Sri Lankan Journal of Applied Statistics, Vol (26-3)

Microarray Survival Analysis of Five Cancer Gene Sequences and its Application using Bayesian Additive Regression Trees (BART)

F. A. Okolie[1,4], **B. O. Fagbemigun**[1,5], and **O. J. Samson**[3,4,*]

[1]Department of Mathematical Sciences, Ajayi Crowther University, Oyo, Nigeria

[2]Department of Microbiology and Biotechnology, Ajayi Crowther University, Oyo, Nigeria

[3]Department of Microbiology, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

[4]Department of Statistics, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

[5]Black in Mathematics Association, Pretoria, South Africa

*Corresponding author: oj.samson@acu.edu.ng

Received: 08th January, 2025/ Revised: 15th October, 2025/Accepted: 01st December, 2025/ Published: 31st December, 2025

ABSTRACT

An important tool in the Genomic approach is the widely used Microarray interventions to effectively and accurately predict survival time and risk analysis of patients. This study is aimed to present and examine the risk impact of five cancer gene sequences with different high-dimensional microarray data sets, through survival analysis. GSE10300; GSE14333; GSE16446; GSE17618 and GSE20685 data sets with 16,183; 54,712; 54,739; 54,726 and 54,739 gene counts from sample sizes 44, 226, 107, 44 and 327 respectively were selected from the NCBI database to identify their risk impact. Then the classical Cox Proportional Hazard (CoxPH), Bayesian Adaptive Spline Surface (BASS) and Bayesian Additive Regression Tree (BART) models were employed to analyze each of the five data sets. Each model's prediction performance was measured using the Root Mean Squared Error (RSME). The following, R packages 'survive' for CoxPH; 'BASS' for BASS and 'bartMachine' for BART were utilized in obtaining the results. Finally, influential variable plots were shown to indicate the most important main effects and interactions

for the five microarray datasets. Varying median hazard values and traces of yellow boxes across the diagonals of the heatmap diagrams indicate multicollinearity between genes, implying misleading results by the CoxPH model, while leveraging superiority of BART model over the BASS and CoxPH models. The RSME plots successfully identify some influential genes among thousands of genes across the five real-life microarray datasets analyzed: CoxPH model shows a moderate level of predictive accuracy while the BASS model exhibits both strengths and weaknesses depending on the number of predictors used, significantly higher than CoxPH. The BART model, on the other hand, with the lowest RSME, consistently demonstrates its superior, stable and consistent predictive performance over CoxPH and BASS models. This study gives a more distinct and nuanced understanding of genes interactions and influence, as well as provide improved prognostic tools and personalized treatment strategies for cancer patients' risks and survival.
Keywords: Genes, Microarray, Survival analysis, Cox regression, BASS, BART.

1 Introduction

Cancer remains a significant public health challenge, with various forms exhibiting distinct genetic profiles and survival outcomes Mohammed et al. (2021). Understanding the molecular mechanisms that underlie cancer progression is crucial to improving prognosis and treatment strategies. One powerful tool in this genomic endeavor is the widely recognized microarray technology, which accurately predicts the survival time and risk profile of patients and allows the simultaneous examination of the expression levels of thousands of genes. By analyzing the expression patterns of cancer-related genes, researchers can identify biomarkers that predict patient survival (death or relapse) and response to therapy Simon (2003).
Naturally, the focus of Survival analysis is on the probability $S(t)$ of nonoccurence of an event by time t for all $t > 0$ while Microarray survival analysis involves correlating gene expression data with patient survival times to identify genes that significantly influence prognosis Van't Veer et al. (2002); PH Model (2021). In regression analysis, variations in the effect of a predictor are typically characterized by changes in the value of its corresponding regressor x. In addition, of paramount importance is the hazard $h(t)$, which is the time rate of the probability of the occurrence of the event, given its nonoccurence by time t. Various aspects of these functions are targets of inference with the availability of cancer data. This approach has been instrumental in uncovering genetic signatures that stratify patients into different risk groups, guiding personalized treatment plans [Ibrahim et al. (2001); Sparapani et al. (2021)].

However, regression relationships in survival data often involve complex interactions, nonproportional hazards, high-dimensional parameter spaces, and nonlinear functions of covariates, which limits traditional methods. Address-

ing these complexities in nonlinear regression relationships requires the use of various advanced statistical models and analysis techniques. For cancer data, this includes a range of methods from parametric to nonparametric approaches such as LASSO (Least Absolute Shrinkage and Selection Operator), Neural Networks, Random Forests, Boosting and Multivariate Adaptive Regression Splines (MARS), all of which are well-documented in the literature [Olaniran and Alzahrani (2023); Baesens et al. (2005); Chipman et al. (2010); Cox (1972); Francom and Sansó (2020)].

Among these popular models, the classic Cox Proportional Hazard model (CoxPH), which assesses the importance of various genes in the survival risk of patients through its hazard function, has been applied in survival analysis over the years [Gelman et al. (1995); Ishwaran et al. (2008)]. Traditional regression methods such as CoxPH assume proportional hazards, linear covariate effects, and manageable dimensionality. However, in microarray survival analysis, the number of predictors (genes) far exceeds the number of samples, leading to severe multicollinearity and the curse of dimensionality. These assumptions limit the reliability of traditional models, as they struggle to capture nonlinear interactions and fail to generalize when predictors are in the tens of thousands. To address these drawbacks, alternative methods such as Bayesian Additive Regression Trees (BART) and Bayesian Adaptive Spline Surface (BASS) have been developed (see [Li and Luan (2005); Tibshirani (1997)]).

BART is a non-parametric Bayesian approach that offers flexibility and robustness in modeling complex relationships between predictors and outcomes Li and Luan (2005). Unlike traditional methods, BART can capture intricate interactions and non-linear effects, making it particularly suitable for high-dimensional genomic data. Integrating BART with microarray data aims to improve the precision of survival predictions and identify critical genes that could serve as therapeutic targets Sparapani et al. (2020).

Although BART has been applied in high-dimensional survival problems, our contribution lies in (i) demonstrating its robustness across five diverse real-world cancer microarray datasets; (ii) directly benchmarking BART against both CoxPH and BASS, which has rarely been done; and (iii) showing that BART not only improves predictive performance (lower RMSE) but also produces interpretable influential genes and interactions, which can inform biological understanding. This dual emphasis on predictive accuracy and interpretability distinguishes our work from earlier studies that focused solely on prediction.

2 Methodology

Based on different number of predictors and for consistency, comparisons of the CoxPH, BASS and BART models are discussed.

For each dataset, the initial number of predictors was determined based on the size of the training set, to satisfy the requirement of the CoxPH model, which

assumes that the number of predictors (p) should not exceed the number of observations (n). In other words, to ensure model identifiability and stable parameter estimation within the training data, predictors were selected such that $p \leq n$ for each dataset. Consequently, smaller datasets (such as GSE10300 and GSE17618 with 44 total samples each, split into smaller training subsets) were modeled with 35 predictors, whereas larger datasets (e.g., GSE14333 with 226 samples and GSE20685 with 327 samples) were modeled with proportionally more predictors. This design ensured that all models, particularly CoxPH, were estimated under statistically valid conditions while allowing fair comparison with BASS and BART.

We selected CoxPH and BASS for comparison because CoxPH represents the classical baseline model in survival analysis, while BASS is a flexible Bayesian method that, like BART, can handle nonlinearities and high-dimensional data. Including these two benchmarks allows us to evaluate BART against both the traditional standard and a modern Bayesian competitor. While other tree-based methods such as Random Survival Forests exist, BASS provides a more direct Bayesian framework for comparison, aligning with our emphasis on Bayesian inference.

The superior predictive ability of the BART models is equally highlighted in the subsequent subsections.

2.1 Standard Cox Regression (CoxPH) Method

The well-known and one of the most popular statistical approaches that been used as a standard for modeling survival is the long-standing traditional Cox Proportional Hazard model (CoxPH) or also known as the Cox regression Gelman et al. (1995). Because of its semi-parametric nature, it is considered to be a fundamental result in the class of survival modeling approaches. The results are of the CoxPH closely approximate those for a correct parametric model.

Given a sample size of n to estimate the relationship between survival time and predictor (genes) $X_1, X_2, \ldots, X_p$, each sample i (where $i = 1, \ldots, n$) is represented by $(t_i, \delta_i, x_{i1}, x_{i2}, \ldots, x_{ip})$. Here, δ_i is the censoring indicator, with t_i being the censoring time if $\delta_i = 0$ or the survival time if $\delta_i = 1$. The vector $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$ contains the genes for the i-th sample.

The CoxPH model illustrates that the hazard function $\eta(t)$, representing the risk of death at time t for an individual with a specific predictor profile, is given by:

$$\eta(t \mid X) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) = \lambda_0(t) \exp^{h(X)}, \quad (1)$$

where $\lambda_0(t)$ is the baseline hazard function, $\beta = \{\beta_1, \ldots, \beta_p\}$ is the vector of regression coefficients, and $X = \{X_1, \ldots, X_p\}$ is the vector of predictor levels.

The last feature of (1) to note is that the baseline hazard depends on t but not in now way with X's. An equally important thing to understand is that the baseline hazard, λ can be a function of time with no explicit form.

The above model (established in Equation (1)) assumes the linearity of X and constant β over time because survival analysis is commonly used to investigate the relationship between covariate-dependent distribution functions. In the context of this study, proportional hazard means that individuals are independent and homogonously given their base-line covariates.

2.2 Overview of Bayesian Approach for survival analysis

Bayesian survival analysis integrates prior knowledge with observed data so that the outcome can be interpreted through both the likelihood function and model complexity. The prior distribution serves as a smoothing mechanism that stabilizes parameter estimation, especially in cases of limited data. This is particularly important when the sample size is very small, as incorporating informative priors helps to reduce the variability of estimates and prevent overfitting. In situations with small samples, one practical approach to control the standard error is to balance estimates by aggregating both overestimates and underestimates of uncertainty, thereby achieving more stable inference.

In addition to providing posterior probability estimates and standard errors for each independent parameter, the Bayesian approach also yields the probability that a parameter lies within a specified range, a feature not directly available in frequentist methods. Through the use of Markov Chain Monte Carlo (MCMC) techniques, Bayesian methods effectively address the problem of over-parameterization that often arises in complex multi-level models. These models may include random effects, time-varying coefficients, and multiple parameters at different hierarchical levels, all of which can be estimated efficiently within the Bayesian framework under the assumption of zero-mean priors for the random components.

The approach naturally deals with censoring by including it in the likelihood function and facilitates hierarchical modeling, allowing for the consideration of multiple levels of variation. Additionally, Bayesian survival analysis aids in model comparison and selection using criteria like the Deviance Information Criterion (DIC), making it a powerful tool for comprehensive survival analysis Saha (2017); Put et al. (2004).

2.2.1 Bayesian Adaptive Spline Surface (BASS)

Bayesian Adaptive Spline Surface (BASS) Francom and Sansó (2020) models are a class of flexible, non-parametric regression and function estimation models. They combine a spline-based model with Bayesian inference, for adaptively fitting complex surfaces. Spline Basis Functions Adptivity priors on the parameters of the spline functions such as include the coefficients and positions of any knots to influence surface smoothness flexibility Residual errors are assumed to be normally distributed about mean zero constant variance.

Let y_i denote the dependent variable and x_i be a vector of p independent variables, scaled between zero and one without loss of generality, where $i = 1, 2, \ldots, n$. Then, the BASS model is defined as:

$$Y = f(x) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2), \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$.
The function $f(x)$ is expressed as

$$f(x) = a_0 + \sum_{m=1}^{M} a_m B_m(X), \tag{3}$$

where

$$B_m(X) = \prod_{k=1}^{K_m} g_{km} \left[s_{km} \left(x_{v_{km}} - t_{km} \right) \right]_+^{\alpha}, \tag{4}$$

$s_{km} \in \{-1, 1\}$ is the sign, $t_{km} \in [0, 1]$ is a knot, v_{km} selects a variable, K_m is the degree of interaction, and g_{km} ensures the basis function has a maximum of one. The function $[\cdot]_+$ is defined as $\max(0, \cdot)$, and α determines the polynomial splines' degree. Each variable can be used only once in a basis function, and M is the number of basis functions. The vector a contains the basis coefficients, including the intercept.

The model (2)–(4) is significant due to its flexibility and adaptive nature, which allows it to effectively capture complex, non-linear relationships between predictors and response variables. Its Bayesian framework enables the incorporation of prior information and quantification of uncertainty, providing a probabilistic approach to inference. BASS models are robust against overfitting through sparsity-inducing priors and are capable of handling high-dimensional data, making them suitable for a wide range of applications across various fields including environmental modeling, finance, and biology. This flexibility, combined with interpretability through spline coefficients, makes BASS a powerful tool for understanding and predicting intricate data patterns.

2.2.2 Bayesian Additive Regression Tress (BART)

One of the most celebrated, important and useful models, that has attracted and continues to attract attention from researchers in regression study in the last few decades, is the famous Bayesian Additive Regression Trees (BART) Chipman et al. (2010), developed by Chipman et al. in 2015. It provides a framework for flexible nonparametric modeling of relationships of covariates to outcomes.

In recent times, BART models have been shown to maintain an excellent survival predictive performance when additional, irrelevant regressors are added, thereby diminishing the need to carry out variable or model selection for both continuous and binary outcomes, thereby exceeding that of its competitors. It equally has a natural quantification of uncertainty that allows construction of credible and prediction intervals.

In addition, BART can address nonlinear relationships of a response variable to a possibly large set of regressors. Possible interactions between these regressors are automatically addressed by tree-based methods, as these relationships are jointly estimated. The model predicts patient status based on gene expressions using the following equation:

$$Y = f(X) + \epsilon \approx \tau_1^M(X) + \tau_2^M(X) + \cdots + \tau_m^M(X) + \epsilon, \tag{5}$$

where $\epsilon \sim N_n(0, \sigma^2 I_n)$.

In (5), Y is the $n \times 1$ vector of responses, X is the $n \times p$ design matrix of predictors, and E is the $n \times 1$ noise vector. The model includes m regression trees, each represented by τ^M, consisting of the tree structure τ and terminal node parameters M. Each tree's structure involves splitting rules $x_j < c$ for internal nodes, directing observations left or right based on these rules until a terminal node is reached. The predicted value for an observation is the sum of the terminal node values. The set of leaf parameters for a tree is $M_t = \{\mu_1, \mu_2, \ldots, \mu_b\}$, where b is the number of terminal nodes.

The BART model is employed in this study to establish its predictive performance of hazard of patients given thousands of predictors (genes).

2.3 Models Prediction Performances

Model prediction performance metrics are crucial in evaluating the effectiveness and reliability of predictive models. Common metrics include accuracy, which measures the proportion of correct predictions out of total predictions, and precision, which assesses the correctness of positive predictions. Recall, or sensitivity, evaluates the model's ability to identify all relevant instances, while F1-score provides a balance between precision and recall. Mean

Absolute Error (MAE) and Mean Squared Error (MSE) quantify prediction errors for continuous outcomes, with MAE representing the average of absolute errors and MSE emphasizing larger errors.

These metrics collectively offer a comprehensive understanding of a model's predictive performance, guiding improvements and ensuring robust decision-making.

2.3.1 Mean Square Error (RMSE)

The Root Mean Squared Error (RMSE) is an effective measure of accuracy widely used and regarded as a reliable general-purpose error metric for numerical predictions. It is used to compare forecasting errors of different models or model configurations for a specific variable. It is calculated by taking the square root of the average of the squared differences between predicted values and observed values as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}, \tag{6}$$

where y_i's are the observations; $\hat{y}_i$'s are the predicted values of the variable y; and n is the number of observations.

2.3.2 Mean Absolute Error (MAE)

The MAE is calculated as the average of the absolute differences between observed values and predicted values as:

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}, \tag{7}$$

where y_i are the observations, $\hat{y}_i$ are the predicted values of the variable y, and n is the number of observations.
MAE (7) is employed in this study alongside RMSE (6) to check the prediction performances of the CoxPH (1) BASS (2) and BART (3) models.

The current study employs the RSME (6) to check the prediction performances of the CoxPH (1) BASS (2) and BART (3) models.

The models were then applied to five real-world cancer microarray datasets to evaluate predictive accuracy and gene interpretability.

3 Results and Discussion

We first summarize dataset characteristics (Table 1) before presenting VIF (Table 2–6).

To show the superior predictive ability of BART model, we considered five cancer gene sequence datasets obtained from `http://www.ncbi.nlm.nih.gov`. These data sets are GSE10300, GSE14333, GSE16446, GSE17618 and GSE20685 with 16,183, 54,712, 54,739, 54,726 and 54,692 genes respectively. The sample sizes are 44, 226, 107, 44 and 327 respectively. The training set consists of 80 percent of the data while the test set consists of 20 percent of the data.

Table 1: Summary of the five datasets

Dataset	No. of Samples	No. of Genes	Median of Hazard
GSE10300	44	16,183	0.362097
GSE14333	226	54,712	0.443691
GSE16446	107	54,739	0.133103
GSE17618	44	54,726	0.631750
GSE20685	327	54,692	0.268935

Table 1 summarizes data from five different datasets, listing the number of samples, number of genes, and median hazard values for each. The datasets, GSE10300, GSE14333, GSE16446, GSE17618, and GSE20685, contain 44, 226, 107, 44, and 327 samples, respectively, with gene counts ranging from 16,183 to 54,739. The median hazard values vary across the datasets, with GSE10300 at 0.3621, GSE14333 at 0.4437, GSE16446 at 0.1331, GSE17618 at 0.6317, and GSE20685 at 0.2689.

Heatmap of genes

The Figure 1 visualizes the correlation between variables. From the heatmap below, traces of yellow boxes can be seen outside the off-diagonal which indicates presence of multicollinearity between the genes. This implies that the CoxPH method will give misleading results such as overestimating the parameters and underestimating the standard errors.

To formally validate the visual evidence of multicollinearity, we computed the Variance Inflation Factor (VIF) and condition number for selected subsets of genes from each dataset. The results, summarized in Tables 2a–2e, show extremely high VIF values (some exceeding 10^4 and even reaching infinity in certain predictors), confirming severe multicollinearity across the datasets. These findings support the adoption of non-parametric and regularized approaches such as BART and BASS, which are less sensitive to collinear predictors and better suited for high-dimensional genomic data.

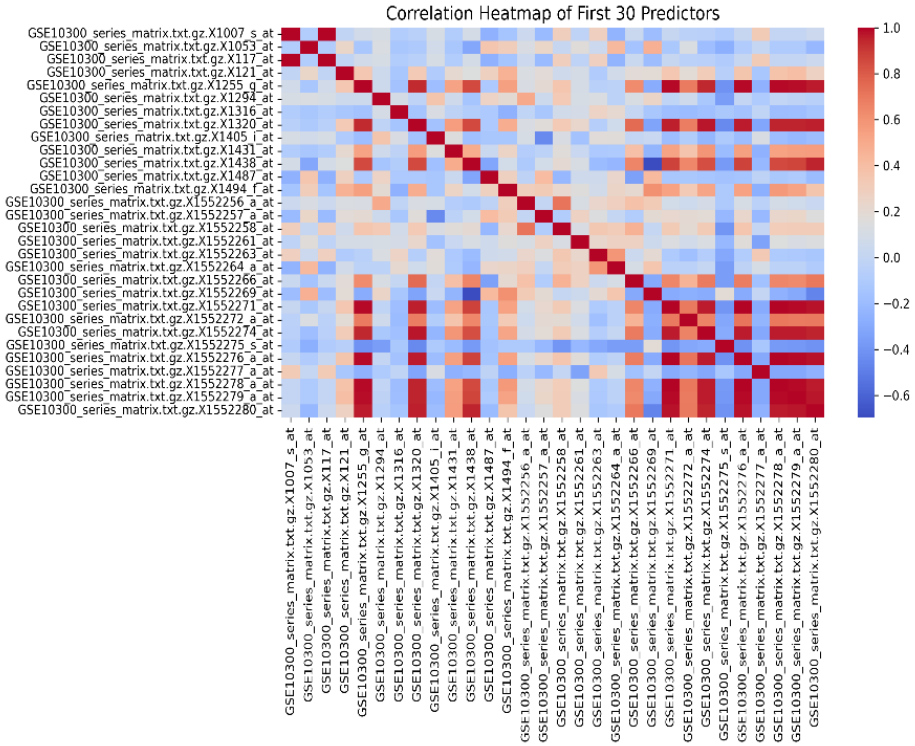

Fig. 1: Heatmap of Gene Expression Profiles with Hierarchical Clustering for dataset GSE10300

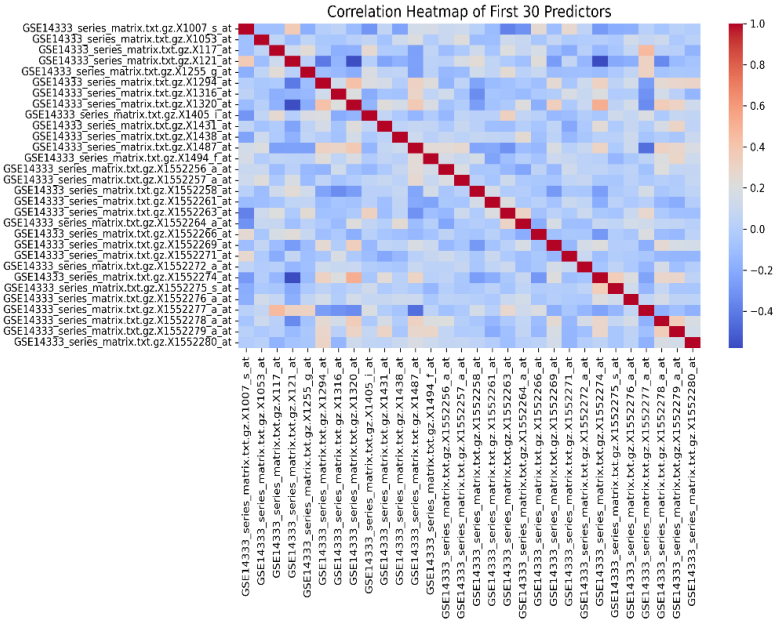

Fig. 2: Heatmap of Gene Expression Profiles with Hierarchical Clustering for dataset GSE14333

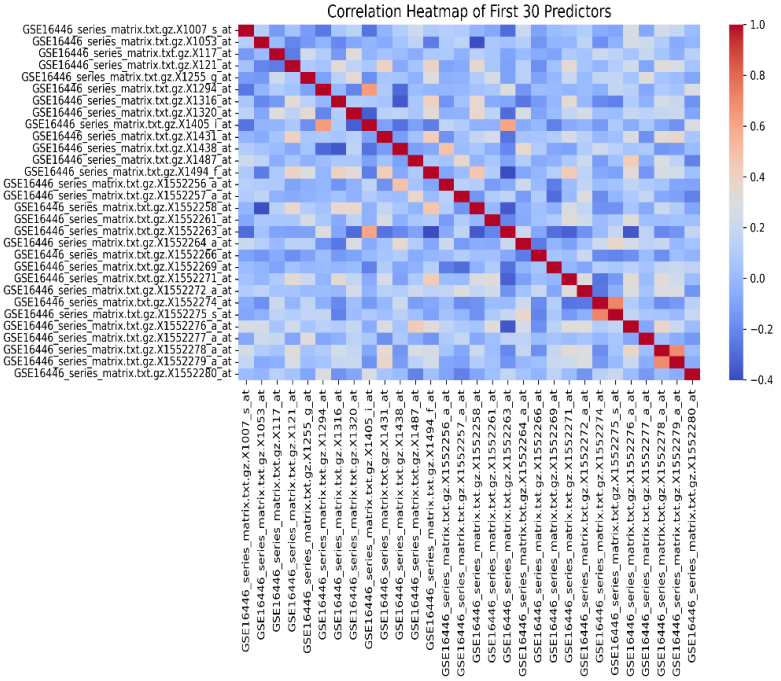

Fig. 3: Heatmap of Gene Expression Profiles with Hierarchical Clustering for dataset GSE16446

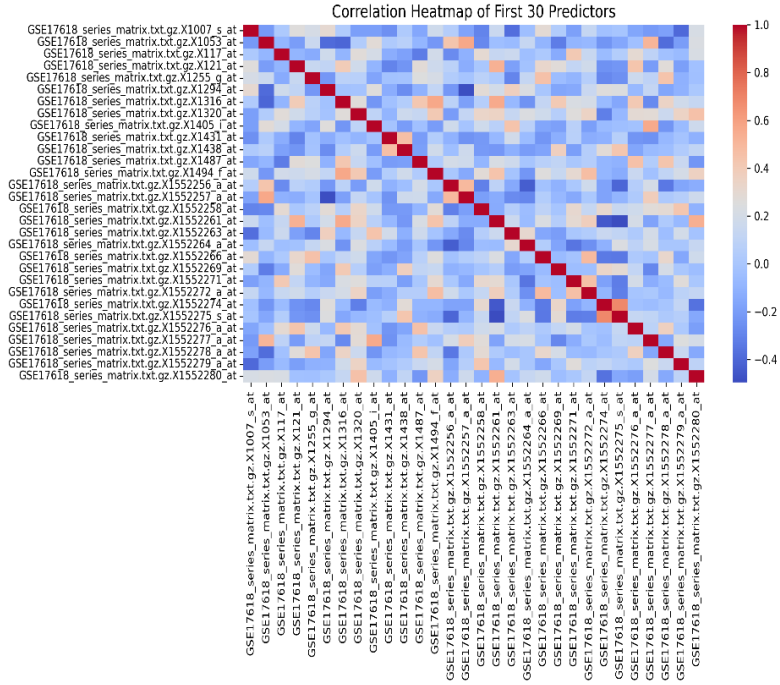

Fig. 4: Heatmap of Gene Expression Profiles with Hierarchical Clustering for dataset GSE17618

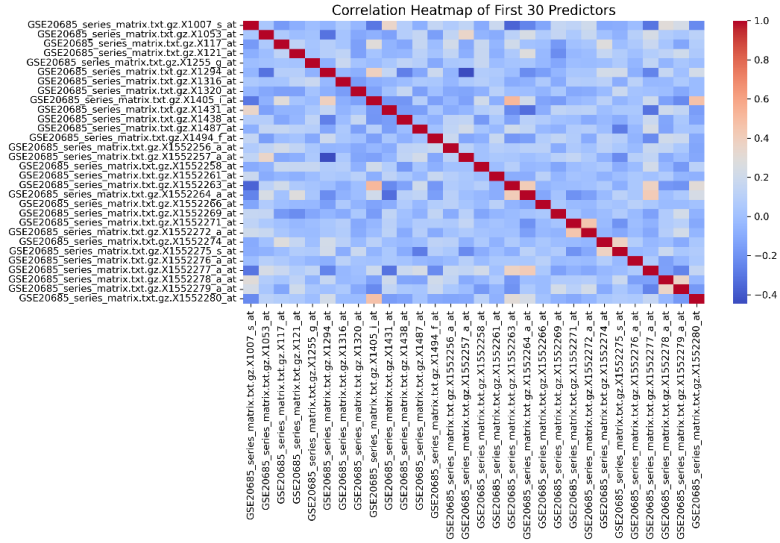

Fig. 5: Heatmap of Gene Expression Profiles with Hierarchical Clustering for dataset GSE20685

Table 2: Variance Inflation Factor (VIF) Results for Selected Predictors of the GSE10300 Dataset

Feature	VIF
GSE10300_series_matrix.txt.gz.X1007_s_at	inf
GSE10300_series_matrix.txt.gz.X1053_at	87.77
GSE10300_series_matrix.txt.gz.X117_at	inf
GSE10300_series_matrix.txt.gz.X121_at	1716.18
GSE10300_series_matrix.txt.gz.X1255_g_at	6757847.84
GSE10300_series_matrix.txt.gz.X1294_at	1872.90
GSE10300_series_matrix.txt.gz.X1316_at	441.37
GSE10300_series_matrix.txt.gz.X1320_at	2762937.83
GSE10300_series_matrix.txt.gz.X1405_i_at	50.66
GSE10300_series_matrix.txt.gz.X1431_at	765.50
GSE10300_series_matrix.txt.gz.X1438_at	17658321.56
GSE10300_series_matrix.txt.gz.X1487_at	6566.46
GSE10300_series_matrix.txt.gz.X1494_f_at	1568063.99
GSE10300_series_matrix.txt.gz.X1552256_a_at	3096.89
GSE10300_series_matrix.txt.gz.X1552257_a_at	489.10
GSE10300_series_matrix.txt.gz.X1552258_at	1047.63
GSE10300_series_matrix.txt.gz.X1552261_at	15953.50
GSE10300_series_matrix.txt.gz.X1552263_at	81.28
GSE10300_series_matrix.txt.gz.X1552264_a_at	2582.75
GSE10300_series_matrix.txt.gz.X1552266_at	11108.29
GSE10300_series_matrix.txt.gz.X1552269_at	5632.14
GSE10300_series_matrix.txt.gz.X1552271_at	71061626.12
GSE10300_series_matrix.txt.gz.X1552272_a_at	38484.82
GSE10300_series_matrix.txt.gz.X1552274_at	14313911.53
GSE10300_series_matrix.txt.gz.X1552275_s_at	2912.19
GSE10300_series_matrix.txt.gz.X1552276_a_at	14326801.88

Table 3: Variance Inflation Factor (VIF) Results for Selected Predictors of the GSE14333 Dataset

Feature	VIF
GSE14333_series_matrix.txt.gz.X1007_s_at	1553.27
GSE14333_series_matrix.txt.gz.X1053_at	436.23
GSE14333_series_matrix.txt.gz.X117_at	123.57
GSE14333_series_matrix.txt.gz.X121_at	1066.65
GSE14333_series_matrix.txt.gz.X1255_g_at	26.20
GSE14333_series_matrix.txt.gz.X1294_at	540.49
GSE14333_series_matrix.txt.gz.X1316_at	262.86
GSE14333_series_matrix.txt.gz.X1320_at	23.15
GSE14333_series_matrix.txt.gz.X1405_i_at	28.21
GSE14333_series_matrix.txt.gz.X1431_at	20.86
GSE14333_series_matrix.txt.gz.X1438_at	76.44
GSE14333_series_matrix.txt.gz.X1487_at	619.54
GSE14333_series_matrix.txt.gz.X1494_f_at	242.61
GSE14333_series_matrix.txt.gz.X1552256_a_at	411.24
GSE14333_series_matrix.txt.gz.X1552257_a_at	843.07
GSE14333_series_matrix.txt.gz.X1552258_at	126.56
GSE14333_series_matrix.txt.gz.X1552261_at	21.38
GSE14333_series_matrix.txt.gz.X1552263_at	104.94
GSE14333_series_matrix.txt.gz.X1552264_a_at	548.53
GSE14333_series_matrix.txt.gz.X1552266_at	60.13
GSE14333_series_matrix.txt.gz.X1552269_at	4.86
GSE14333_series_matrix.txt.gz.X1552271_at	51.28
GSE14333_series_matrix.txt.gz.X1552272_a_at	108.66
GSE14333_series_matrix.txt.gz.X1552274_at	322.51
GSE14333_series_matrix.txt.gz.X1552275_s_at	170.33
GSE14333_series_matrix.txt.gz.X1552276_a_at	12.50
GSE14333_series_matrix.txt.gz.X1552280_at	21.77

Table 4: Variance Inflation Factor (VIF) Results for Selected Predictors of the GSE16446 Dataset

Feature	VIF
GSE16446_series_matrix.txt.gz.X1007_s_at	482.02
GSE16446_series_matrix.txt.gz.X1053_at	502.67
GSE16446_series_matrix.txt.gz.X117_at	231.20
GSE16446_series_matrix.txt.gz.X121_at	1318.42
GSE16446_series_matrix.txt.gz.X1255_g_at	296.59
GSE16446_series_matrix.txt.gz.X1294_at	566.82
GSE16446_series_matrix.txt.gz.X1316_at	922.07
GSE16446_series_matrix.txt.gz.X1320_at	521.51
GSE16446_series_matrix.txt.gz.X1405_i_at	214.90
GSE16446_series_matrix.txt.gz.X1431_at	490.13
GSE16446_series_matrix.txt.gz.X1438_at	113.57
GSE16446_series_matrix.txt.gz.X1487_at	645.28
GSE16446_series_matrix.txt.gz.X1494_f_at	646.81
GSE16446_series_matrix.txt.gz.X1552256_a_at	235.36
GSE16446_series_matrix.txt.gz.X1552257_a_at	567.10
GSE16446_series_matrix.txt.gz.X1552258_at	736.74
GSE16446_series_matrix.txt.gz.X1552261_at	446.83
GSE16446_series_matrix.txt.gz.X1552263_at	538.00
GSE16446_series_matrix.txt.gz.X1552264_a_at	764.67
GSE16446_series_matrix.txt.gz.X1552266_at	158.61
GSE16446_series_matrix.txt.gz.X1552269_at	147.05
GSE16446_series_matrix.txt.gz.X1552271_at	716.10
GSE16446_series_matrix.txt.gz.X1552272_a_at	525.27
GSE16446_series_matrix.txt.gz.X1552274_at	590.85
GSE16446_series_matrix.txt.gz.X1552275_s_at	381.35
GSE16446_series_matrix.txt.gz.X1552276_a_at	986.87

Table 5: Variance Inflation Factor (VIF) Results for Selected Predictors of the GSE17618 Dataset

Feature	VIF
GSE17618_series_matrix.txt.gz.X1007_s_at	8871.71
GSE17618_series_matrix.txt.gz.X1053_at	8425.93
GSE17618_series_matrix.txt.gz.X117_at	685.50
GSE17618_series_matrix.txt.gz.X121_at	22474.41
GSE17618_series_matrix.txt.gz.X1255_g_at	6353.66
GSE17618_series_matrix.txt.gz.X1294_at	4507.75
GSE17618_series_matrix.txt.gz.X1316_at	9530.36
GSE17618_series_matrix.txt.gz.X1320_at	34666.91
GSE17618_series_matrix.txt.gz.X1405_i_at	345.87
GSE17618_series_matrix.txt.gz.X1431_at	811.24
GSE17618_series_matrix.txt.gz.X1438_at	3165.40
GSE17618_series_matrix.txt.gz.X1487_at	5190.61
GSE17618_series_matrix.txt.gz.X1494_f_at	1637.15
GSE17618_series_matrix.txt.gz.X1552256_a_at	6566.47
GSE17618_series_matrix.txt.gz.X1552257_a_at	14105.54
GSE17618_series_matrix.txt.gz.X1552258_at	24902.85
GSE17618_series_matrix.txt.gz.X1552261_at	35420.51
GSE17618_series_matrix.txt.gz.X1552263_at	1672.84
GSE17618_series_matrix.txt.gz.X1552264_a_at	5177.46
GSE17618_series_matrix.txt.gz.X1552266_at	2987.37
GSE17618_series_matrix.txt.gz.X1552269_at	1690.08
GSE17618_series_matrix.txt.gz.X1552271_at	7840.66
GSE17618_series_matrix.txt.gz.X1552272_a_at	14916.24
GSE17618_series_matrix.txt.gz.X1552274_at	16277.12
GSE17618_series_matrix.txt.gz.X1552275_s_at	2711.86
GSE17618_series_matrix.txt.gz.X1552276_a_at	4187.07

Table 6: Variance Inflation Factor (VIF) Results for Selected Predictors of the GSE20685 Dataset

Feature	VIF
GSE20685_series_matrix.txt.gz.X1007_s_at	689.86
GSE20685_series_matrix.txt.gz.X1053_at	715.45
GSE20685_series_matrix.txt.gz.X117_at	225.06
GSE20685_series_matrix.txt.gz.X121_at	955.05
GSE20685_series_matrix.txt.gz.X1255_g_at	23.89
GSE20685_series_matrix.txt.gz.X1294_at	713.83
GSE20685_series_matrix.txt.gz.X1316_at	364.20
GSE20685_series_matrix.txt.gz.X1320_at	29.94
GSE20685_series_matrix.txt.gz.X1405_i_at	136.68
GSE20685_series_matrix.txt.gz.X1431_at	74.99
GSE20685_series_matrix.txt.gz.X1438_at	112.71
GSE20685_series_matrix.txt.gz.X1487_at	903.44
GSE20685_series_matrix.txt.gz.X1494_f_at	47.36
GSE20685_series_matrix.txt.gz.X1552256_a_at	360.88
GSE20685_series_matrix.txt.gz.X1552257_a_at	433.00
GSE20685_series_matrix.txt.gz.X1552258_at	294.88
GSE20685_series_matrix.txt.gz.X1552261_at	29.26
GSE20685_series_matrix.txt.gz.X1552263_at	372.25
GSE20685_series_matrix.txt.gz.X1552264_a_at	681.57
GSE20685_series_matrix.txt.gz.X1552266_at	43.03
GSE20685_series_matrix.txt.gz.X1552269_at	31.66
GSE20685_series_matrix.txt.gz.X1552271_at	32.55
GSE20685_series_matrix.txt.gz.X1552272_a_at	190.80
GSE20685_series_matrix.txt.gz.X1552274_at	517.63
GSE20685_series_matrix.txt.gz.X1552275_s_at	181.87
GSE20685_series_matrix.txt.gz.X1552276_a_at	25.53

Table 7: Models' performance measured using the Root Mean Squared Error (RMSE)

Models	Base Predictors	500	1000	2000	5000	10000	All Genes
GSE10300							
CoxPH	1.8708	-	-	-	-	-	-
BASS	12.1504	21.3305	12.5179	11.1919	52.1055	30.2020	-
BART	0.2008	0.1824	0.1780	0.1682	0.1655	0.1759	0.1889
GSE14333							
CoxPH	4.0620	-	-	-	-	-	-
BASS	42.1571	31.1415	30.7103	31.5447	29.8666	29.5800	-
BART	0.9027	0.1789	0.1736	0.8378	0.8351	0.8316	0.8438
GSE16446							
CoxPH	2.1213	-	-	-	-	-	-
BASS	2.0748	2.1331	2.2270	2.0824	2.0967	2.1328	-
BART	0.0699	0.0715	0.0711	0.0703	0.0704	0.0700	0.0646
GSE17618							
CoxPH	1.0000	-	-	-	-	-	-
BASS	10.1689	11.5877	9.8517	9.9896	10.2345	11.1126	-
BART	0.3711	0.3761	0.3824	0.3696	0.3750	0.3773	0.3644
GSE20685							
CoxPH	3.9370	-	-	-	-	-	-
BASS	3.3274	2.9261	3.5251	2.9133	3.0910	5.0341	-
BART	0.0985	0.1046	0.1020	0.1028	0.0921	0.0990	0.0902

Table 8: Models' performance measured using the Mean Absolute Error (MAE)

Models	Base Predictors	500	1000	2000	5000	10000	All Genes
GSE10300							
CoxPH	0.6667	-	-	-	-	-	-
BASS	0.2555	0.2985	0.7111	0.2423	0.6122	0.4965	-
BART	0.2262	0.2178	0.2233	0.2241	0.2305	0.2280	0.2216
GSE14333							
CoxPH	0.5167	-	-	-	-	-	-
BASS	0.8181	0.7079	0.7184	0.7286	0.6930	0.6954	-
BART	0.6880	0.6758	0.6708	0.6774	0.6483	0.6885	0.6847
GSE16464							
CoxPH	0.4286	-	-	-	-	-	-
BASS	0.0485	0.0495	0.0526	0.0486	0.0491	0.0495	-
BART	0.0456	0.0465	0.0643	0.0459	0.0477	0.0461	0.0456
GSE17618							
CoxPH	0.4750	-	-	-	-	-	-
BASS	0.2646	0.2660	0.2602	0.2644	0.2686	0.2644	-
BART	0.2263	0.1996	0.2093	0.2070	0.2072	0.2085	0.2457
GSE20685							
CoxPH	0.4923	-	-	-	-	-	-
BASS	0.0814	0.0739	0.0789	0.0738	0.0762	0.0953	-
BART	0.0751	0.0668	0.0661	0.0664	0.0655	0.0639	0.0668

Table 7–8 presents a comparative analysis of the three models (1)–(5); evaluated across five different gene expression datasets (GSE10300, GSE14333, GSE16464, GSE17618, GSE20685) with varying numbers of predictors. Each model's performance is measured using the RMSE (6). The R statistical packages used in achieving this were "bartMachine" for the BART model, package "BASS" for the BASS model and "survive" for the CoxPH model.

In the GSE10300 dataset, the CoxPH model had an RMSE of 1.8708 with 35 predictors, which shows a moderate level of predictive accuracy. In contrast, the BASS model exhibits a wide range of RMSE values, demonstrating both strengths and weaknesses depending on the number of predictors used. For instance, with 35 predictors, BASS achieves an RMSE of 12.1504, significantly higher than CoxPH, indicating poor performance. However, as the number of predictors increases, BASS's RMSE fluctuates, peaking at 52.1055 with 5000 predictors before stabilizing at lower values with larger predictor counts. The BART model, on the other hand, consistently performs better than BASS, with RMSE values ranging from 0.1655 to 0.2008 across various predictor counts, suggesting that BART is better suited for this dataset.

In the GSE14333 dataset, CoxPH again shows a modest RMSE of 4.0620 with 181 predictors. BASS's RMSE values remain significantly high, with a peak of 42.1571 at 35 predictors, but generally reduce as more predictors are added, indicating its adaptability. The BART model shows more consistent RMSE values, ranging from 0.1736 to 0.9027, which demonstrates its superiority.

For the GSE16464 dataset, CoxPH records an RMSE of 2.1213, while BASS maintains an RMSE range of 2.0748 to 2.2270 across the varying predictor counts. This performance suggests that BASS remains competitive but does not outperform CoxPH in some cases. BART, however, again displays a lower RMSE compared to both CoxPH and BASS, indicating a consistent performance across predictors.

The GSE17618 dataset presents a similar pattern, with CoxPH's RMSE at 1.0000, indicating strong performance with 35 predictors. BASS's RMSE values fluctuate, having a maximum of 11.5877 at 500 predictors and a minimum of 9.8517 at 1000 predictors, which shows its weak predictive power among the other two models considered. BART's RMSE values remain consistent with higher performance levels than both CoxPH and BASS, reflecting a potential advantage in its predictive capacity.

Finally, in the GSE20685 dataset, CoxPH shows an RMSE of 3.9370, while BASS continues to demonstrate its poor performance with a peak RMSE of 5.0341 at 10,000 predictors. BART again shows lower RMSE values, ranging from 0.0902 to 0.1046, which is still lower than CoxPH and BASS, indicating

a more stable performance than in previous datasets.

A complementary assessment using the Mean Absolute Error (MAE) is provided in Appendix Table 4. The MAE evaluates the average magnitude of prediction errors and, unlike RMSE, is less sensitive to large deviations. The MAE results show a similar pattern to the RMSE outcomes, with the BART model consistently yielding the lowest error values across all datasets. This reinforces the robustness and reliability of BART's predictive performance relative to the CoxPH and BASS models.

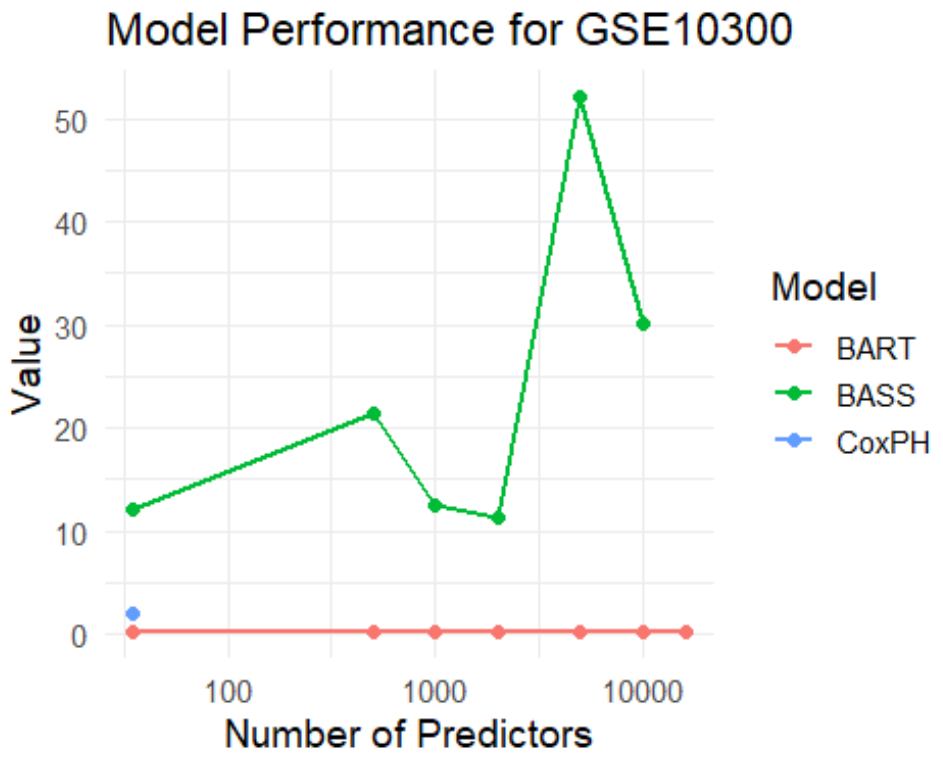

Fig. 6: Comparison of RMSE Across Models (CoxPH, BASS, and BART) Against Number of Parameters for dataset GSE10300

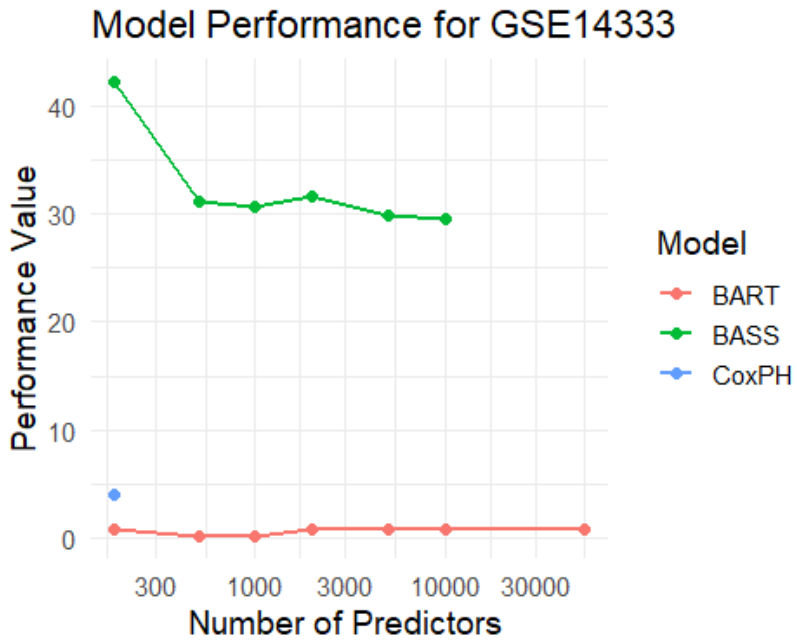

Fig. 7: Comparison of RMSE Across Models (CoxPH, BASS, and BART) Against Number of Parameters for dataset GSE14333

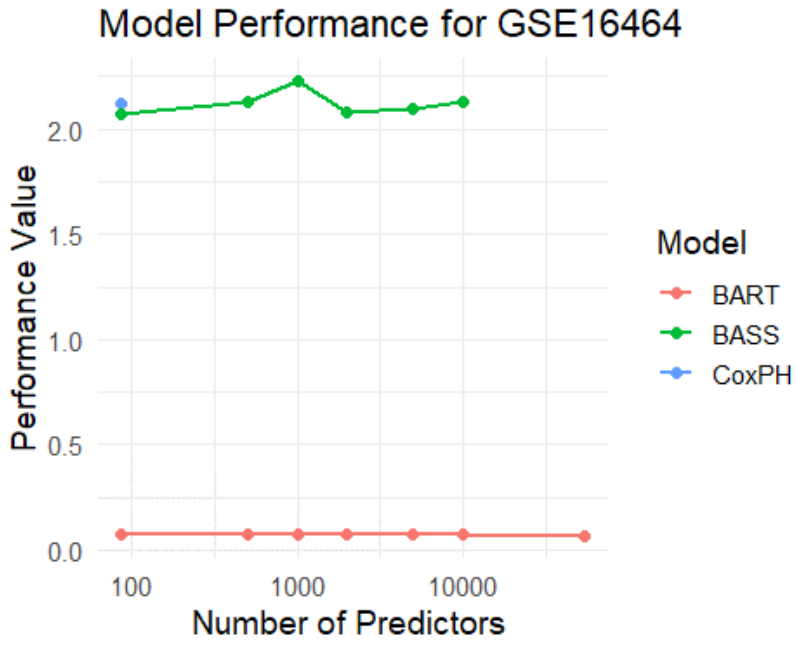

Fig. 8: Comparison of RMSE Across Models (CoxPH, BASS, and BART) Against Number of Parameters for dataset GSE16446

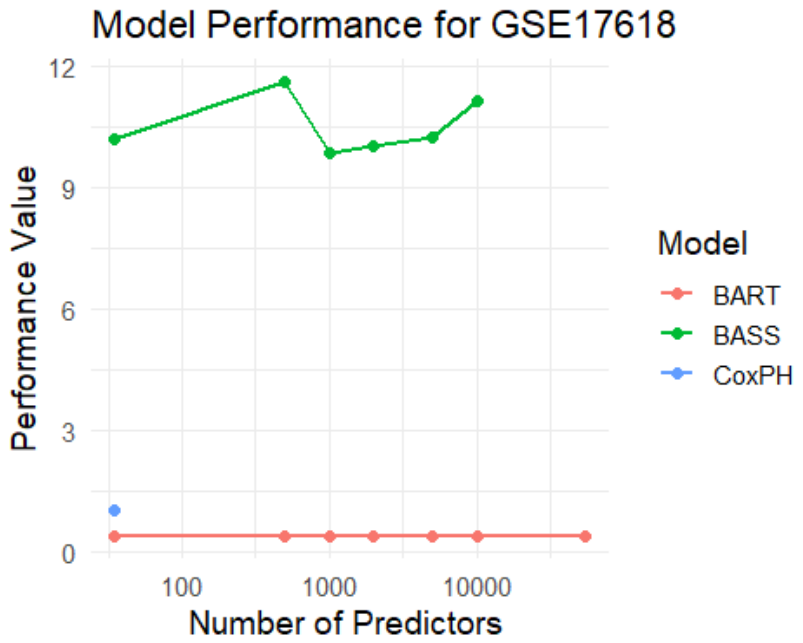

Fig. 9: Comparison of RMSE Across Models (CoxPH, BASS, and BART) Against Number of Parameters for dataset GSE17618

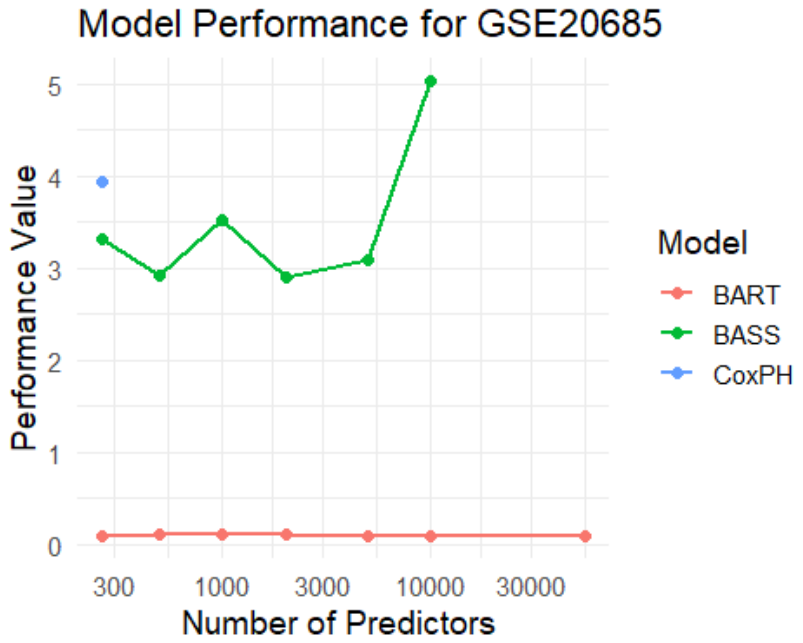

Fig. 10: Comparison of RMSE Across Models (CoxPH, BASS, and BART) Against Number of Parameters for dataset GSE20685

Figures 6 through 10 compare the Root Mean Square Error (RMSE) of three models; Cox Proportional Hazards (CoxPH), Bayesian Additive Regression Trees (BART), and Bayesian Adaptive Spline Surfaces (BASS) across varying numbers of parameters for different datasets (GSE10300, GSE14333, GSE16464, GSE17618, and GSE20685). Across the datasets, the BASS model exhibits a significant variation in RMSE with increasing parameter numbers, suggesting sensitivity to model complexity. Conversely, BART demonstrates consistently low RMSE, indicating robustness across different datasets and parameter scales. CoxPH generally maintains stable and low RMSE but is limited in scalability. These comparisons highlight the performance variability of the models, with BART emerging as a reliable choice across datasets and parameter settings.

Proportion of Variance Explained Across Variables

The provided figures (Figures 3–3) below depict the proportion of variance explained across variables for different datasets: GSE10300, GSE14333, GSE16464, GSE17618, and GSE20685. Each figure likely highlights the principal components within each dataset, showcasing how variance is distributed across variables.

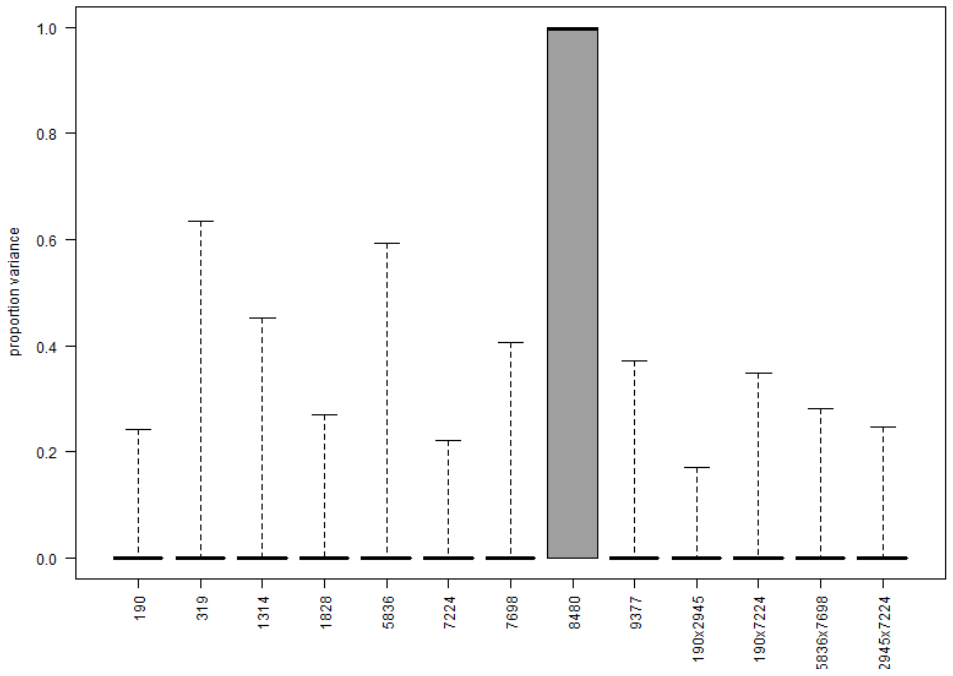

Fig. 11: Proportion of Variance Explained Across Variables for dataset GSE10300

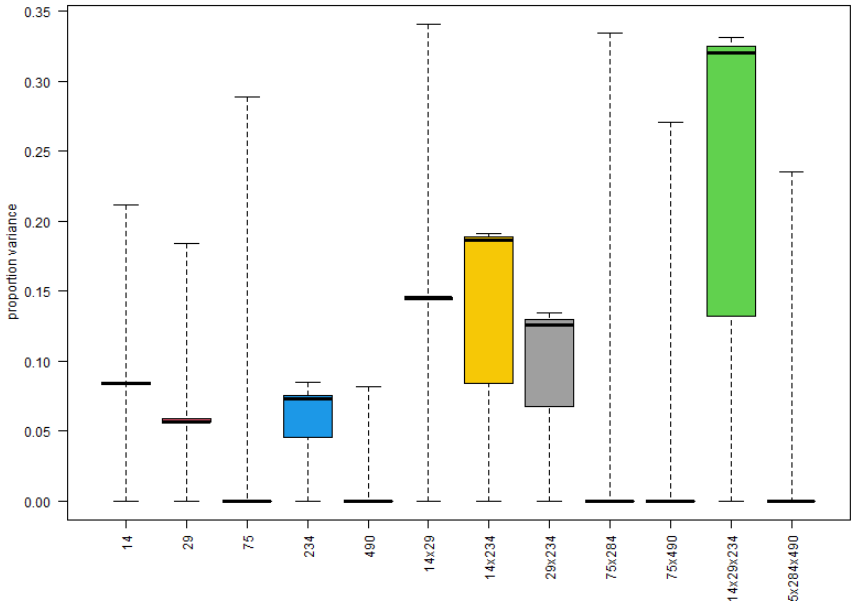

Fig. 12: Proportion of Variance Explained Across Variables for dataset GSE14333

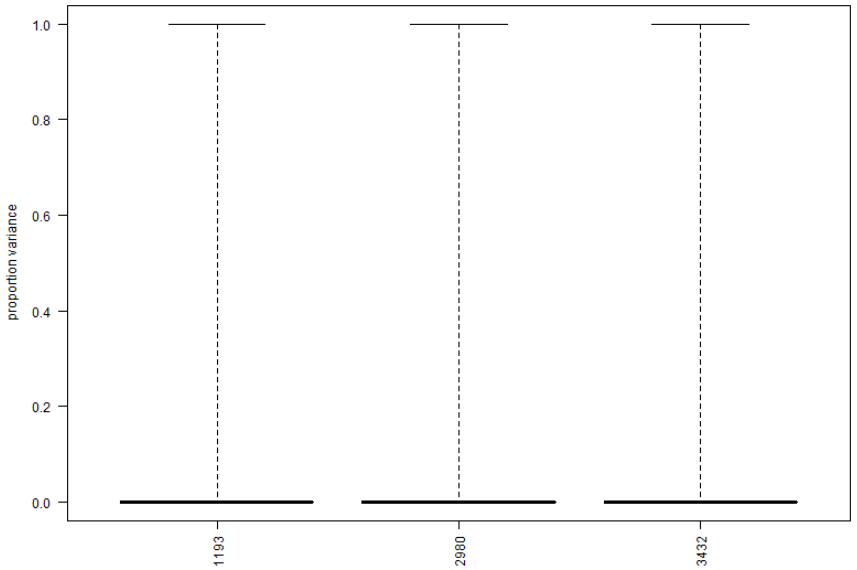

Fig. 13: Proportion of Variance Explained Across Variables for dataset GSE16446

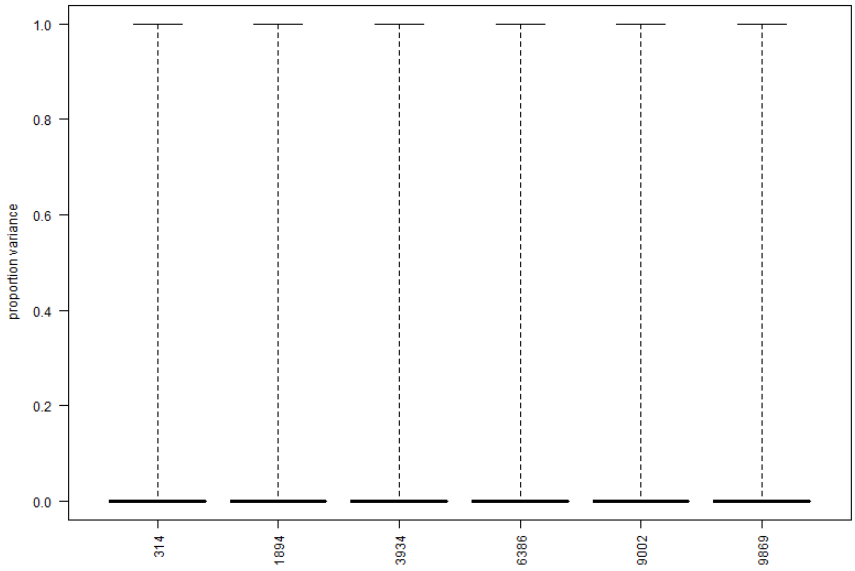

Fig. 14: Proportion of Variance Explained Across Variables for dataset GSE17618

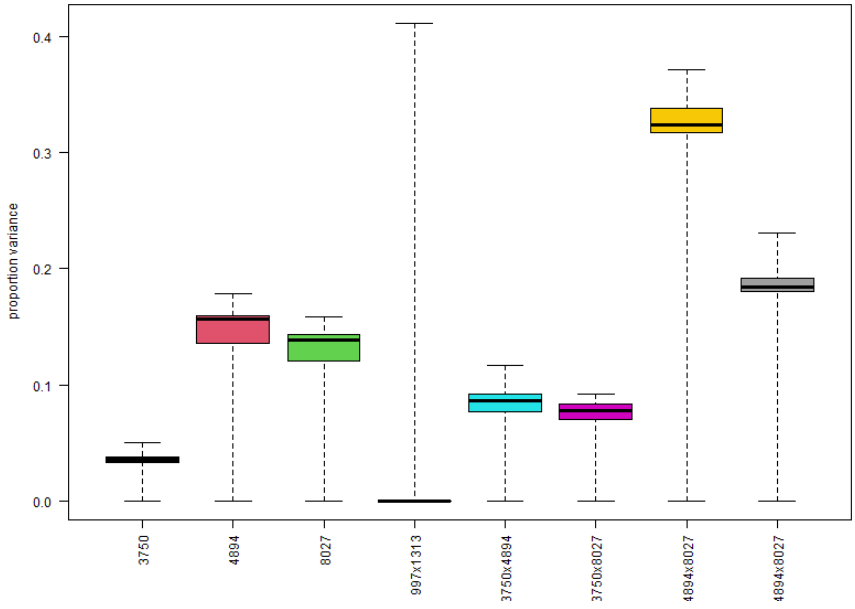

Fig. 15: Proportion of Variance Explained Across Variables for dataset GSE20685

4 Discussion

BART having the lowest RMSE (0.2008, 0.9027, 0.0699, 0.3711, and 0.0985), has demonstrated its predictive superiority over the CoxPH and BASS models with BASS having the highest RMSE in most cases for the five microarray survival data respectively. Also, at 10,000 parameters, BART still performed better than BASS in the different microarray data sets.

Our findings suggest that BART achieves stable performance even as the number of predictors increases beyond 10,000, whereas BASS shows high variability. The results indicate that BART does not require aggressive dimensionality reduction, but in practice, selecting $\approx 1,000$–$10,000$ predictors appears to balance accuracy and computational efficiency.

Finally, the number of predictors with high influence among the thousands of genes in each data set are 9 genes and 4 interactions (GSE10300), 5 genes and 7 interactions (GSE14333), 3 genes with no interactions (GSE16446), 6 genes with no interaction (GSE17618) and 3 genes with 5 interactions (GSE20685).

5 Conclusion

The impact of five key cancer gene sequences on the risk outcomes of cancer patients has been analyzed using BASS and BART models, as compared with the classical CoxPH model. By evaluating the models' performances through the RSME, BART outperformed BASS and CoxPH survival models at different numbers of covariates(genes) included in the model. This revealed the limitations of the BASS method when the predictors(genes) are more than 10,000. By leveraging BART's capabilities, a more precise and nuanced understanding of gene interactions and their influence on cancer patients' survival is discovered to have paved the way for improved prognostic tools and personalized treatment strategies.

Abbreviations
CoxPH Cox Proportional Hazards
BASS Bayesian Adaptive Spline Surface BART Bayesian Additive Re gression Trees
RSME Root Mean Square Error
MCMC Marcov Chain Monte Carlo
NCBI National Center for Biotechnology Information
LASSO Least Absolute Shrinkage and Selection Operator
MARS Multivariate Adaptive Regression Splines
MAE Mean Absolute Error
DIC Deviance Information Criterion
MSE Mean Square Error

Author Contributions
BO, Fagbemigun was responsible for conception and design of this study. OJ, Samsom and BO, Fagbemigun wrote the main manuscript text. FA, Okolie collected, analyzed all dataset, prepared all figures and interpreted the data. OJ, Samson prepared the final draft of manuscript. All authors reviewed and approved the manuscript.

Acknowledgement
The Authors wish to thank the anonymous referees, whose comments helped to improve the final version of this manuscript.

Funding
The authors received no funding for this study.

References

Mohammed M., Mwambi H. and Omolo B. (2021) Colorectal cancer classification and survival analysis based on an integrated RNA and DNA molecular signature. *Current Bioinformatics*, 16(4):583–600.

Simon R. (2003) Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *British Journal of Cancer*, 89(9):1599–1604.

Van't Veer L. J., Dai H., Van De Vijver M. J., He Y. D., Hart A. A., Mao M., Peterse H. L., Van Der Kooy K., Marton M. J., Witteveen A. T., Schreiber G. J. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

PH Model (2021) Cox proportional hazards model.

Ibrahim J. G., Chen M.-H. and Sinha D. (2001) *Bayesian Survival Analysis*. Springer.

Sparapani R., Spanbauer C. and McCulloch R. (2021) Nonparametric machine learning and efficient computation with Bayesian Additive Regression Trees: The BART R package. *Journal of Statistical Software*, 97:1–66.

Olaniran O. R. and Alzahrani A. R. R. (2023) On the oracle properties of Bayesian random forest for sparse high-dimensional Gaussian regression. *Mathematics*, 11(24):4957.

Baesens B., Van Gestel T., Stepanova M., Van den Poel D. and Vanthienen J. (2005) Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, 56(9):1089–1098.

Chipman H. A., George E. I. and McCulloch R. E. (2010) BART: Bayesian Additive Regression Trees.

Cox D. R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–202.

Francom D. and Sansó B. (2020) BASS: An R package for fitting and performing sensitivity analysis of Bayesian Adaptive Spline Surfaces. *Journal of Statistical Software*, 94:1–32.

Gelman A., Carlin J. B., Stern H. S. and Rubin D. B. (1995) *Bayesian Data Analysis*. Chapman and Hall/CRC.

Ishwaran H., Kogalur U. B., Blackstone E. H. and Lauer M. S. (2008) Random survival forests.

Li H. and Luan Y. (2005) Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, 21(10):2403–2409.

Tibshirani R. (1997) The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Sparapani R., Logan B. R., McCulloch R. E. and Laud P. W. (2020) Nonparametric competing risks analysis using Bayesian Additive Regression Trees. *Statistical Methods in Medical Research*, 29(1):57–77.

Saha S. (2017) Survival analysis with Bayesian Additive Regression Trees and its application.

Put R., Xu Q., Massart D. and Vander Heyden Y. (2004) Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies. *Journal of Chromatography A*, 1055(1–2):11–19.