PREDICTING DISTANT METASTASIS SURVIVAL FUNCTION IN BREAST CANCER USING COBRA

A Project Report Submitted
for the Course

MA691 Advanced Statistical Algorithms

Asst. Prof. Arabin Kumar Dey

by

Adit Jain (180102003)
Kousik Rajesh (180101094)
Eklavya Jain (180123065)
Drishti Chouhan (180101021)
J. Neeraja (180123017)

to the

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, INDIA

November 2021

Abstract

Metastatic Breast Cancer can develop when breast cancer cells break away from the primary tumor and enter the bloodstream or lymphatic system. Metastases are the primary cause of death of cancer patients, and improving the means of foretelling their development is a major goal of current clinical research. Here we analyse the TRANSBIG dataset and obtain the gene signatures of 198 breast cancer patients along with their characteristics. We predict the survival function for distant metastasis of lymph-node-negative primary breast cancer using multiple regression models and combining their predictions using COmBined Regression Alternative (COBRA). We compare our findings with the existing ensemble learning technique, Random Survival Forest, and reported the accuracy score for both methodologies. We report an improved accuracy of 0.78 in comparison to 0.57 offered by the Random Survival Forest. We also present our pipeline to predict the time for distant metastasis, which can be further extended to apply COBRA Technique on any healthcare dataset.

1 Literature Review

1.1 TRANSBIG dataset

Gene signatures of cancer patients have been extensively used in clinical research to predict distant cancer metastasis. There have been regressive attempts to accurately predict diagnosis and treatment for these patients by analysing appropriate gene expression data. One such instance is the research corpus that uses the TRANSBIG dataset(Gene expressions of 198 systematically untreated patients) to predict distant metastases in lymph node-negative (N-) breast cancer patients. Several methodologies have been used overtime to improve the accuracy of predictions and understand the genetics that influences these prognostics. The research pipeline consists of the following steps:

- Correctly identifying the appropriate gene signature that is relevant to breast-cancer.
- Validating the performance of the 76-gene signature to support the growing evidence that gene expression signatures are of clinical relevance, especially for identifying patients at high risk of early distant metastases.
- Performing a better gene-selection, eg. feature selection to select 12 genes which showed a higher area under the receiver operating characteristic curve of 0.730 compared with 0.579 yielded by previously reported 76 genes.

1.2 Identifying 76-gene signature for predicting metastatis

22,000 gene-transcripts from the RNA 286 lymph-node-negative patients who had not received systemic treatment, was analysed. In a training set of 115 tumours, 76-gene signature consisting of 60 genes for patients positive for oestrogen receptors (ER) and 16 genes for ER-negative patients was identified. This signature showed 93% sensitivity and 48% specificity in a subsequent independent testing set of 171 lymph-node-negative patients. The methodology adopted was as follows:

1. 17,819 genes were 'present' in two or more samples and were eligible for hierarchical clustering.
2. To identify genes that discriminated patients who developed distant metastases from those remaining metastasis-free for 5 years, two supervised class prediction approaches were used. The first being randomly building a training dataset of 80 and testing dataset of 206 patients. Kaplan-Meier survival curves for the two sets were examined to ensure that there was no significant difference. The second approach was to randomly allocate patients to one of the two subgroups based on Er status.
3. Univariate and multivariate analyses with Cox's proportional-hazards regression were done on the individual clinical variables with and without the gene signature. The hazard ratio and its 95% CI were derived from these results.
4. After 5 years, absolute differences in distant-metastasis-free and overall survival between the patients with the good and poor 76-gene signatures were 40% and 27%, respectively. [Wang et al., 2005]

1.3 Validating the prognosis signature

A study conducted by TRANSBIG to validate the 76-gene prognostic signature for distant metastasis and compare the results with clinical risk assessment. The results of this experiment showed that there was a

strong time dependence leading to an adjusted hazard ratio of 13.58 (1.85-99.63) and 8.20 (1.10-60.90) at 5 years and 5.11 (1.57-16.67) and 2.55 (1.07-6.10) at 10 years for time to distant metastasis and overall survival, respectively. [Desmedt et al., 2007]

1.4 Correlation-centred gene selection to reduce gene-set

A robust feature-selecting strategy with a correlation-centred approach to select minimal gene sets was developed using a multiple logistic regression model. This method selected 12 genes which showed a higher area under the receiver operating characteristic curve of 0.730 compared with 0.579 yielded by previously reported 76 genes. In conclusion, a smaller gene-set that has higher predictive capabilities and increased applications in cancer-treatment has been identified. There have been various other methods explored as described below which can be used to improve the ROC. [Hikichi et al., 2020]

1.5 Different machine-learning and deep-learning models on TRANSBIG

One of the many approaches was an Ensemble machine learning technique [Zakharov and Dupont, 2011] that combines multiple classifiers to improve performance by choosing feature subsets and learning predictive models.
DeepProg [Poirion et al., 2019], is another novel ensemble framework of deep-learning and machine-learning approaches that robustly predicts patient survival subtypes. DeepProg is highly accurate, with a c-index of 0.68-0.73.

Other papers [Djebbari et al., 2008] , [Jiao and Vert, 2015], [Mirsadeghi et al., 2019] have also analysed the TRANSBIG dataset and made significant contribution to improve predictive performance but none have explored the impact of Combined Regression on the same.

1.6 COBRA

Multiple initial estimators of the regression function can be combined, instead of building a linear or convex optimized function over a selection of basic estimators. This can be used as a collective indicator of the proximity between training data and test results. This local-distance approach is fast and efficient, which performs asymptotically in the L^2 sense as the best combination of the basic estimators in the collective. The increased accuracy and reduced Brier-score of the COBRA strategy can help achieve better classifier performance on the TRANSBIG dataset. [Biau et al., 2016]

2 Brief overview

To predict the survival function for distant metastasis we initially create indicator variables for time instances ranging from 0 to 8000 days with a step size of 100 days, and add them as separate columns in the dataset. Each of these columns denote if metastasis has occurred until a certain point. These indicator variables are binary variables and are estimated by applying classification models on the training data. Subsequently, these columns are combined to estimate the survival function. We train different classification models such as Support Vector Classifier, K Nearest Neighbours Classifier, Logistic Regression, Gaussian Naive Bayes

and Linear Discriminant Analysis each of which output a value between 0 and 1 denoting the probability of the indicator variable being true. For each indicator variable, we then combine these predictions through a Combined Regression (COBRA) Strategy. Our COBRA technique is soft in nature, meaning rather doing a hard match for each data point of the the test dataset, it combines the predictions using an epsilon threshold. This epsilon parameter used by COBRA is a measure of the selectivity of the model and we perform a grid search with brier score as the scoring function to choose the optimal value of epsilon. We benchmark our results by comparing the average accuracy obtained by COBRA over all indicator functions with the average accuracy obtained by logistic regression, Random Survival Forest and a simple average over models. We achieved an accuracy of 0.78 in comparison to Random Survival Forest's accuracy of 0.57.

3 Methodology

3.1 Dataset

The dataset for this project was obtained from the original study conducted by TRANSBIG to test the possibility of predicting distant metastases in lymph node-negative breast cancer patients using 76-gene signature's ability. Gene Expression profiling was done from frozen samples of 198 node negative untreated patients. [1] The dataset has three aspects, gene expression, clinical data and genomic risk which are obtained independent to each other. The gene expression contains expressions data for 22283 genes, out of which only 76 are relevant for breast cancer metastatis prediction, which are extracted. The clinical data contains various attributes of the patient, importantly the age, size of tumor, time to distant metastasis and censoring variables for the same. The genomic risk are different diagnostic risks including the NPI Score, score by Adjuvant Online (AOL) and Veridex, and Clinical Risk Group by St. Gallen Criteria. These datasets are extracted and combined for each of the 198 patients to create a single dataset.

3.2 Train-Test split

For performing our model analysis we split the dataset into training and testing data using a split ratio of 0.66 and doing a random split. This gives us 118 data points in the trainset and 70 in the test set. This split has been taken because the initial number of data points is itself very small, and to get some statistically significant results we need enough test data points.

3.3 Creating intermediate variables

We create 80 indicator variables which act as intermediate variables for time points ranging from 0 days to 8000 days that denote the event that metastasis has taken place. This indicator variables are spaced at distance of 100 days each.

The equation for creating an indicator variable at time point T is $I_{t<T} = 1$ if time to distant metastasis is less than the value T, otherwise 0.

The intermediate variables that were created are used later on used to approximate the cummulative survival function since predicting the survival function is otherwise not possible. The indicator functions are

[1] The dataset is publically available online at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse7390

predicted using the approach highlighted below.

3.4 Estimating intermediate variables

We use estimators especially COBRA (explained below) to predict each of these indicator variables. When working on the test data we use the trained model to generate predictions in the form of probabilities for each indicator variable. We then combine these estimated indicator functions to form the survival function.

3.5 Calculation of survival function

The survival function is calculated using the indicator variables that we predicted in the last step. The equations for calculation are detailed below

$$S(t) = Pr\{T \geq t\} = 1 - F(t) = \int_t^\infty f(x)dx$$

Here $F(t)$ is the cumulative probability that metastasis has occurred, in our case the predictied probability of the indicator function at point t serves as an estimate for $F(t)$.

3.6 Combined Regression Strategy (COBRA)

3.6.1 Introduction to COBRA

COBRA which stands for Combined Regression is a method used to combine multiple weak learners. Given a set of preliminary estimators $r_1, .., r_M$, the idea behind this combining method is an unanimity concept. It creates a prediction mapping for each weak learner on the training data. These are then used while predicting on the test data to find existing datapoints that are close to the considered point. The prediction corresponding to these datapoints are used to generate the final prediction by taking help of some summary metric (Mean in our implementations).

3.6.2 Overview of model

We first split the training dataset into two groups of equal sizes, and then our implementation initializes different models to predict the indicator functions. These models are trained using one of these halves and the models predict the indicator function over the other half. These prediction results are stored for each of these models and are used for predicting outcomes, these are referred to as reference-training data. For predicting the indicator outcome for an input from the test data, we do the following:

- For each machine, we find its prediction on the considered test point and we iterate through the prediction table to find ϵ close predictions. We mark these predictions for those machines
- We pick up the reference-training data points which have all machines giving ϵ close predictions (The total number of machines/models is given as α)
- Taking the mean of prediction probabilities corresponding to these data-points, which we already have since this is the training data, we get the outcome corresponding to our required test data point.

3.6.3 Constituent classification models

- **Support Vector Classifier** - The Support Vector Classification algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.
- **K Nearest Neighbours Classifier** - The KNN classification algorithm uses a distance metric to find k closest neighbors of a given test point and classifying it into the majority class.
- **Decision Tree Classifier** - The Decision Tree Classifier builds a Decision tree to classify samples into different classes.
- **Logistic Regression** - Logistic Regression which is actually a classification algorithm generates a decision boundary between two classes based on the training data.
- **Gaussian Naive Bayes** - Gaussian Naive Bayes is a model based on Bayes' Theorem and has a strong assumption that predictors should be independent of each other.
- **Linear Discriminant Analysis** - LDA is closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data.

  We use these models as a constituent model with the default hyperparameters provided by sklearn.

3.6.4 Metrics for evaluation

- **Brier Score** - Brier score is a strictly proper scoring rule that measures the accuracy of probabilistic predictions. The Brier score is calculated as follows: Across all items $i \in 1..N$ in a set of N predictions, the Brier score measures the mean squared difference between: the predicted probability assigned to the possible outcomes for item i and the actual outcome.

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

  Here f_t is the probability that was predicted, o_t is the actual outcome of the event at instance t (0 if it does not happen and 1 if it does happen) and N is the number of predicting instances.

- **Accuracy** - The accuracy score is the percentage of labels predicted that match the corresponding set of true labels in the evaluating dataset.

3.6.5 Finding optimal epsilon

Epsilon as discussed earlier is used to control the selectivity of the COBRA model, this epsilon can take real values within the range of 0 to 1, Values near to 0 signifying a highly selective model. This epsilon can be hard coded to a particular value but that prevents the possibility of improving the model. And so to find the optimal value, cross validation is done on the training set using a continuous range of different epsilons, this technique is often referred to as grid search in optimization literature. For each value of epsilon, we calculate the brier score for a subset of the training data. The epsilon corresponding to the minimum brier score is taken to be the optimal epsilon is then used to run the final model on the test data-set for each of the indicator functions to finally obtain the survival function shown in Results.

3.7 Random Survival Forest

Random Survival Forests are a class of Random Forests which were introduced for specific analysis of right censored data and are used in predicting different kinds of survival metrics, including mortality, survival function and hazard function. A basic overview of the random survival function can be given as:

- Drawing B bootstrap samples from the original data
- Growing a survival tree for each of the bootstrap sample, at each node of the tree p candidate variables are selected. The node is then split using the candidate variables that maximizes the survival difference between daughter nodes.
- The tree is grown to full size until the constraint that any terminal node doesn't have less than $d_0 > 0$ deaths.
- A CHF is calculated for each tree and an average of all CHFs is taken as the ensemble CHF
- Using out-of-bag data a prediction error is calculated

4 Implementation and Results

The experiments were conducted in python with the help of libraries like Scikit-Learn, Scikit-Survival, Numpy, Pandas, and Matplotlib. Scikit-Learn was used to train classification models like Support Vector Classifier, K Nearest Neighbors, Decision Tree Classifier, Logistic Regression, Gaussian Naive Bayes, and Linear Discriminant Analysis. Furthermore, Sklearn metrics like accuracy, brier score, and RMSE were used.

The experiment is executed on a system having the following configuration:
(i) Intel® Core™ i5-8250U CPU @ 1.60GHz x 4, (ii) Ubuntu 20.04 LTS OS, and (iii) 12 GB Memory.

The 76 gene features along with the age, size, NPI score and Adjuvant Online (AOL) were considered as the explanatory variables, and the constructed indicator variables for time to distant metastasis as the response variable.

The predictions of the indicator variables were extended to calculate the survival function, and further the cumulative hazard function. The graphs for both the survival function and the cumulative hazard function are plotted for our native implementation of COBRA as well as Random Survival Forest present in Scikit-Survival library. Figure 1 plots the cumulative hazard function of 5 patients against time in days for our native implementation of COBRA. Figure 2 plots the same function calculated using Random Survival Forest. Figure 3 plots the survival function of the same 5 patients against time in days for our native implementation of COBRA. Figure 4 plots the same function calculated using Random Survival Forest.

In Figure 5, we compare the accuracies offered by various models. In figure 6 we compare the brier score corresponding to each model. We see that our method (using COBRA) provides the best accuracy and the least brier score among all the machines, even Random Survival Forest.

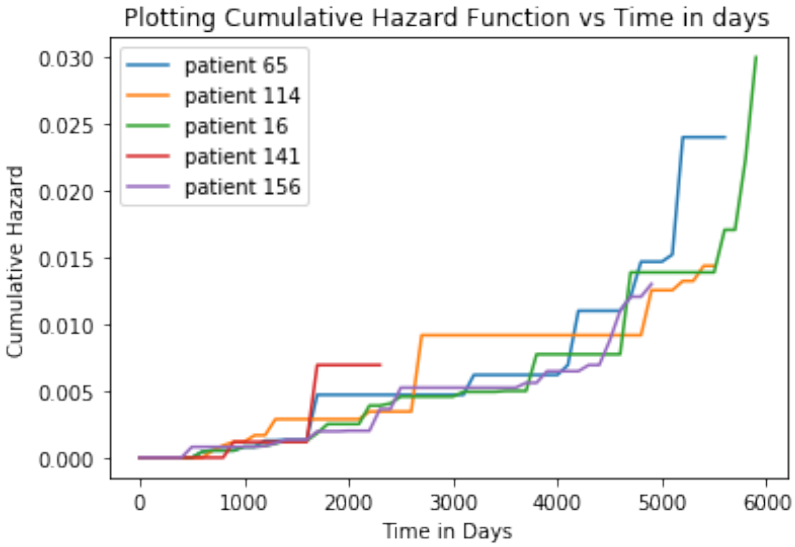

Figure 1: Cumulative Hazard Function of 5 patients on applying COBRA

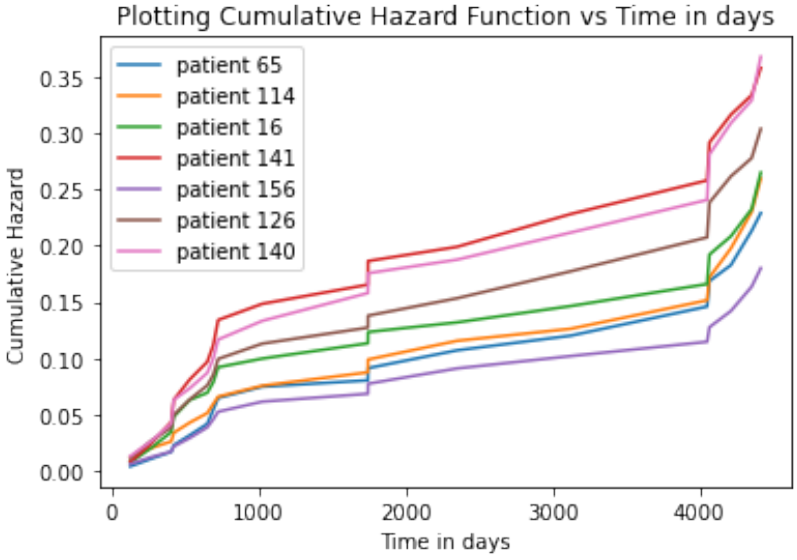

Figure 2: Cumulative Hazard Function of 5 patients on applying Random Survival Forest

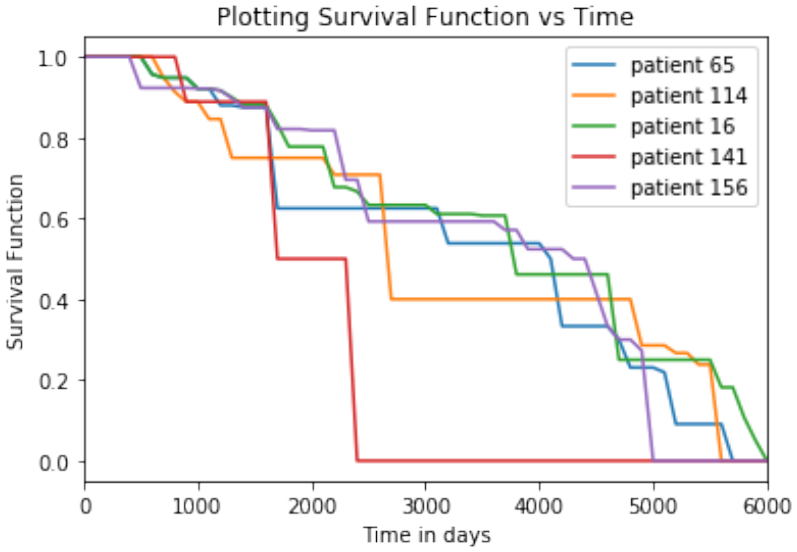

Figure 3: Survival Function of 5 patients on applying COBRA

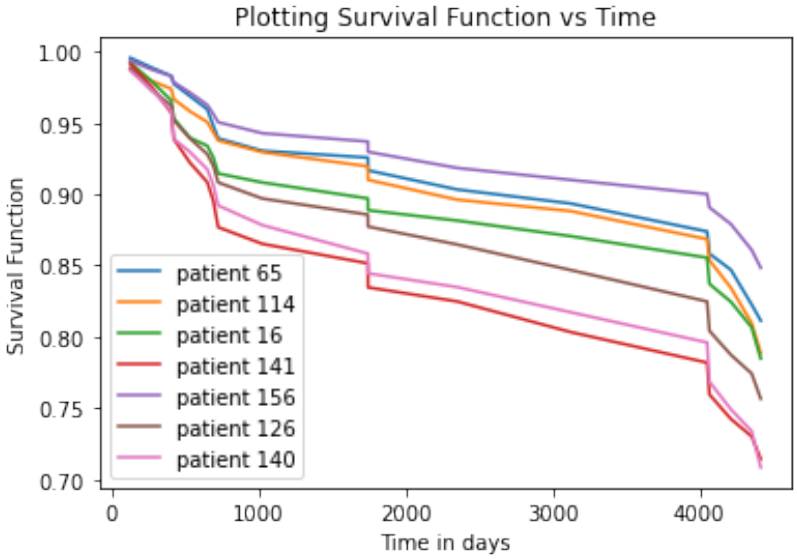

Figure 4: Survival Function of 5 patients on applying Random Survival Forest

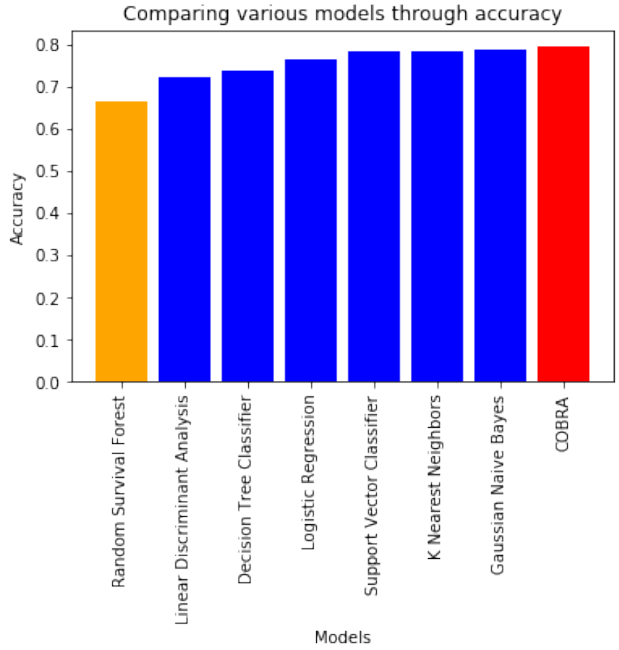

Figure 5: Comparison using Accuracy of various models used in our Analysis

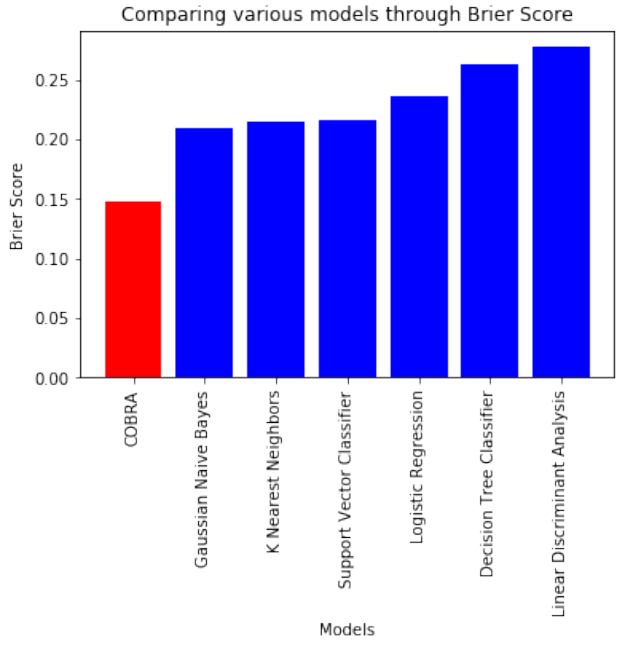

Figure 6: Comparison using Brier Score of various models used in our Analysis

References

[Biau et al., 2016] Biau, G., Fischer, A., Guedj, B., and Malley, J. D. (2016). Cobra: A combined regression strategy. *Journal of Multivariate Analysis*, 146:18–28.

[Desmedt et al., 2007] Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., et al. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series. *Clinical cancer research*, 13(11):3207–3214.

[Djebbari et al., 2008] Djebbari, A., Liu, Z., Phan, S., and Famili, F. (2008). An ensemble machine learning approach to predict survival in breast cancer. *International journal of computational biology and drug design*, 1(3):275–294.

[Hikichi et al., 2020] Hikichi, S., Sugimoto, M., and Tomita, M. (2020). Correlation-centred variable selection of a gene expression signature to predict breast cancer metastasis. *Scientific reports*, 10(1):1–8.

[Jiao and Vert, 2015] Jiao, Y. and Vert, J.-P. (2015). The kendall and mallows kernels for permutations. In *International Conference on Machine Learning*, pages 1935–1944. PMLR.

[Mirsadeghi et al., 2019] Mirsadeghi, L., Banaei-Moghaddam, A. M., Beh-Afarin, S. R., Hosseini, R. H., and Kavousi, K. (2019). A post-method condition analysis of using ensemble machine learning for cancer prognosis and diagnosis: a systematic review.

[Poirion et al., 2019] Poirion, O. B., Chaudhary, K., Huang, S., and Garmire, L. X. (2019). Multi-omics-based pan-cancer prognosis prediction using an ensemble of deep-learning and machine-learning models. *medRxiv*, page 19010082.

[Wang et al., 2005] Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679.

[Zakharov and Dupont, 2011] Zakharov, R. and Dupont, P. (2011). Ensemble logistic regression for feature selection. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 133–144. Springer.