Bayesian Networking for Prediction of Breast Cancer Survivability

Isha Mukundan[1,+], Sofia Polychroniadou[1,+], and Kara Walp[1,+]

[1]Boston University, Department of Biomedical Engineering, Boston, MA, 02215, USA
[+]these authors contributed equally to this work

ABSTRACT

This study investigates the application of Bayesian Networks (BNs) to predict breast cancer survival outcomes using gene expression and mutation data from the METABRIC dataset. While traditional prognostic tools like the Breast Cancer Index Test analyze a limited number of genes, this project leverages modern genomic technologies to develop a more comprehensive predictive model. We implemented a mixed-distribution Bayesian network incorporating gaussian gene expression nodes, binomial gene mutation nodes, and a Bernoulli survival outcome node. The network topology was learned using a hill climbing algorithm with Bayesian Information Criterion optimization, and parameters were estimated through Maximum Likelihood Estimation. Model validation through 5-fold cross-validation yielded an average Area Under Curve (AUC) score of 0.53, falling short of the target 0.7 threshold. Comparative analysis with existing literature revealed that this performance aligns with other machine learning approaches using this dataset, suggesting that these specific gene expression profiles alone may not be sufficient predictors of breast cancer survival. Our findings indicate that while the Bayesian Network effectively captured available genetic relationships, the fundamental predictive power of gene expression data for survival outcomes may be inherently limited, highlighting the potential need to integrate clinical variables for more accurate prognostic modeling.

1. Introduction

1.1 Background

Breast cancer remains one of the most significant public health challenges worldwide, with concerning statistics that highlight its prevalence and impact. In the United States, it represents the most common cancer type among women (excluding skin cancers), with approximately 30% of all female cancer cases. Current data indicates a 13% lifetime risk of developing breast cancer for women in the US, with a median diagnosis age of 62 years. The disease's impact is particularly evident in its mortality rates, being the second leading cause of cancer death in women with a 2.5% mortality rate, though this rate has shown encouraging improvement with a 42% decline from 1989 through 2021.[1]

While significant medical advances have improved outcomes, several critical challenges persist. Most notably, about 50% of breast cancer patients have no known risk factors beyond age and biological sex, creating a substantial gap in our understanding of disease development and progression. This gap is particularly problematic in low-HDI (Human Development Index) countries, where limited access to diagnostic resources contributes to higher mortality rates. These challenges underscore the urgent need for more sophisticated analytical approaches that can better predict and understand breast cancer outcomes.[2]

The traditional clinical approach to breast cancer prognosis, exemplified by tools like the Breast Cancer Index (BCI) Test, while valuable, has limitations. The BCI Test analyzes only 11 genes to generate predictions about severity, longevity, and treatability. However, with modern genomic technologies capable of measuring expression levels for hundreds of genes, there's an opportunity to develop more comprehensive predictive models that could uncover previously unknown genetic interactions and prognostic indicators.

1.2 Bayesian Networks

Bayesian Networks (BNs) emerge as a powerful probabilistic framework uniquely suited to analyzing complex biological systems. These directed acyclic graphs model probabilistic relationships between variables, offering a sophisticated approach to understanding the intricate interactions between genetic markers, mutations, and clinical outcomes in breast cancer.

The power of Bayesian networks in gene expression analysis lies in their ability to address multiple critical research challenges simultaneously. Unlike traditional statistical methods, BNs can effectively capture complex variable dependencies while maintaining interpretability. They excel at naturally handling uncertainty and missing data—a persistent challenge in genetic research—and provide a robust framework for both predictive and diagnostic reasoning. Moreover, these networks can seamlessly integrate diverse data types, from continuous gene expression values to discrete clinical outcomes, creating a comprehensive analytical approach.

In the specific context of breast cancer analysis, Bayesian networks provide transformative insights that extend beyond conventional statistical methodologies. By modeling probabilistic relationships between genetic markers and cancer outcomes, researchers can uncover previously unidentified genetic interactions, pinpoint key prognostic factors overlooked by simpler linear models, and generate interpretable predictions that can inform clinical decision-making.

Our research model strategically harnesses these capabilities by synthesizing three critical research dimensions. We integrate genetic markers and their probabilistic relationships with breast cancer outcomes, examine gene expression patterns and survival dependencies, and explore complex interactions associated with cancer progression and prognosis.

Using the METABRIC dataset[3], a comprehensive repository containing genetic information from 1,980 primary breast cancer samples, our model transcends traditional clinical variables to provide deeper biological insights. This approach represents a significant methodological advancement in understanding breast cancer genetics and their relationship to patient outcomes.

Although our work establishes a foundational framework for potential personalized medicine by identifying relationships between genetic markers and survival, we acknowledge that applying these findings in clinical practice extends beyond the immediate scope of this research. Translating these insights into actionable clinical strategies represents a critical long-term objective for subsequent research initiatives.

2. Methods

2.1 Data Preprocessing

The dataset[3] used in this study contains breast cancer gene expression profiles, and is divided into three main categories: clinical data, gene expression profiles, and gene mutation data.

The first 31 features consist of miscellaneous clinical data; such as age, tumor class, tumor size, therapeutic administered, etc. These features were not used since they are irrelevant to our question. The next 489 features consist of gene expression profiles, where each feature is a gene and its value is its Z-score. Z-score is a measure used to compare gene expression levels of a specific gene to the average observed expression across all genes. The interpretation of a Z-score is relatively straightforward–if the value, Z, is less than zero, the gene was expressed less than average, and vice versa. Z-scores typically range from negative two to positive two. Implementing a Bayesian network with this data that aims to predict cancer survival comes with the opportunity of uncovering undiscovered pathways which affect survival; upon other biological insights. The last 173 features consist of gene mutation profiles; where similarly to the expression profile data, each feature is a gene; and the value of the feature is a classification of either zero if there is no mutation, or the name of the mutation if there is a mutation. Given that the specific mutations are unique to the gene, we discretized the mutation data from a multinomial set into a binomial set, which classified whether each gene is mutated or not.

Unfortunately, upon attempting to learn our network topology with hill climbing, we ran into the issue of very large runtime complexity. Hill climbing is a new enough method that all of the literature does not agree on its specific runtime complexity, however multiple sources seemed to agree that hill climbing has $O(n^3)$ complexity[4]. This meant that it was infeasible to use all 489 gene expression profiles, and all 173 mutation profiles.

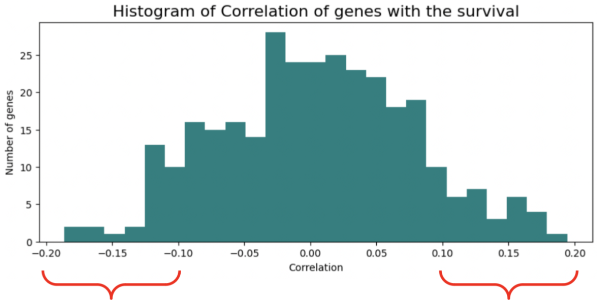

Figure 1. Histogram of correlation between gene expression value / mutation state and patient survival. All features were pruned except for ones with highly negative or highly positive correlation with death from cancer (indicated by the red brackets).

To condense our dataset, we pruned out all genes/features except for ones with highly negative or highly positive correlation with patient survival (< |0.10|, as seen in Figure 1. It should be noted that this is not a mathematically sound method of dataset pruning for two reasons. Firstly, correlation only indicates linear probabilistic dependence, meaning that this method may have filtered out genes which patient survival has strong, but nonlinear probabilistic dependence on. Secondly, aside from only considering linear dependence, our pruning method only considers the direct probabilistic dependence between gene nodes and our node of interest, patient survival. The benefit of implementing a Bayesian network is to approximate inter-node dependencies, which will hopefully cascade down to the node of interest– however our data pruning method only took into account linear dependencies between our node of interest and other nodes, meaning that we may have pruned out very important features unintentionally. For example, in Figure 2, our method could potentially prune out the “Gene C Mutation” feature if it doesn’t have correlation with the “Death From Cancer”, however in doing so, we would have pruned an integral node in the “Gene 1 Expression” to “Gene C Mutation” to “Gene 2 Expression” to “Death From Cancer” pathway, which could be a very important pathway in survival prediction. A more mathematically sound pruning method would be to prune stochastically, and run hill climbing for multiple iterations.

Following our data preprocessing, we can discuss how our model will look. Our model is a mixed distribution Bayesian network, with gaussian Gene Expression nodes (Z-score), binomial Gene Mutation nodes, and a Bernoulli node of interest which predicts death from cancer.

2.2 Building our Bayesian Network

To actually build the Bayesian network to model the relationships between the variables in our dataset, the learning process involves two primary tasks: structure learning and parameter learning.

2.2.1 Structure Learning

For a Bayesian network, structure learning involves finding the underlying direct acyclic graph (DAG) structure that encodes the dependencies between the data being modeled. In this way, the goal of structure learning is to determine and define which variables in a dataset are conditionally dependent on others and is typically done through evaluating possible graph structures and selecting the structure that best explains the data.

There are two main approaches to learn the structure of a Bayesian network: constraint-based methods and score-based methods. Constraint-based methods identify the conditional independence relationships between variables in a dataset through beginning with a fully connected undirected graph and using hypothesis testing to eliminate edges that indicate unconditional independence between variables. Directed edges are then added based on the d-separation criterion, which determines if conditioning on a variable blocks all paths between two others, implying conditional independence[5]. Score-based methods for structure learning evaluate various possible BN configurations and assign a score to each network based on how well it explains observed data. The structure that either maximizes or minimizes the score, depending on the chosen scoring function, is selected as the best model. As the number of possible network structures grows exponentially with the number of nodes, exhaustive searches become infeasible even for small networks, so to address this, score-based methods typically rely on heuristic search techniques to more efficiently explore the space.

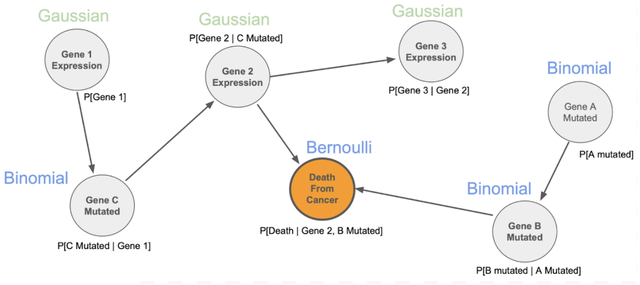

Figure 2. Schematic of a potential simplified Bayesian network topology of our model. Gene expression nodes are gaussian and contain the Z-score of their specific gene, Gene Mutation nodes are binomial and contain whether or not the gene is mutated, and the Death From Cancer node (the node of interest) gives a probability of death from cancer.

We chose a score-based method, specifically the hill climbing algorithm, to learn the Bayesian network (BN) structure, as it provides advantages over constraint-based methods. Constraint-based methods can struggle with detecting conditional independencies, as they are sensitive to errors in hypothesis testing. Additionally, since these methods rely on the d-separation criterion to determine edge directions, they may fail to assign directions to all edges, potentially leaving key dependencies undetermined. Hill climbing was selected because it strikes a good balance between computational efficiency and model quality[6]. The algorithm begins with a random initial structure and iteratively refines it by making small, incremental changes. Each change is evaluated using a scoring function, and the algorithm continues until it reaches a local maximum, where further changes do not improve the model’s score[6].

The scoring function used in our study was the Bayesian Information Criterion (BIC), which is a scoring function that looks to balance goodness of fit with model complexity[7]. The BIC score is defined as:

$$BIC = kln(n) - 2ln(\hat{L}) \tag{1}$$

Where k is the number of parameters estimated by the model, n is the number of data points in the dataset, and $\hat{L}$ is the maximum likelihood of the model defined as Equation 2.

$$\hat{L} = p(x|\hat{\theta}, M) \tag{2}$$

Based on the definition of BIC, a better model will be one that has a lower BIC score, which can be shown in Equation 1, where BIC penalizes more complex structures (those with a larger k) by incorporating a penalty term to the likelihood score. This method of scoring can help to combat overfitting as the penalty placed on complex networks lead the algorithm towards picking lower complexity models that are able to better generalize to unseen data.

2.2.2 Parameter Learning

Once the structure was learned, the next step was to estimate the parameters of the BN, specifically the conditional probability distribution (CPD) for each node, which quantifies the likelihood of a node's value given the values of its parent nodes. We used Maximum Likelihood Estimation (MLE) to estimate these parameters, as it maximizes the likelihood of the observed data without incorporating prior beliefs. While Maximum A Priori (MAP) is another approach that combines observed data with prior distributions, we chose MLE due to our lack of background regarding prior information about these probability distributions. MLE seeks to find the parameter values that maximize the likelihood of the data given the learned network structure. Mathematically, it aims to maximize the likelihood function:

$$\hat{\theta} = argmax_{\theta} P(D|\theta, S) = \prod_{i=1}^{n} P(X_i|pa(X_i), \theta) \tag{3}$$

Where θ is the model parameters, D is the data, and S is the structure of the BN. In this way, with MLE we want to choose the parameters of our model that will maximize the likelihood of our data given the network structure and parameters, which due to the nature of BNs looks specifically at the probability of a node given the parents of that node as seen in Equation 3.

2.3 Model Validation

To evaluate the performance and generalizability of our Bayesian Network model, we employed k-fold cross-validation as a method of model validation. This technique helps mitigate the risk of overfitting, which can occur when a model is too closely fit to the training data and performs poorly on unseen data. By using k-fold cross-validation, we ensure that the model's performance is robust and generalizes well across different subsets of the data.

2.3.1 K-Fold Cross Validation

In k-fold cross-validation, the dataset is split into k subsets (or folds). The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, each time using a different fold as the test set, and the remaining folds as the training set. The results are averaged to obtain a final performance measure. We used k = 5 folds, training and evaluating the model five times (each with a different train-test split), and recorded the Area Under Curve (AUC) scores for each fold to assess performance variability.

2.3.2 Register Operating Characteristic Curve and Area Under Curve

To quantitatively understand the success of the model at predicting overall survival, we calculated the Area under the curve (AUC) for a Receiver Operating Characteristic (ROC) curve for each fold. The ROC and AUC were utilized as the metric of comparison for this case as the ROC curve is a graphical representation used to evaluate the performance of binary classifiers which align with the task of our model to predict between two classes: survival versus non-survival of breast cancer. AUC plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings[8].

The ROC curve was generated using the probabilities predicted by the Bayesian Network for each instance in the test set. The BN model predicts the probability of a patient's survival based on the gene states encoded as nodes in the network. These probabilities were then thresholded to generate binary classification predictions, which were compared against the actual survival outcomes. By comparing the predicted classifications with the true survival outcomes, we calculated the corresponding TPR and FPR. The ROC curve plots TPR vs. FPR across different thresholds to illustrate the model's discrimination ability between the two classes[8].

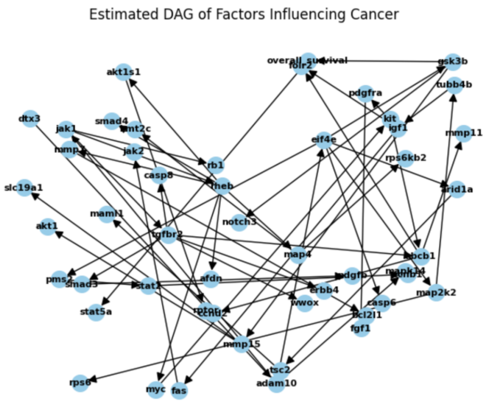

Figure 3. Schematic of our approximated Bayesian network topology. The node of interest can be seen in the top center.

The AUC provides a scalar value that summarizes the ROC curve's performance by quantifying the overall ability of the model to distinguish between the two classes. It takes a value of 0 to indicate a perfectly inaccurate test and a value of 1 to indicate a perfectly accurate test, with 0.5 indicating no discriminative ability of the model (equivalent to random guessing[8]). In the context of binary classification, such as predicting survival, AUC is particularly valuable because it considers all possible classification thresholds. This is important because survival predictions do not always have a clear-cut threshold for classification and so by calculating the value across a range of thresholds can provide a more accurate validation method of the model's capabilities.

For our case we aimed for a target AUC of 0.7, which based on prior research is considered to be an acceptable value in the field suggesting that the model has a reasonable ability to distinguish between survivors and non-survivors[9].

3. Results

3.1 Final Bayesian Network Structure

We developed a mixed-distribution Bayesian network with Gaussian gene expression nodes, binomial gene mutation nodes, and a Bernoulli survival outcome node. The final Bayesian network structure, shown in Figure 3, was learned using the hill climbing algorithm with the Bayesian Information Criterion (BIC) as the scoring function. This approach allowed us to efficiently explore the space of possible network configurations while balancing model complexity and predictive power. The learned network represents the dependencies between the variables through directed edges, indicating which genes and mutations have conditional dependencies with the survival outcome (the "overall_survival" node).

3.2 Model Validation Through AUC

To assess and validate the model's predictive capabilities and attempt to prevent the model from overfitting, we employed 5-fold cross-validation. Figure 4 presents a comparison of the calculated AUC scores across the five validation folds. We had wanted our model to achieve a target AUC of 0.7 as this value is considered the acceptable standard in the field. However, as seen in Figure 4 our generated model produced an average AUC of 0.53 across the 5 folds.

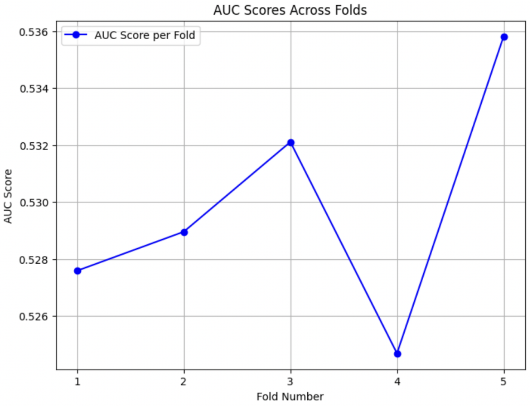

Figure 4. Area under the ROC curve (AUC) scores across 5 validation folds.

4. Discussion

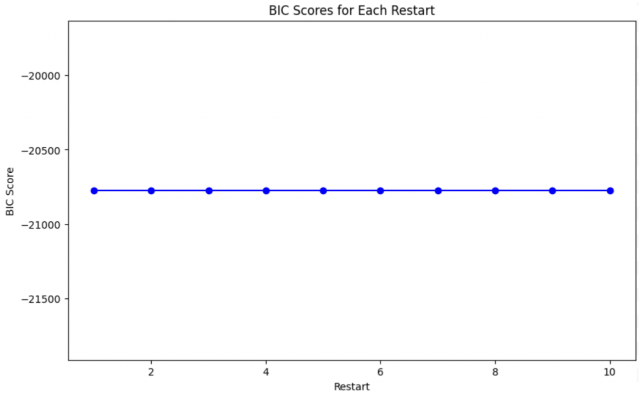

Figure 5. Bayesian Information Criterion (BIC) scores for each stochastic restart of the hill climbing algorithm to approximate the BN structure. The consistent low score may indicate that there is a very limited search space in the data.

Hill Climbing as a means of network topology estimation leaves us prone to a suboptimal model. Hill climbing initializes a directed acyclic topology randomly, and moves in the direction which optimizes the objective function. It is a greedy algorithm, however, meaning that it is susceptible to getting "stuck" at local maxima or shoulders. To reduce the chance of getting stuck at a local maxima, we reinitialized hill climbing stochastically, which as seen in Figure 5, did not improve the Bayesian information criterion. This indicates that there may be a deeper issue with the data or our preprocessing/pruning method which could explain why our results were not good.

In addition to the challenges present within our model that may have contributed to the low prediction accuracy, we hypothesize that there may be more fundamental issues with the research question itself and the dataset we are working with that could additionally cause this low performance.

If we look at the created network structure after training our BN using the gene expression data from the full genetic data set in Figure 6 compared to the structure created from cross validation where 80% of the full data was used for each fold in Figure 3 a significant loss in structure. Such a large loss in structure occurring from just removing 20% of the full data can suggest that the dependencies between the genes and survival outcomes were not well-defined or stable. This could imply that the relationships between gene expression and survival, which we initially expected to show complex and meaningful dependencies, are in fact weak or perhaps nonlinear in nature and difficult to capture using the chosen model.

The model's inability to reliably capture the structure of the data and its weak predictive performance may point to a more fundamental issue with gene expression as a predictive variable for overall survival. Specifically, the lack of strong dependencies between the variables—namely, gene expression profiles and survival outcomes—raises the possibility that gene expression alone, or this collection of genes in particular, may not be a strong predictor of survival in breast cancer. This is further supported by our review of existing literature on the same dataset which explored multiple other machine learning models on the same genetic data to predict overall survival[10]. As seen in Figure 7a, the authors found that when using only genetic data, the AUC scores ranged from 0.51 to 0.6 across five models, placing our model's AUC score within the same range. However, when the authors used clinical data to train the same models, as seen in Figure 7b, the AUC scores were significantly higher (all above the acceptable range of 0.7), indicating much better predictive performance. This comparison suggests that rather than the model itself being the cause of the lower predictive power this specific gene expression data alone may not be an adequate or reliable predictor of breast cancer survival as clinical data tested on the same exact models were able to make much better predictions. Therefore, the low AUC in our study may reflect an inherent limitation in the data—not necessarily a failure of our model.

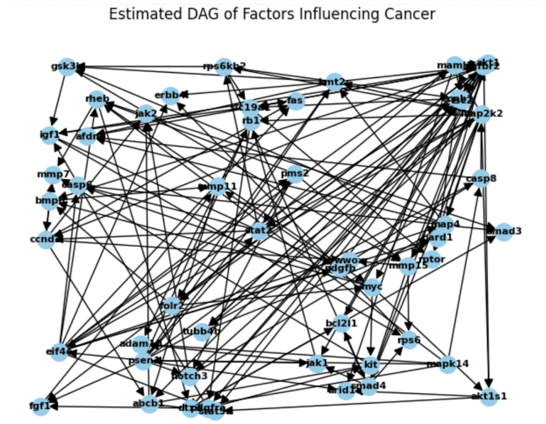

Figure 6. Approximated Bayesian network topology using 100% of the data, rather than the 80% allocated as training data through cross validation. The addition of many more dependencies show that the dependencies between nodes in our dataset may be very weak.

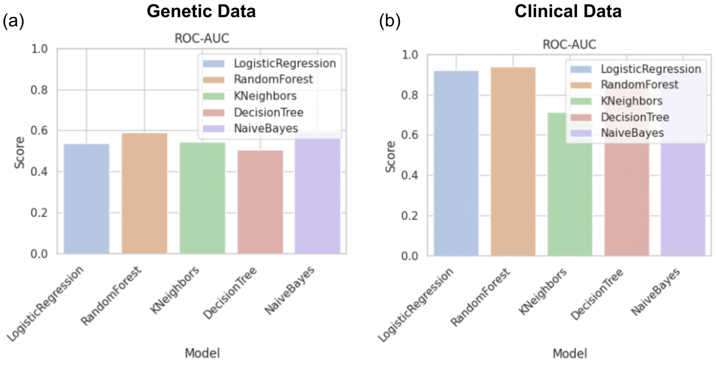

Figure 7. Survival predictor scores across many models using (a) genetic data, versus (b) clinical data (Adapted from Brzezanski et al., 2024). The performance of other models that use genetic data is similar to ours, and the increased performance of models using clinical data across models indicates that the genes contained in this dataset may not have any effect on cancer survival.

In light of these findings, we must consider the possibility that the likelihood space of our data might be fundamentally constrained. This would mean that our Bayesian network model might actually be capturing the available information optimally, but the information itself (i.e., the genetic data) may simply not contain enough meaningful variation to predict survival outcomes effectively. This could be a likely scenario as for the genetic information the performance of our model very closely matches the performance of the other model types on the genetic data. This could suggest that our BN is doing well, or at least comparable to others, at modeling the data and relationships that are present within the data but that the data itself is insufficient to answer the research question, or that the question—predicting survival from gene expression profiles alone—might not be well-suited for the dataset.

Acknowledgements

This is a project for ENG BE 562: Computational Biology (Machine Learning Fundamentals). We would like to thank Professor James Galagan, and our TF, Irene Zhou for their teaching and mentorship.

References

1. Breast Cancer Statistics | How Common Is Breast Cancer? American Cancer Society.
2. Breast Cancer | World Health Organization.
3. Alharabi, R. Breast Cancer Gene Expression Profiles (METABRIC).
4. Scutari, M., Vitolo, C. & Tucker, A. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Stat. Comput.* **29**, 1095–1108, DOI: 10.1007/s11222-019-09857-1 (2019).
5. Su, C., Andrew, A., Karagas, M. R. & Borsuk, M. E. Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Min.* **6**, 6, DOI: 10.1186/1756-0381-6-6 (2013).
6. Gámez, J. A., Mateo, J. L. & Puerta, J. M. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. *Data Min. Knowl. Discov.* **22**, 106–148, DOI: 10.1007/s10618-010-0178-6 (2011).
7. Kriegeskorte, N. Crossvalidation. In Toga, A. W. (ed.) *Brain Mapping*, 635–639, DOI: 10.1016/B978-0-12-397025-1.00344-4 (Academic Press, Waltham, 2015).
8. Nahm, F. S. Receiver operating characteristic curve: overview and practical use for clinicians. *Korean J. Anesthesiol.* **75**, 25–36, DOI: 10.4097/kja.21209 (2022).
9. Mandrekar, J. N. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *J. Thorac. Oncol.* **5**, 1315–1316, DOI: 10.1097/JTO.0b013e3181ec173d (2010).
10. Brzeżański, M. Breast Cancer Survival - Accuracy 0.86.

Additional information

The data can be accessed at https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric