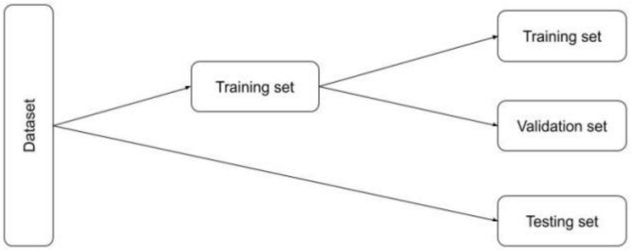
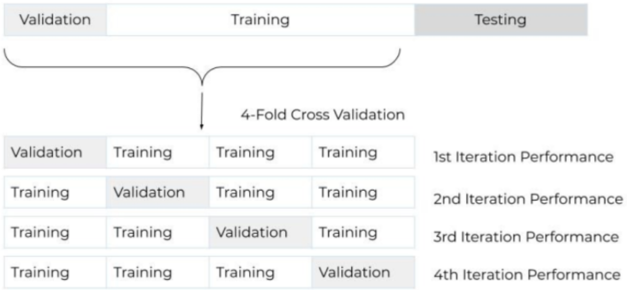


Figure 2. Prior to feature selection, the dataset is separated into training and testing sets before being further separated into training and validation sets via 4-fold cross validation method.

The model is trained on the training set and applied to the validation and testing set to generate survival risk scores. Then the process is repeated for every possible combination of training and validation set. This provides a more realistic performance that shows how well the model can be generalized to new data. We used a standard 4-fold method that has been used in previous studies on gene expression data (Mitchel, et al).



Validation	Training	Testing

Validation	Training	Training	Training	1st Iteration Performance
Training	Validation	Training	Training	2nd Iteration Performance
Training	Training	Validation	Training	3rd Iteration Performance
Training	Training	Training	Validation	4th Iteration Performance

Figure 3. After training data is split 4-fold, every combination of validation and training portions will be tested and averaged to yield a combined performance.

The MATLAB function crossvalind() was used to generate indices for the training, validation, and testing sets. K-fold cross validation was used so that each of the four folds was used as a testing set once and performance was compared for all four of the models.

5.2 Feature Selection

The selection of features for analysis is imperative to reducing the dimensionality of data fed into the survival model. We used feature selection to identify genes of interest within each data modality and use those as the inputs to our model. Our feature selection method to determine genes of interest consisted of a combination of reviewing past literature and performing hypothesis testing to determine the significance of each gene within the modalities on survival. More quantitatively, we used the log rank test to analyze each gene by splitting the gene expression levels by the median. We then performed a hypothesis test to gauge how significantly different the high and low expression groups were, using an alpha of .05. If a gene's p-value is below .05, we deem the gene significant to breast cancer and will include it in our dataset. Genes with p-values above .05 will be omitted from this study.

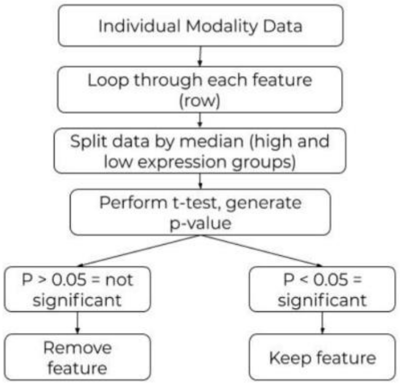
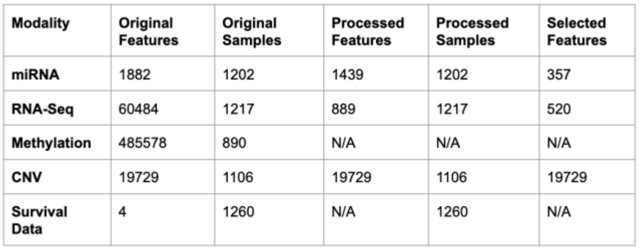

Individual Modality Data
Loop through each feature (row)
Split data by median (high and low expression groups)
Perform t-test, generate p-value
P > 0.05 = not significant
P < 0.05 = significant
Remove feature
Keep feature

Figure 4. Log-rank feature selection was used to remove features with low variability in expression levels.

The log rank test was chosen because of its use in previous cancer studies for the purpose of analyzing which genes are relevant to survival. For example, Attallah et al used the log-rank test on an Endovascular Aortic Repair dataset to evaluate features based on an average of the p-values produced during the stepwise process. The log-rank test is commonly used to evaluate significance between two groups in gene expression studies, stepwise or not (Gyorffy et al). The reduction of features post pre-processing, quality control, and feature selection can be seen in Table 2.



Modality	Original Features	Original Samples	Processed Features	Processed Samples	Selected Features
miRNA	1882	1202	1439	1202	357
RNA-Seq	60484	1217	889	1217	520
Methylation	485578	890	N/A	N/A	N/A
CNV	19729	1106	19729	1106	19729
Survival Data	4	1260	N/A	1260	N/A

Table 2. The number of features per modality was reduced using preprocessing, quality control, and feature selection. The modalities of relevance are miRNA, RNA seq, and the survival data.

5.3 Cox Proportional Hazards Model

The Cox survival model was used because it accounts for how long it takes for an event to occur, not just if the event occurred or not. This model can tell us the simultaneous effect of one of more predictor variables on the time it takes for an event to occur. This effect is known as the hazard, or risk, and is the probability that an individual at any given time will have an event. The Cox proportional hazard model is also particularly useful when analyzing censored data. Censored data refers to data in which the event of interest is not observed during the follow up, or end of study. Many time-to-event data forms in the biomedical field use censored data, for example, in a cancer survival study, if a patient drops out of the study, or dies before the follow up, their survival time is considered censored. Previous binary classifiers for multi-modal integration are not equipped to handle censored time-to-event data, but the Cox model is, making it a promising avenue for survival prediction. In order to use the Cox model, the proportional hazards assumption needs to be met. This assumption states that the ratio between any two hazards remains proportional over time, which can be verified using Kaplan-Meier curves, as seen in Figure 5.

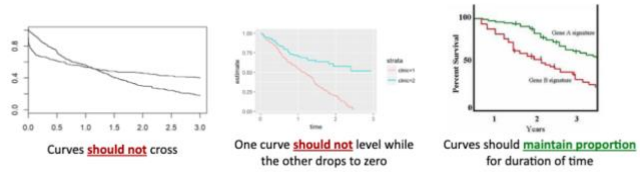


Figure 5. Kaplan-Meier curves can be used to verify that the proportional hazards assumption has been met.

The Cox proportional hazards model follows the Cox Regression equation, as seen in Figure 6.

$$h(X_i, t) = h_0(t)\exp\left[\sum_{j=1}^{p} x_{ij} b_j\right],$$

Figure 6. Cox proportional hazards regression equation to output risk score h(X,t) using features (X), feature weights (B), and the baseline hazard $h_0(t)$

The regression function outputs a hazard h(X,t) at a given time t, as determined by a set of features Xi = (x_1, x_2, x_3, … x_p). The coefficients, or beta values, (b_1, b_2, b_3, … b_p) measure the weight or impact size of each feature. The baseline hazard, $h_0(t)$ is used to scale all the hazards with relationship to a control hazard. In our model, we were able to ignore this baseline hazard h0(t) because we did not have a control hazard that we were comparing our results to. We essentially wanted to compare all the hazards to each other, so multiplying each hazard by a baseline hazard would essentially retain the proportionality of the values, just adjust the magnitude of the hazards themselves. Therefore, it was not necessary to include the baseline hazard.

The implementation of the Cox model in MATLAB began with the use of the coxphfit() function from the MATLAB Statistics and Machine Learning toolbox, as seen in Figure 7.

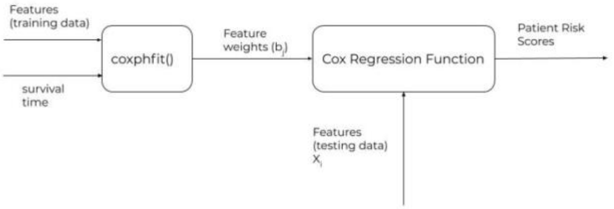


Figure 7. Cox Proportional Hazards Model Pipeline to output patient risk scores using features and feature weights.

We used this function to input the feature expression levels, the survival time, and the censored data for overall survival for each patient. Using this function on the training data, we were able to generate weights, or beta values, for each feature in a modality with respect to its impact on survival. After producing beta values for each feature, we used the Cox regression equation, Figure 6, on testing data to loop through the features, X, and beta values, to generate a risk score per patient. We then averaged the risk scores using decision-level integration for all of the modalities to get an overall risk score per patient. We then found the median of the overall patient risk scores and separated the patients into high and low risk groups based on their presence above or below the median, respectively. We then input the survival times into the high and low risk groups and using an empirical cumulative distribution estimate (Kaplan-Meier estimate), we output an approximation of survival probability.

6 Performance Metrics

To evaluate how well our assigned risk scores correlate with actual test data survival times, the output risk scores were separated into high and low risk groups by the median risk score. Each group can then be plotted in terms of survival probability estimate via a Kaplan Meier survival curve. Kaplan Meier plots are commonly used to visualize separation of groups in terms of their survival times. To implement this, we used the ecdf() function in MATLAB to estimate the survival curves of high and low risk groups. In addition to visual separation, a log-rank test was used to evaluate if the separation between the two groups was actually significant.

Another widely adopted metric to evaluate the performance of survival models is the C-index. It is equivalent to the ROC curve measuring accuracy in many binary classification models (Huang et al). In the context of our Cox survival model, it can be interpreted as the fraction or percentage of all pairs of patients whose predicted risk scores correctly correspond to the true survival time. For example, if patients with higher survival times are given a higher risk score, then the C-index of the model will be higher (F. E. Harrell, 1996). In our model, C-index was calculated in MATLAB using inputs of survival times and risk scores for patients. A C-index of 0.5 indicates a predictive capability of slightly better than a random classifier (Huang et al). State of the art Cox survival models integrating genomic data have reported C-indices higher than 0.7 to indicate a good model.

7 Robust Modeling Parameter Selection

While our Cox proportional hazards model does not have hyperparameters, the beta values, or weights, for each feature as calculated by coxphfit() are considered learnable parameters. These beta values are learned by our Cox model during the training process and optimized over 4 folds of cross validation. Each combination of validation and training data produces slightly different beta values which affect model performance. Cross validation is used to compare combinations in order to optimize the feature coefficients. The combinations of data that optimized these parameters the best, therefore producing higher C-indices, can be seen in Figure 8.

RESULTS OF FINAL MODELING AND DATA ANALYSIS

After all models were trained, high and low risk groups were visualized with Kaplan Meier plots and their performance was evaluated with log-rank t-test and c-index. The values for these performance metrics for each model through cross validation can be seen in Figures 10 and 11. A heat map of model performance across these folds can be seen in Figure 8. According to the heat map, miRNA and miRNA + mRNA models performed slightly better than a random classifier about half of the time, while mRNA model performance was slightly less. All 3 models performed better during fold 1, 2, and 4, while performing worse during fold 2. The miRNA model scored the highest model performance out of the 3 models during fold 3, and on average performed best out of all 3. Ultimately, for all of our models, these C-index values indicate a low predictive power.

In all 3 models, poor separation of high and low risk groups can be observed in Figures 9, 10, and 11. In addition, log rank p-values for all models indicate that the separation between high and low risk groups was not significant. Interestingly, the mRNA model was able to produce significant separation in fold 2, indicated by a p-value of 0.0076 (Table 3). Based on log-rank separation, it can be concluded that the integrated model exhibited no significant superiority to individual modalities by separating risk scores in terms of survival times.

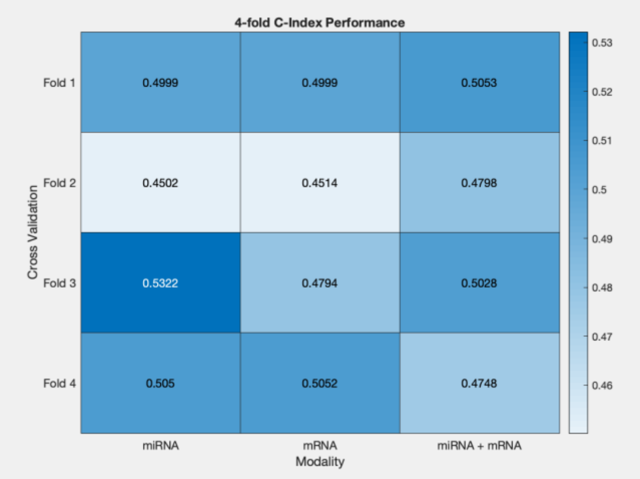
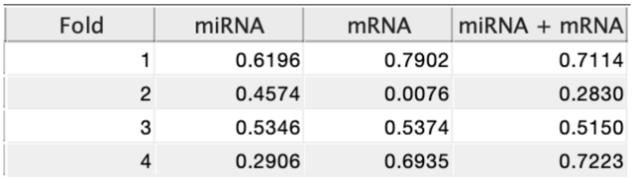
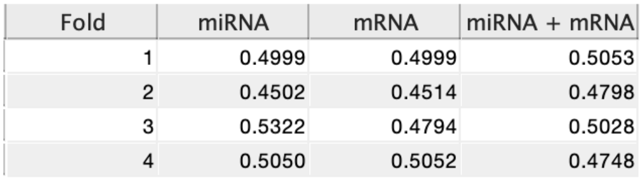


Figure 8. Heatmap of c-index values of each model for a range of combinations of testing and training data. The x axis displays models including miRNA, mRNA, and integrated miRNA and mRNA. The y axis displays cross validation folds. The miRNA only model scored the highest performance on fold 3 of cross validation.



Fold	miRNA	mRNA	miRNA + mRNA
1	0.6196	0.7902	0.7114
2	0.4574	0.0076	0.2830
3	0.5346	0.5374	0.5150
4	0.2906	0.6935	0.7223

Table 3. Log Rank p-values values are shown for each model across 4 folds of cross validation



Fold	miRNA	mRNA	miRNA + mRNA
1	0.4999	0.4999	0.5053
2	0.4502	0.4514	0.4798
3	0.5322	0.4794	0.5028
4	0.5050	0.5052	0.4748

Table 4. C-index values are displayed for each model across 4 folds of cross validation.

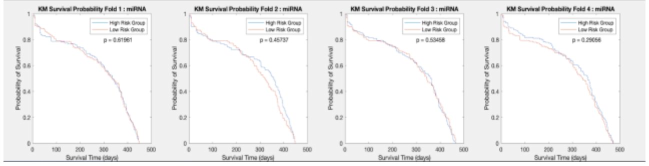


Figure 9. Separation of high and low risk groups visualized Kaplan Meier plot for miRNA model across 4 folds of cross validation.

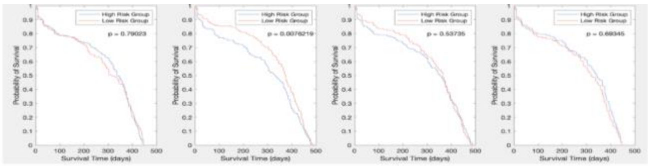


Figure 10. Separation of high and low risk groups visualized with Kaplan Meier plot for mRNA model across 4 folds of cross validation.

8 Modeling Results

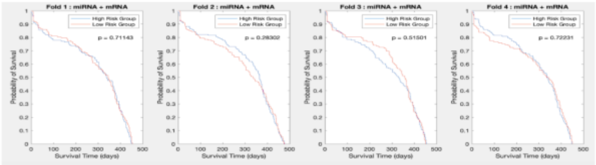


Figure 11. Separation of high and low risk groups visualized with Kaplan Meier plot for miRNA + mRNA integrated model across 4 folds of cross validation.

9 Quantitative Data Analysis Results

Top performing metrics of our integrated model for cross validation and external validation can be seen in Table 5. Model performance among both validation and testing is oddly identical. Further analysis may be needed to determine how this may be possible. Overall, the integrated model did not have good separation indicated by a top performance log-rank p-value of 0.2830 and was slightly better at predicting survival risk than random chance.

Dataset	Test	Validation	Training
Sample Size	254	233	773
C-Index	0.5053	0.5053	n/a
P-Value	0.2830	0.2830	n/a

Table 5. Sample size of testing, validation, and training data sets, including top performing metric for testing and training across 4 folds of cross validation.

10 Predictability of Model

Cross-validation, through model training, was graphed against external validation, C-index, to show the predictability of the model, in Figure 12. As training performance improved, testing performance also improved.

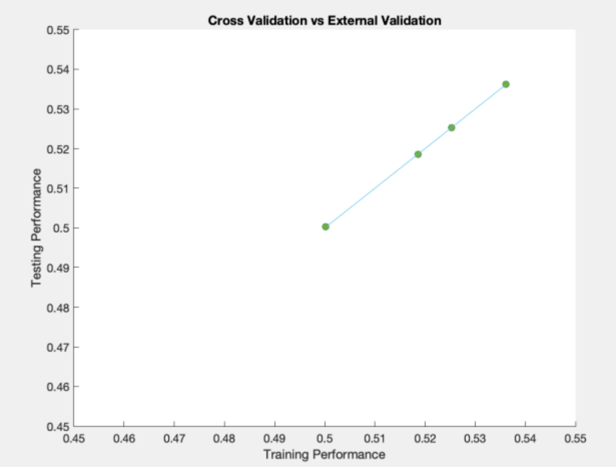


Figure 12. Comparison of c-index performance of during cross validation to performance during external validation, for the miRNA model.

11 GUI Mockup

A GUI mockup that was constructed to give users the ability to select which modalities they want integrated into the model can be seen in Figure 13. Users can repeatedly train these models and view results as a table of risk scores assigned to each patient. Users of this interface may include doctors or physicians, and they will benefit from viewing these risk scores in order to personalize treatment plans for breast cancer patients, who are highly diverse.

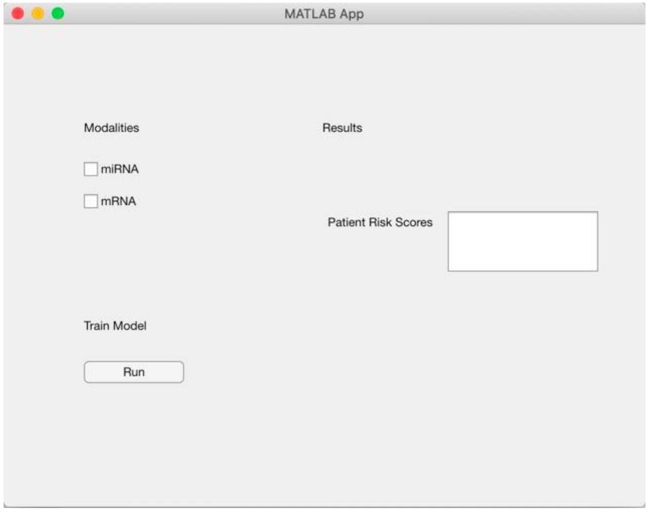


Figure 13. GUI mockup of our model for decision-level integration of multi-omics data for survival risk prediction of breast cancer patients.

INTERPRETATION, NOVELTY, AND CONCLUSION

12 Discussing and Interpreting Results

Survival modeling was performed using two integrated modalities: miRNA and RNA sequencing. These modalities each had their own set of features with values representing the level of gene expression. Log-rank feature selection was used to filter out features in both modalities on a feature by feature basis. The top features selected for making predictions were ones with a large range of variability, meaning that there was a significant difference between high and low expression groups relative to the median for that particular feature. Features were removed if p > .05. For miRNA, features refer to individual miRNA identifiers. For mRNA sequencing, chosen features refer to individual genes.

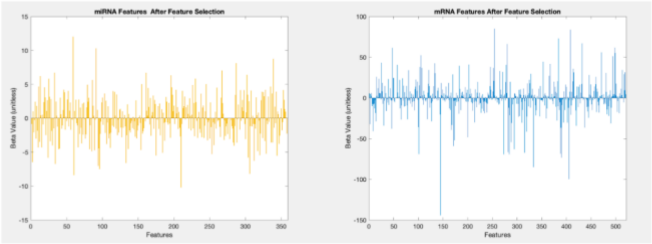


Figure 15. Beta values for selected features calculated by coxphfit() function in MATLAB.

Figure 17 reveals the distribution of beta values for each of the top features selected during log-rank feature selection for each modality. Due to the fact that our feature selection method chose over 300 features for miRNA, and over 500 features for mRNA, further examination is needed to determine the significance of individual features on survival, instead of in terms of all the selected features. The beta values of features in both modalities exhibit a wide range of magnitude, suggesting the future possibility of examining the impact of beta value magnitude on survival prediction. Further selection methods that incorporate using coxphfit() to filter out features with low beta values may have an impact on, or even increase, survival prediction performance.

13 Comparison to Literature

The state-of-the-art model from literature that we have used as comparison is the study by Liu et al on the "Multi-omics facilitated variable selection in Cox regression model for cancer prognosis prediction". As previously discussed, this study integrated mRNA expression, methylation, and copy number variation glioblastoma and lung adenocarcinoma data to predict cancer prognosis. Through modeling, they found that mRNA expression data was the most predictive of cancer prognosis. This team then developed several novel methods that fit the Cox regression model and were able to output risk scores specific to the mRNA data. Ultimately, their results had higher predictive power than previous studies.

Our model did not meet this standard set forth in literature because our C-index and p-values indicate that our model has predictive power similar to random chance. In addition, our model did not have the capabilities to predict which modality had the most predictive power. The planned pipeline for integration follows literature justification, however further refinement is needed at each step to ensure that implementation is correct.

14 Novelty of Project

The aim of this project was to complete the integration pipeline in order to predict survival time. Although this aim in itself is not groundbreaking, we attempted to distinguish our project from predecessors by incorporating temporal data into our survival modeling. Our TCGA dataset included binary, where each patient had a 0 or 1 prescribed to indicate survival or death, and temporal survival data, indicating the amount of time of observation per patient. In modeling, we analyzed both the binary and temporal data points to generate a risk score reflective of the survival probability per patient. This approach was novel because (1) classification prediction has traditionally been limited to binary data, and (2) despite our model evaluation metrics being used to examine how well our risk scores separated into high and low risk groups, the overall aim in developing risk scores was to calculate a prediction for continuous survival time, not a binary survival outcome, as previous models have done.

15 Weaknesses

There are several inherent weaknesses within the model itself that should be improved on. The first major weakness is in the method used for feature selection. Currently, log-rank is used on each feature within a modality. Log-rank separates the feature into high and low expression groups based on the presence above or below the median and calculates a p-value for that feature. Features with $p > .05$ are removed from the dataset. The reasoning behind this is because we only want to examine features with a wide distribution because it is easier to make a judgement call on their impact on survival if expression is varied across patients. The weakness in this technique is that features are only removed based on their own expression levels, without accounting for its impact on survival. The model would be stronger if survival data was incorporated into the analysis of each feature. For example, separating the feature into high and low expression groups and then graphing these groups using the Kaplan-Meier method where a survival probability is given using survival data. A p-value could then be calculated based on the survival probabilities of the high and low groups and features could be filtered out.

Another weakness in the model is that in the decision-level integration of risk scores, all modalities were weighted the same. In the model, miRNA and mRNA seq were integrated and based on literature these modalities are equally important to survival. However, the eventual goal is to integrate CNV and DNA methylation data as well to predict survival probabilities. Based on literature, we know that biologically, copy number variations are not as important as other modalities in relation to their impact on survival from breast cancer. Our model could be improved by incorporating an integration method that accounts for the dynamic range of each modality, or an appropriate machine learning algorithm that can learn which modalities have a stronger correlation with survival. Using such an algorithm, we would be able to assign a weight to each modality when averaging the risk scores to get a more biologically accurate representation of risk.

16 Future Work

Currently, our model integrates miRNA expression, mRNA sequencing, and survival data to produce a survival model and complete the full pipeline. Due to time constraints, we were unable to integrate DNA methylation and copy number variation data into the model. We performed preprocessing on both of these modalities but had difficulties with feature selection, so we elected to proceed with the other two modalities in order to complete the learning objective. The first future step would be to incorporate the full range of the four modalities into the model in order to produce a more comprehensive risk score and give us insight into how to conceptually improve our model.

Another goal that we have is to investigate why our model has such poor predictive power. Our model had both p-value and C-index scores that reflected a lack of significance in the classification of survival risk and poor predictive power. We aim to investigate the values being produced at each step in the pipeline to figure out issues with code implementation of the methods chosen in the pipeline. Based on our experience having difficulties with feature selection for DNA methylation and copy number variation, we believe that the main issue with our results is due to code implementation as opposed to glaring methodology choices in the pipeline. In addition, after examining the code, it is important to go back and reassess the chosen pipeline.

17 Conclusion

Ultimately, our model was unsuccessful in predicting patient survival risk and that random classification has about the same likelihood of success. This conclusion does not necessarily invalidate the model, it just means that refinement of feature selection methods and the full integration of modalities is needed to make a better conclusion. We have seen the importance of multi-modality integration throughout literature. We also know that biologically, survival from breast cancer is due to many different genetic and environmental factors, so a model that can account for these successfully incorporate these factors and modalities will be stronger.

Through the completion of this project, we gained familiarity with the methods needed to handle modality and survival data. Prior to this, the overall machine learning pipeline was completely foreign to us and we gained a deep appreciation for the importance of bioinformatics as a field, especially during the COVID-19 circumstances. It is clear that there is a great need for modeling big data in biomedical applications and we are grateful for the opportunity to learn about this field.

REFERENCES

[1] Attallah, O., Karthikesalingam, A., Holt, P.J.E. et al. Feature selection through validation and un-censoring of endovascular repair survival data for predicting the risk of re-intervention. *BMC Med Inform Decis Mak* 17, 115 (2017).

[2] Cheerla, A., Gevaert, O., Deep learning with multimodal representation for pancancer prognosis prediction, *Bioinformatics*, Volume 35, Issue 14, July 2019, Pages i446–i454.

[3] Ching T, Zhu X, Garmire LX (2018) Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLOS Computational* Biology 14(4): e1006076.

[4] Chow, M. L., Winn, M. E., Li, H. R., April, C., Wynshaw-Boris, A., Fan, J. B., Fu, X. D., Courchesne, E., & Schork, N. J. (2012). Preprocessing and Quality Control Strategies for Illumina DASL Assay-Based Brain Gene Expression Studies with Semi-Degraded Samples. *Frontiers in genetics*, 3, 11.

[5] DeSantis, C.E., Ma, J., Gaudet, M.M., Newman, L.A., Miller, K.D., Goding Sauer, A., Jemal, A. and Siegel, R.L. (2019), *Breast cancer statistics*, 2019. CA A Cancer J Clin, 69: 438-451.

[6] Györffy, B., Lanczky, A., Eklund, A.C. et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat* 123, 725–731 (2010).

[7] Huang, Z., Zhan, X., Xiang, S., Johnson, T., Helm, B., Yu, C., Zhang, J., Salama, P., Rizkalla, M., Han, Z., Huang, K. "SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer" in *Frontiers in Genetics*, vol. 10.

[8] Liu, C., Wang, X., Genchev, G., Lu, H. "Multi-omics facilitated variable selection in Cox-regression model for cancer prognosis prognosis prediction," 2017. *Methods*, pp. 100-107.

[9] Mayr, A., & Schmid, M. (2014). Boosting the concordance index for survival data--a unified framework to derive and evaluate biomarker combinations. *PloS one*, 9(1), e84483.

[10] Mitchel, J., Chatlin, K., Tong, L., Wang, M., "A Translational Pipeline for Overall Survival Prediction of Breast Cancer Patients by [11] Decision-Level Integration of Multi-Omics Data," *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019, pp. 1573-1580.

[12] Phan, J., Quo, C., Cheng, C., Wang, M., "Multiscale Integration of -Omic, Imaging, and Clinical Data in Biomedical Informatics*," IEEE Rev Biomed Eng.* 2012 ; 5: 74–87.

[13] Sun, D., Wang, M. Li, M., "A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data" in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 03, pp. 841-850, 2019.

[14] Rapaport, F., Khanin, R., Liang, Y. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 14, 3158 (2013). https://doi.org/10.1186/gb-2013-14-9-r95

[15]
https://www.ncbi.nlm.nih.gov/pubmed/8668867