Article

Not peer-reviewed version

# On the Oracle Properties of Bayesian Random Forest for Sparsed High-Dimensional Gaussian Regression

Oyebayo Ridwan Olaniran * and Ali Rashash R Alzahrani

Posted Date: 19 October 2023

doi: 10.20944/preprints202310.1272.v1

Keywords: Random Forest; Oracle Property; Variable Selection; Bayesian Analysis; Asymptotic Normality

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2023 doi:10.20944/preprints202310.1272.v1

*Article*

# On the Oracle Properties of Bayesian Random Forest for Sparsed High-Dimensional Gaussian Regression

**Oyebayo Ridwan Olaniran** [1,*,†], **Ali Rashash R Alzahrani** [2,†]

[1] Department of Statistics, Faculty of Physical Sciences, University of Ilorin, Ilorin, PMB 1515, Kwara State, Nigeria; olaniran.or@unilorin.edu.ng
[2] Department of Mathematical Sciences, College of Applied Sciences, Umm Al-Qura University, Mecca, Saudi Arabia; arrzahrani@uqu.edu.sa
[*] Correspondence: olaniran.or@unilorin.edu.ng
[†] These authors contributed equally to this work.

**Abstract:** Random Forest (RF) is a widely used data prediction and variable selection technique. However, the variable selection aspect of RF can become unreliable when there are more irrelevant variables than relevant ones. In response, we introduced the Bayesian Random Forest (BRF) method specifically designed for high-dimensional datasets with a sparse covariate structure. Our research demonstrates that BRF possesses the oracle property, which means it achieves strong selection consistency without compromising efficiency or bias.

**Keywords:** random forest; oracle property; variable selection; bayesian analysis; asymptotic normality

**MSC:** 62F15; 62G20; 62G08

## 1. Introduction

Several techniques for handling high-dimensional data have been proposed from different areas of research, such as in oncology (modelling and identification of relevant genetic biomarkers for tumourous cancer cells) [1–5]. The methodologies of the techniques differ, but the collective standpoint is to find an efficient way to analyze high-dimensional data [6]. In a broader sense, high-dimensionality (HD) refers to a modelling situation where the number of unknown parameters $p$ are far greater than the sample size $n$ that's $p \gg n$ [7]. This scenario includes supervised regression and classification with several explanatory variables or features largely greater than sample size, unsupervised learning with more attributes than samples and hypothesis testing parlance with more considered hypotheses than observations [8]. [9] identified the need for developing robust methods for high-dimensional data. Classical methods like ordinary least squares, logistic regression, and $k - NN$ often break down due to an ill-conditioned design matrix when $p \gg n$. [10] described two major approaches to analyzing high-dimensional data, namely: modification of $n > p$ approaches to accommodate high-dimensional data or developing a new approach. Modifying approaches involves moving from complex to simple models by selecting relevant subsets of the $p$ variables. This approach is widely referred to as variable selection.

Variable selection is an approach used to adapt existing low-dimensional data modeling techniques for high-dimensional data. Simultaneously, penalized regression involves imposing constraints on dimensionality to achieve a similar objective. The primary advantage of variable selection methods is their ability to preserve the desirable qualities of low-dimensional approaches like the Maximum Likelihood Estimator (MLE), even though they may struggle to address the complexity of high-dimensional datasets. Penalized methods such as LASSO [11] and SCAD [12], among others, offer a partial solution to the problem but introduce bias in estimation. Both approaches share the drawback of not fully capturing the complexities of high-dimensional datasets, including interactions, non-linearity, and non-normality [13]. One robust procedure that has been shown to overcome these challenges in both low and high-dimensional scenarios is Classification and Regression Trees (CART) [14,15]. CART is a non-parametric statistical method that relaxes dimensionality assumptions and naturally accommodates modeling of interactions and non-linearity.

The strength of CART in terms of simplicity and interpretability is offset by a significant drawback, which often leads to a loss of accuracy. In the late 20th century, a new methodological framework emerged for combining multiple models to create a more comprehensive model, known as ensemble modeling. One of the earliest ensemble techniques within the CART framework is Bagging (Bootstrap Aggregating) [16]. The Bagging process involves taking multiple versions of the bootstrap sample [17] from the training dataset and fitting an unpruned CART to each of these bootstrap samples. The final predictor is derived by averaging these different model versions. Remarkably, this procedure works well and typically outperforms its competitors in most situations. Some intuitive explanations for why and how it works were provided in [18]. This concept has spurred subsequent work, including the development of Random Forests (RF) [19], which presents a broader framework for tree ensembles. RF enhances Bagging by replacing all covariates in the CART's splitting step with a random sub-sampling of covariates. This adjustment helps reduce the correlation between adjacent trees, thereby enhancing predictive accuracy.

The complexity of dealing with high-dimensional data has led to the development of multiple versions of Random Forest (RF) algorithms in the context of regression modeling. One prominent characteristic of high-dimensional datasets is their sparsity, which means there are relatively few relevant predictors within the predictor space. This sparsity is often observed in microarray data, where only a small number of genes are associated with a specific disease outcome [13,20]. The traditional approach of RF, which involves randomly subsampling either the square root of the predictors $\sqrt{p}$ or $\frac{p}{3}$, fails to effectively capture this sparsity [21]. This failure arises from RF's unrealistic assumption that the predictor space should be densely populated with relevant variables to achieve reasonable accuracy. In contrast, boosting techniques, as introduced by [20], specifically address this issue by boosting weak trees rather than averaging all the trees, as done in RF. However, boosting comes at the cost of reduced predictive accuracy compared to RF. The Bayesian modified boosting approach, known as Bayesian Additive Regression Trees (BART) proposed by [22], provides an alternative method for estimating a boosting-inspired ensemble of Classification and Regression Trees (CART) models. Nevertheless, BART does not focus on solving the problem of sparsity and is not robust to an increase in dimensionality, as reported by [23].Therefore, in this paper, we introduce a new framework for Random Forest (RF) by incorporating Bayesian estimation and a hybrid variable selection approach within the splitting step of RF. This innovation aims to enhance prediction efficiency and improve the consistency of variable selection in high-dimensional datasets.

## 2. Random Forest and Sum of Trees Models

Suppose we let $D = [y_i, x_{i1}, x_{i2}, \ldots, x_{ip}], i = 1, 2, \ldots, n$ be an $n \times p$ dataset with $y_i$ assuming continuous values and $x = [x_{i1}, x_{i2}, \ldots, x_{ip}]$ be the vector of $p$ covariates. Thus, we can define a single regression tree using the formulation of [22]

$$y_i = f(x_{i1,}, x_{i2}, \ldots, x_{ip}) + \epsilon_i. \tag{1}$$

where the random noise seen during estimation is denoted by the variable $epsilon_i$, which is assumed to have an independent, identical Gaussian distribution with a mean of 0 and a constant variance of $sigma2$. Consequently, in the same formulation by [22], a sum of trees model can be defined as:

$$y_i = h(x_{i1,}, x_{i2}, \ldots, x_{ip}) + \epsilon_i. \tag{2}$$

where $h(x_{i1,}, x_{i2}, \dots, x_{ip}) = \sum_{j=1}^{J} f(x_{i1,}, x_{i2}, \dots, x_{ip})$ and the total number of trees in the forest is $J$. In the notation of a tree, we have;

$$y = \sum_{j=1}^{J} \Im_j(\beta_{mj} : x \in R_{mj}) \tag{3}$$

where $\beta_m$ is an estimate of $y$ in region $R_m$, $\Im_j(\beta_{mj} : x \in R_{mj})$ is the single regression tree.

The model's parameters $\beta_m$ and $\sigma^2$ (3) are often estimated using the frequentist approach. These approaches include Bagging [16], Stacking [24], Boosting [25] and Random forest [19]. Stacking improves the performance of the single tree in the model (1) by forming a linear combination of different covariates in $x$. Similarly, boosting improves (1) by fitting a single tree model on data sample points not used by an earlier fitted tree. Bagging and random forest iteratively randomize the sample data $D$ and fitting (1) on each uniquely generated sample. Random forest improves over bagging by using a random subset of $p$ covariates, often denoted $mtry$, to fit (1) instead of all $p$ covariates. This procedure has been shown to dramatically lower the correlations between adjacent trees and thus reduce the overall prediction risk of (3). Freidman [26] defined the risk of using the random forest for estimating (3) as:

$$Var(\hat{y}_{RF}) = \rho(x)\sigma^2(x) \tag{4}$$

where $\rho(x)$ is the pairwise correlation between adjacent trees and $\sigma^2(x)$ is the variance of any randomly selected tree from the forest. Equation (4) implies that $\rho(x)$ plays a vital role in shrinking the risk towards 0 as $J \rightarrow \infty$. Achieving infinite forests is rarely possible due to computational difficulty. This drawback has necessitated the development of Bayesian [27] alternatives that are adaptive in nature in terms of averaging many posterior distributions. Chipman [22] proposed the Bayesian Additive Regression Trees (BART) that average many posterior distributions of single Bayesian Classification and Regression trees (BCART, [28]). The approach is specifically similar to a form of boosting proposed in [20]. BART boosts weak trees by placing some form of deterministic priors on them. Although the empirical bake-off results of 42 datasets used to test the BART procedure showed improved performance over RF, there is no theoretical backup on why BART is better than RF. Several authors have queried the improvement, especially in a high-dimensional setting where RF still enjoys moderate acceptance. Hernandez [13] claimed the BART algorithm implemented in R as package "bartMachine" is memory hungry even at moderate $p$ with $T$ the number of MCMC iterations for posterior sampling fixed at 1000.

Taddy [29] proposed the Bayesian and Empirical Bayesian Forests (BF and EBF) to maintain the structure of RF and modify the data randomisation technique using a Bayesian approach. BF replaces the uniform randomisation with the Dirichlet posterior distribution of sampled observation. Similarly, EBF considered the hierarchical parameter structure of estimating the next stage prior hyperparameters using current data. The results from the empirical analysis showed that BF and EBF are not different from RF except in the aspect of model interpretation.

### *2.1. Variable Selection Inconsistency and Inefficiency of Random Forest in Sparse High-Dimensional Setting*

**Definition 1.** *Let $D_{hd} = [Y|X]$ be a partitioned matrix composed of $n \times 1$ response variable $Y$ vector and $X$ be an $n \times p$ matrix with $y_i, x_{ik} \in \Re$, for $i = 1, \dots, n$ and $k = 1, \dots, p$ then a rectangular matrix $D_{hd}$:*

$$D_{hd} = \left[\begin{array}{c|cccccc} y_1 & x_{11} & x_{12} & x_{13} & x_{14} & \dots & x_{1p} \\ y_2 & x_{21} & x_{22} & x_{23} & x_{24} & \dots & x_{2p} \\ y_3 & x_{31} & x_{32} & x_{33} & x_{34} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & x_{n1} & x_{n2} & x_{n3} & x_{n4} & \dots & x_{np} \end{array}\right] \tag{5}$$

*is referred to as high-dimensional data matrix if $p \gg n$ [8].*

If we redefine some of the columns in $D_{hd}$ such that the entries are zeros, thus truncating the matrix structure, we have a sparsed HD. Sparsity is inherent in HD where only a few of the $p$ covariates $x$S are usually related to the response $y$.

**Definition 2.** *A typical sparsed HD matrix is given by*

$$D_{shd} = \left[\begin{array}{c|cccccc} y_1 & x_{11} & 0 & x_{13} & 0 & 0 & x_{1p} \\ y_2 & x_{21} & 0 & x_{23} & 0 & 0 & x_{2p} \\ y_3 & x_{31} & 0 & x_{33} & 0 & 0 & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & x_{n1} & 0 & x_{n3} & 0 & 0 & x_{np} \end{array}\right] \tag{6}$$

*in Olaniran & Abdullah [30].*

The risk of random forests (RF), as indicated in equation (4), increases in high-dimensional situations with a sparse covariate matrix. This increase is a consequence of the random selection of the variable set $mtry$ used to build equation (1). The method of hypergeometric sampling for selecting $mtry \subset p$ is notably sensitive to the mixing parameter $\pi$, which represents the proportion of relevant variables. Specifically, when the proportion of relevant variables ($\pi$) is higher, the risk associated with RF is lower, and vice versa.

It's worth noting that many software implementations of RF regression commonly set $mtry$ as $p/3$. This choice is expected to yield a satisfactory set of covariates for predicting the target variable $y$ under two conditions: when the data matrix $D$ is of low dimensionality or when the number of observations ($n$) is greater than the number of covariates ($p$).

In the specific scenario where $D = D_{shd}$, theorem (1) is employed to establish an upper bound for the probability of correctly selecting at least one relevant covariate from the set of $p$. This theorem is the foundation for defining the selection consistency of random forests in high-dimensional, sparse settings.

**Theorem 1.** *Given p covariates, $r \subset p$ relevant variables, if we set the RF subsample size $mtry = p/3$ as r, then the probability that at least one of the covariates in r is relevant converges to $1 - e^{-1}$ as $p \to \infty$.*

**Proof.**

**Proposition 1.** *Let $R_1, R_2, \ldots R_r$ denote the event that the $R_k$ covariate is relevant. Then, the event that at least one covariate in r is relevant is $R_1 \cup R_2 \cup \ldots R_r$, the required probability is $P(R_1 \cup R_2 \cup \ldots R_r)$. It is worthy of note that the subsample selection done by RF is without replacement, implying that the sample space of p covariates can be partitioned into two (R relevant covariates and $p - R$ irrelevant covariates). This partitioning done without replacement is often referred to as a hypergeometric or non-mutually exclusive process [31]. Thus, by the generalization of the principle of non-mutually exclusive events, we have*

$$\begin{aligned} P(R_1 \cup R_2 \cup \ldots R_r) = \sum_{k=1}^{r} P(R_k) - \sum_{j,k=1;k>j}^{r} P(R_j \cap R_k) + \sum_{i,j,k=1;k>j>i}^{r} P(R_i \cap R_i \cap R_k) \\ - \cdots + (-1)^{r-1} P(R_1 \cap R_2 \cap \cdots \cap R_r) \end{aligned} \tag{7}$$

doi:10.20944/preprints202310.1272.v1

$$
\begin{aligned}
P(R_1 \cup R_2 \cup \ldots R_r) &= \binom{r}{1}\left(\frac{1}{r}\right) - \binom{r}{2}\left(\frac{1}{r}\right)\left(\frac{1}{r-1}\right) + \binom{r}{3}\left(\frac{1}{r}\right)\left(\frac{1}{r-1}\right)\left(\frac{1}{r-3}\right) \\
&\quad - \cdots + (-1)^{r-1}\binom{r}{r}\left(\frac{1}{r!}\right) \\
&= 1 - \left(\frac{1}{2!}\right) + \left(\frac{1}{3!}\right) - \cdots + (-1)^{r-1}\left(\frac{1}{r!}\right)
\end{aligned} \tag{8}
$$

Recall that the exponential function $e^{\xi} = 1 + \xi + \frac{\xi^2}{2!} + \frac{\xi^3}{3!} + \cdots + \frac{\xi^r}{r!}$, if $\xi = -1$,

$$
e^{-1} = 1 - \left(1 - \frac{1}{2!} + \frac{1}{3!} + \cdots + (-1)^{r-1}\frac{1}{r!}\right) \tag{9}
$$

Thus,

$$
P(R_1 \cup R_2 \cup \ldots R_r) = 1 - e^{-1} = 0.6321 \tag{10}
$$

□

Theorem (1) implies that when $p$ grows infinitely, the maximum proportion of relevant covariates that would be selected is 63.2% assuming the number of subsample $mtry$ chosen equals the number of relevant covariates in $p$. Figure 1 shows the convergence over varying $p$ covariates.

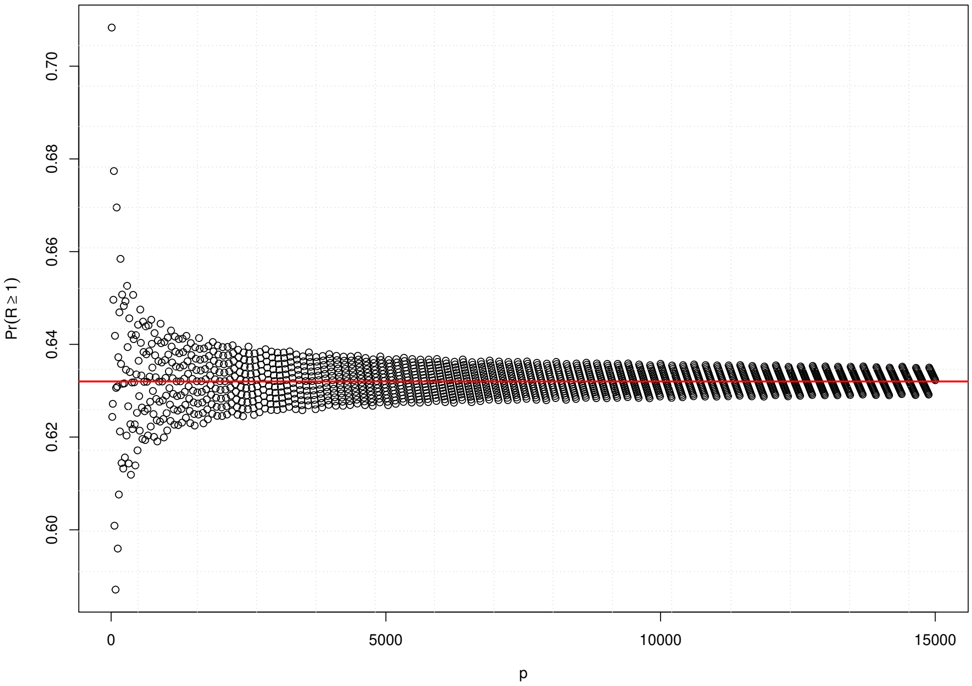

**Figure 1.** Probability of selecting relevant variables for RF at varying dimensionality $p$.

**Lemma 1.** *RF variable selection is consistent if* $\lim_{p\to\infty} P(\hat{\mathcal{M}} = \mathcal{M}) = 1$.

**Remark 1.** *Lemma (1) indicates that for an RF model $\hat{\mathcal{M}}$ fitted using r subsampled covariates, RF variable selection is consistent if the fitted model converges almost surely to the true model $\mathcal{M}$ that contains all relevant variables R.*

**Corollary 1.** $\lim_{p\to\infty} P(\bigcup_{k=1}^{r\subset p} R_k) = 0.6321 < 1$, *then RF variable selection is inconsistent in HD with large p.*

Now that we have established the inconsistency of RF in the HD setting, the following lemma presents RF variance for $D_{shd}$ matrix.

**Lemma 2.** *Let $\pi$ be the proportion of relevant covariates and $0 \le \pi \le 1$, the variance of a single tree $\sigma^2(x)$ can be decomposed into random and noise variance such that the risk of RF in a high-dimensional setting with $p \gg n$ can be defined as:*

$$Var(\hat{y}_{RF}) = \rho(x)\left[\pi\sigma_1^2(x) + (1-\pi)\sigma_2^2(x)\right] \tag{11}$$

*where $\sigma_1^2(x)$ and $\sigma_2^2(x)$ are the random and noise variances respectively.*

**Remark 2.** *It is clear from lemma (2) that the risk of RF in (11) is larger than (4) when $\pi < 1$, thus RF violates the oracle property conditions defined by [32–34] among others as:*

i. *Identification of the right subset model $\mathcal{M}$ such that $P(\hat{\mathcal{M}} = \mathcal{M}) \to 1$.*
ii. *Achievement of the optimal estimation rate, $\sqrt{n}(\hat{\mathcal{M}} - \mathcal{M}) \xrightarrow{d} N(0, Var(\mathcal{M}))$*

Many authors have argued that a good estimator $\hat{\mathcal{M}}$ should satisfy these oracle properties. However, from theorem (1) and (2), RF fails to achieve these conditions in the sparse HD setting. Thus, there is a need to propose an alternative procedure that enjoys these attractive properties.

## 3. Priors and Posterior Specification of Bayesian Random Forest for Sparse HD

The Bayesian Random Forest (BRF) proposed here has three major prior parameters. The first is model uncertainty prior defined over tree $\mathfrak{I}$. Here we propose a uniform prior $\mathfrak{I} \sim U(0,1)$ by [35] such that $Pr(\mathfrak{I}) = 1$ for any candidate tree. We used this prior specification to retain the average weighing procedure of RF so that each tree $\mathfrak{I}_j$ has an equal voting right. The core advantage of this prior is to retain RF's strength in correcting the over-fitting problem by averaging over all trees. The second form of prior is terminal node parameter $\beta_M$ and $\sigma^2$ prior, here we propose the Normal Inverse Gamma prior $NIG(\mu_M, \sigma^2 \sum, a_0, b_0)$ by [35], where $\mu_M$ and $\Sigma$ are the prior mean and covariance for parameter $\beta_M$, $a_0$ and $b_0$ are the prior sample size and sum of squares for response $y_i$ on parameter $\sigma^2$. Furthermore, we assumed a conditional prior of trees parameters $\beta_M$ on $\sigma^2$, that's for a single tree with $M$ terminal nodes:

$$P(\beta_M, \sigma^2) = P(\beta_M|\sigma^2)P(\sigma^2). \tag{12}$$

This can be easily extended to $J$ trees with the assumption of constant model variance $\sigma^2$ over all trees. Thus we have;

$$P(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) = \prod_{j=1}^{J} P(\mathfrak{I}_j, \beta_{Mj})P(\sigma^2) \tag{13}$$

$$P(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) = \prod_{j=1}^{J} P(\beta_{Mj}|\sigma^2)P(\mathfrak{I}_j)P(\sigma^2) \tag{14}$$

with $Pr(\mathfrak{I}_j) = 1$

$$P(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) = \left[\prod_{j=1}^{J} P(\beta_{Mj}|\sigma^2)\right]P(\sigma^2) \tag{15}$$

$$P(\beta_{Mj}|\sigma^2) = \frac{\exp\left[\frac{-1}{2\sigma^2}(\beta_{Mj} - \mu_{Mj})' \sum_j^{-1}(\beta_{Mj} - \mu_{Mj})\right]}{(\sqrt{2\pi})^M |\sigma^2 \sum_j|^{\frac{1}{2}}} \tag{16}$$

If we assumed that the trees are independent and identically distributed, then,

$$\prod_{j=1}^{J} P(\beta_{Mj}|\sigma^2) \sim N(J\mu_M, J\sigma^2\Sigma)$$

$$\prod_{j=1}^{J} P(\beta_{Mj}|\sigma^2) = \frac{\exp\left[\frac{-1}{2J\sigma^2}(\beta_{Mj} - J\mu_{Mj})'\Sigma_j^{-1}(\beta_{Mj} - J\mu_{Mj})\right]}{(\sqrt{2\pi})^{JM}|\sigma^2 J \sum_j|^{\frac{J}{2}}} \tag{17}$$

$$Pr(\sigma^2) = \frac{b_0^{a_0}(\sigma^2)^{-a_0-1}\exp(-b_0/\sigma^2)}{\Gamma(a_0)} \tag{18}$$

$$\begin{aligned} Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) &= \frac{\exp\left[\frac{-1}{2J\sigma^2}(\beta_{Mj} - J\mu_{Mj})'\sum_j^{-1}(\beta_{Mj} - J\mu_{Mj})\right]}{(\sqrt{2\pi})^{JM}|\sigma^2 J \sum_j|^{\frac{J}{2}}} \\ &\times \frac{b_0^{a_0}(\sigma^2)^{-a_0-1}\exp(-b_0/\sigma^2)}{\Gamma(a_0)} \end{aligned} \tag{19}$$

$$\begin{aligned} Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) &= \frac{b_0^{a_0}(\sigma^2)^{-(a_0+(J/2)+1)}}{\Gamma(a_0)(\sqrt{2\pi})^{JM}|J\sum_j|^{\frac{J}{2}}} \\ &\times \frac{\exp\left\{\frac{-1}{2\sigma^2}\left[(\beta_{Mj} - J\mu_{Mj})'J^{-1}\sum_j^{-1}(\beta_{Mj} - J\mu_{Mj}) + 2b_0\right]\right\}}{\Gamma(a_0)(\sqrt{2\pi})^{JM}|J\sum_j|^{\frac{J}{2}}} \end{aligned} \tag{20}$$

The Bayes theorem leads to the posterior density of trees;

$$\begin{aligned} &Pr(\mathfrak{I}_1, \ldots, \mathfrak{I}_J|y, x) = \\ &\frac{\left[\prod_{j=1}^{J} Pr(\beta_{Mj}|\sigma^2)\right] Pr(\sigma^2) L(y, x|\mathfrak{I}_1, \ldots, \mathfrak{I}_J)}{\int_{\beta_{Mj}}\int_{\sigma^2}\left[\prod_{j=1}^{J} Pr(\beta_{Mj}|\sigma^2)\right] Pr(\sigma^2) L(y, x|\mathfrak{I}_1, \ldots, \mathfrak{I}_J) d\beta_{Mj} d\sigma^2} \end{aligned} \tag{21}$$

The integral at the denominator of equation (21) cannot be solved analytically thus, it is often dropped in most Bayesian analyses suggested by [35] and hence, we proceed as;

$$Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J|y, x) \propto \left[\prod_{j=1}^{J} Pr(\beta_{Mj}|\sigma^2)\right] Pr(\sigma^2) L(y, x|\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J). \tag{22}$$

The likelihood of $J$ trees can be defined as;

$$L(y, x|\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) = \prod_{j=1}^{J} L(y, x|\mathfrak{I}_j) \tag{23}$$

$$L(y, x|\mathfrak{I}_j) = \frac{\exp\left[\frac{-1}{2\sigma^2}(y - \beta_{Mj})'(y - \beta_{Mj})\right]}{(\sqrt{2\pi\sigma^2})^n} \tag{24}$$

$$L(y, x|\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J) = \frac{\exp\left[\frac{-1}{2\sigma^2}\sum_{j=1}^{J}(y - \beta_{Mj})'(y - \beta_{Mj})\right]}{(\sqrt{2\pi\sigma^2})^{Jn}}. \tag{25}$$

Therefore, the posterior of Bayesian Random Forest regression is;

$$\begin{aligned} Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J | y, x) \propto\ & \frac{(\sigma^2)^{-(a_0+(J/2)+1)}}{\Gamma(a_0)(\sqrt{2\pi})^{JM}|J\Sigma_j|^{\frac{J}{2}}} \\ & \times \frac{\exp\left\{\frac{-1}{2\sigma^2}\left[(\beta_{Mj} - J\mu_{Mj})' J^{-1}\Sigma_j^{-1}(\beta_{Mj} - J\mu_{Mj}) + 2b_0\right]\right\}}{\Gamma(a_0)(\sqrt{2\pi})^{JM}|J\Sigma_j|^{\frac{J}{2}}} \\ & \times \frac{\exp\left[\frac{-1}{2\sigma^2}\sum_{j=1}^{J}(y - \beta_{Mj})'(y - \beta_{Mj})\right]}{(\sqrt{2\pi\sigma^2})^{Jn}}. \end{aligned} \tag{26}$$

$$\begin{aligned} Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J | y, x) \propto\ & (\sigma^2)^{-(a_1+(J/2)+1)} \\ & \times \exp\left\{\frac{-1}{2\sigma^2}\left[(\beta_{Mj} - J\mu_{Mj}^1)'(J\Sigma_j^1)^{-1}(\beta_{Mj} - J\mu_{Mj}^1) + 2b_1\right]\right\} \end{aligned} \tag{27}$$

where;

$$J\mu_{Mj}^1 = [(J\Sigma_j)^{-1} + (JV_j)]^{-1}\left[(J\Sigma_j)^{-1}\mu_{Mj} + (JV_j)^{-1} n_{Mj}\bar{y}_{Mj}\right] \tag{28}$$

where $V_j$ is an $m \times m$ matrix of data information such that the diagonal of $V_j$ is $\sigma_{mj}^{-2}$ which is defined as;

$$\sigma_{mj}^{-2} = \frac{n_{mj} - 1}{\sum_{i=1}^{n_{mj}}(y - \bar{y}_{mj})^2} \tag{29}$$

$$J\Sigma_j^1 = [(J\Sigma_j)^{-1} + (JV_j)]^{-1} \tag{30}$$

$$a_1 = a_0 + n/2 \tag{31}$$

$$b_1 = b_0 + \left(\frac{\mu_{Mj}'(\Sigma_j)^{-1}\mu_{Mj} + y'y - \mu_{Mj}^{1'}(\Sigma_j^1)^{-1}\mu_{Mj}^1}{2}\right) \tag{32}$$

The marginal densities of $Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J | y, x)$ is important when performing inference about $\beta_M$ and $\sigma^2$. The marginal density of $\beta_M$ is given by;

$$Pr(\beta_M | \mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J, y, x) = \int_{\sigma^2} Pr(\mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J | y, x) d\sigma^2 \tag{33}$$

$$\begin{aligned} Pr(\beta_M | \mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J, y, x) = & \int_{\sigma^2} (\sigma^2)^{-(a_1+(J/2)+1)} \\ & \times \exp\left[\frac{-1}{2\sigma^2}(\beta_{Mj} - J\mu_{Mj}^1)'(J\Sigma_j^1)^{-1}\right. \\ & \left. \times (\beta_{Mj} - J\mu_{Mj}^1) + 2b_1\right] d\sigma^2 \end{aligned} \tag{34}$$

According to [36], the marginal distribution is identical to the $student - t$ distribution defined as:

$$Pr(\beta_M | \mathfrak{I}_1, \mathfrak{I}_2, \ldots, \mathfrak{I}_J, y, x) \sim t(\mu_{Mj}^1, s^2\Sigma_j^1, 2a_1)$$

where,

$$s^2 = \frac{b_1}{a_1 - 1} = \frac{b_0 + \left(\frac{\mu_{Mj}'(\Sigma_j)^{-1}\mu_{Mj} + y'y - \mu_{Mj}^{1'}(\Sigma_j^1)^{-1}\mu_{Mj}^1}{2}\right)}{a_0 + n/2 - 1} \tag{35}$$

Therefore, the posterior mean for $\beta_M$ is;

$$\hat{\beta}_M = J^{-1}\left\{[(J\Sigma_j)^{-1} + (JV_j)]^{-1}\left((J\Sigma_j)^{-1}\mu_{Mj} + (JV_j)^{-1}n_{Mj}\bar{y}_{Mj}\right)\right\} \tag{36}$$

and the posterior variance for $\beta_M$ is;

$$var(\beta_M) = J^{-1}\left\{\frac{2a_1s^2}{2a_1-2}[(J\Sigma_j)^{-1} + (JV_j)]^{-1}\right\} \tag{37}$$

The posterior mean of $\beta_M$ can be interpreted as the weighted average of prior mean $\mu_M$ and data mean $\bar{y}_M$. The scaling factor is the joint contribution of prior and data information. Similarly, the posterior variance of $\beta_M$ is the scaled form of the joint contribution of data and prior information matrix. The parameters $\beta_M$ and $\sigma^2$ can be extracted from their posterior density using a hybrid of Metropolis-Hastings and Gibbs sampler algorithms described as follows:

### 3.0.1. Hybrid Gibbs and MH procedure for extracting posterior information from Bayesian Random Regression Forest with Gaussian Response

1. **Step 0:** Define initial values for $\beta^0_{Mj}$ and $(\sigma^2)^0$ such that $Pr(\beta^0_{Mj}|y,x) > 0$ and $Pr[(\sigma^2)^0] > 0$.
2. **Step 1:** For $v = 1, 2, \ldots, V$
3. **Step 2:** Sample $\tilde{\sigma}^2$ from lognormal distribution;$q_1(\tilde{\sigma}^2, \nu_1) = LN[(\tilde{\sigma}^2)^{v-1}, \nu_1]$.
4. **Step 3:** For $j = 1, 2, \ldots, J$ trees
5. **Step 4:** Sample $\tilde{\beta}_{Mj}$ from independent multivariate normal distribution $q_2(\tilde{\beta}_{Mj}, \nu_2) = INM(\beta^{v-1}_{Mj}, \nu_2)$.
6. **Step 5:** Calculate the moving probability for $\beta_{Mj}$ by;

$$\pi_1(\beta^v_{Mj}, \tilde{\beta}_{Mj}) = \min\left[\frac{Pr(\tilde{\beta^v}_{Mj}|y,x)}{Pr(\beta^v_{Mj}|y,x)}, 1\right]$$

7. **Step 6:** Sample $U_1 \sim U(0,1)$; then

$$\beta^v_{Mj} = \begin{cases} \tilde{\beta}_{Mj} & \text{if } U_1 \leq \pi_1(\beta^v_{Mj}, \tilde{\beta}_{Mj}); \\ \beta^{v-1}_{Mj} & \text{if } U_1 > \pi_1(\beta^{v-1}_{Mj}, \tilde{\beta}_{Mj}). \end{cases}$$

8. **Step 7:** Compute the residuals $\epsilon_i = y_i - J^{-1}\sum_{j=1}^{J}\Im_j(\beta^v_{Mj} : x \in R_{Mj})$
9. **Step 8:** Calculate the moving probability for $\sigma^2$ by;

$$\pi_2[(\sigma^2)^v, \tilde{\sigma}^2] = \min\left\{\frac{Pr(\tilde{\sigma}^2|\epsilon_i)q_1[(\sigma^2)^{v-1}|\tilde{\sigma}^2, \nu_1]}{Pr[(\sigma^2)^{v-1}|\epsilon_i]q_1[\tilde{\sigma}^2|(\sigma^2)^{v-1}, \nu_1]}, 1\right\}$$

10. **Step 9:** Sample $U_2 \sim U(0,1)$; then

$$(\sigma^2)^v = \begin{cases} \tilde{\sigma}^2 & \text{if } U_2 \leq \pi_2[(\sigma^2)^v, \tilde{\sigma}^2]; \\ (\sigma^2)^{v-1} & \text{if } U_2 > \pi_2[(\sigma^2)^v, \tilde{\sigma}^2]. \end{cases}$$

The proposed algorithm is a combination of the Metropolis-hasting algorithm and Gibbs sampler. It is a Metropolis Hasting algorithm with further updates on $\sigma^2$ using Gibbs sampler. The algorithm's validity was demonstrated with a simulated response variable $y$ scenario with no predictor variable. A regression tree with three terminal nodes was assumed. Each terminal node of the regression tree consists of 5 observations with mean $\beta_m = 10$; variance $\sigma^2_m = 4$; $m = 1, 2, 3$. The regression tree was replicated 5 times using bootstrapping to make a forest. The MCMC iteration results are shown below. The first column shows the trace plot of the parameter for each node over the trees. The trace plot for the four parameters shows that the iterations converge very sharply at 10000. The autocorrelation plots show an exponential decay, suggesting independent MCMC chains. The acceptance rate lies within the tolerable range of $20\% - 40\%$ as suggested by [35]. These features established the validity of the algorithm. In addition, the histogram supports the analytical densities proposed for the posterior distribution of parameters. The posterior densities for $\beta_m$ is very much closer to student $t$ distribution with values at the tail end while $\sigma^2$ density is very much closer to Gamma. The overall model standard error estimate using the Bayesian Random Forest (BRF) algorithm is $\sigma_{brf} = 1.42$ while that of frequentist estimate is $\sigma_{rf} = 2.04$. This established that empirically, BRF is more efficient than the frequentist RF method.

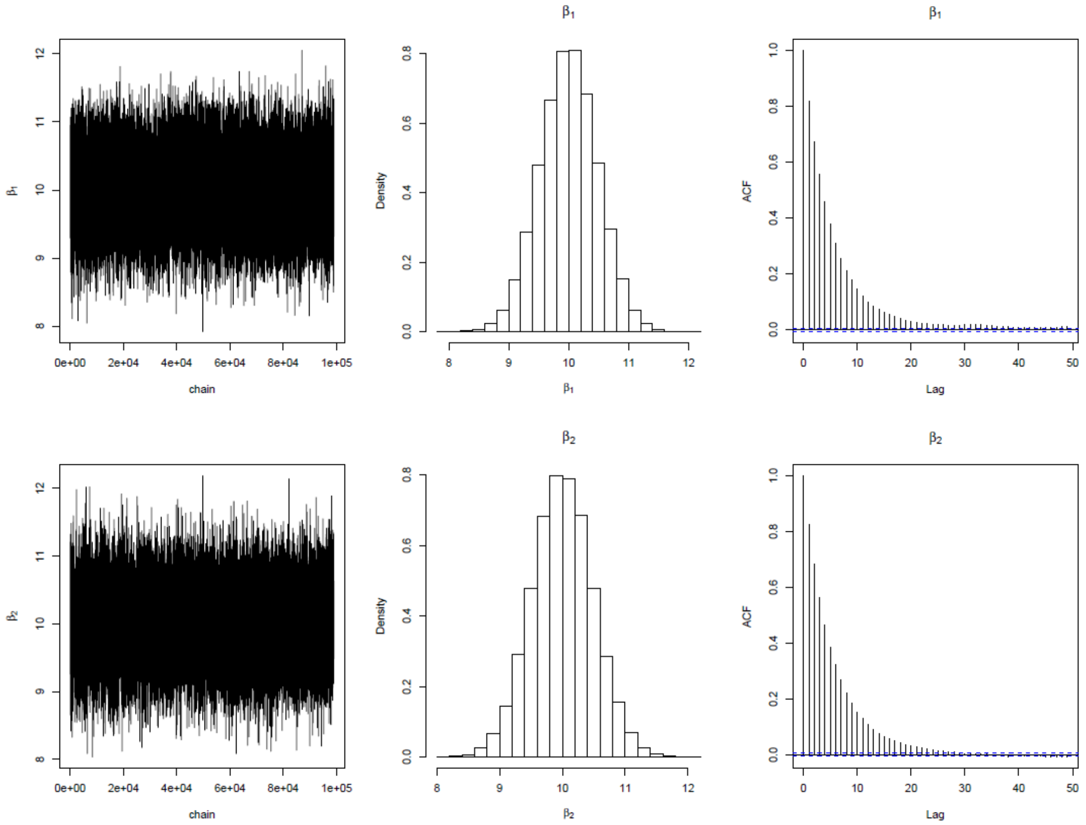

**Figure 2.** Simulation plot 1 of the hybrid algorithm for Bayesian Random Forest regression.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 19 October 2023 doi:10.20944/preprints202310.1272.v1

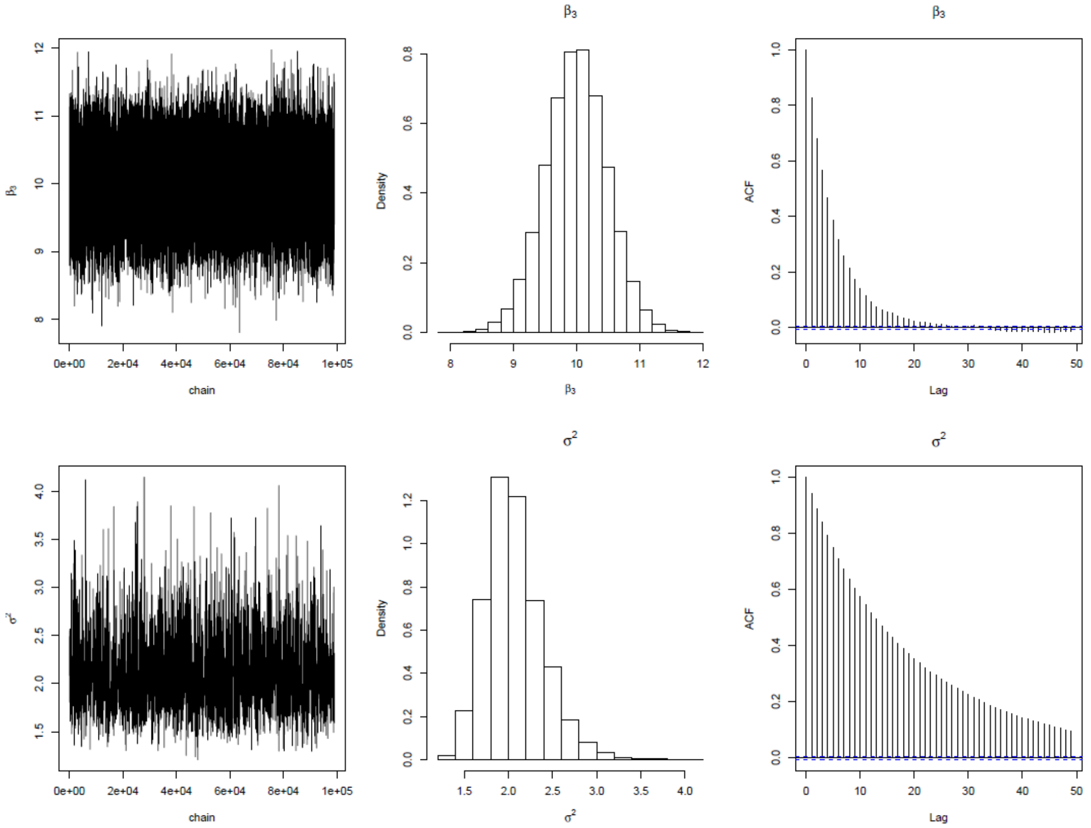

**Figure 3.** Simulation plot 2 of the hybrid algorithm for Bayesian Random Forest regression.

The performance of the BRF procedure was also examined for predicting response $y$ given three covariates that correspond to genes. The three covariates are denoted as $Gene1$, $Gene2$, $Gene3$. The functional relation between the response and genes was defined as:

$$y_i = 5 + 10 \times Gene1 + 20 \times Gene2 + 30 \times Gene3 + \epsilon_i$$

where $i = 1, 2, \ldots, 30$. Figures 4 and 5 shows a single tree diagram from a Bayesian random forest consisting of 5 trees. The plot shows simulated 20 terminal nodes.

The first white box in Figure 4 housed value 2.4, corresponding to the split decision value. Variable Gene3 split into two daughter nodes with the left node terminating on predicted $\hat{y} = 13.39$ based on the condition that $Gene3 < 2.4$. Subsequently, the right node further splits into another two daughter nodes using condition $Gene3 < 0.69$. Again, the left daughter node terminates on predicted $\hat{y} = 17.17$. The process continues until the maximum number of nodes set to be 20 is reached. The number of nodes condition is also referred to as maximal tree depth by [6]. The reason for more splits on $Gene3$ can be easily observed from how the response was simulated such that $Gene3$ is the most relevant in terms of weights. If the variable importance score is desired, $Gene3$ will be the most important variable since it occurs more frequently than others in the predictor set. The posterior estimates in equations (36) and (37) obtained earlier rely on the accuracy of prior parameter values assumed. In most occasions, searching for appropriate prior parameters may be difficult, especially in the case of the sum of trees model. A data-driven approach is often used, such as those used in [22,28]. Another alternative is the Empirical Bayes. Empirical Bayes [37–39] allows the experimenter to estimate the prior hyperparameter values from the data. It is a hierarchical modelling approach where the parameter of a second stage or later model depends on initial stage data. The empirical Bayes approach is often used when hierarchical data are available. However, it can also be applied for non-hierarchical situations as extended by [40] using bootstrapped data to construct confidence intervals. The sum of trees modelling strategy is thus further simplified using the bootstrap prior technique. The approach was used to obtain the prior hyperparameters $\mu_M$, $\Sigma$ for each tree. The major advantage of bootstrap prior is that it guarantees an unbiased estimate of $\beta_M$ for each tree. To achieve a fast grasp of the approach, we consider Bayesian inference of a single tree $j$ with one of the terminal node parameters defined by $\beta_m$ and $\lambda_m$. The likelihood of a single tree $\Im_j(\beta_m : x \in R_m)$, $L[y, x|\Im_j(\beta_m : x \in R_m)]$ can be written as;

$$L[y,x|\Im_j(\beta_m : x \in R_m)] = \left(\frac{\lambda_{mj}^{\frac{1}{2}}}{\sqrt{2\pi}}\right)^{n_{mj}} \exp\left[-\frac{\lambda_{mj}}{2}\sum_{i:x\in R_{mj}}^{n_{mj}}(y_i-\beta_{mj})^2\right] \tag{38}$$

where $\lambda_{mj} = \sigma_{mj}^{-2}$ interpreted as precision for node $m$. Correspondingly, we can write the prior density for the parameters of a single tree as;

$$\begin{aligned} Pr(\beta_{mj}, \lambda_{mj}|n_0, a_0, b_0) = & \frac{(\lambda_{mj}n_0)^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda_{mj}}{2}(\beta_{mj}-\mu_0)^2\right] \\ & \times \frac{b_0^{a_0}\lambda_{mj}^{a_0-1}\exp\Big(-\lambda_{mj}b_0\Big)}{\Gamma(a_0)} \end{aligned} \tag{39}$$

where $n_0$ is the prior sample size for terminal node $m$, $\mu_0$ is the prior mean obtained from $n_0$, $a_0$ is the prior sample size for the precision $\lambda_{mj}$ and $b_0$ is the prior sum of squares deviation from prior mean $\mu_0$. The posterior distribution of a single tree thus follows from the Bayes theorem:

$$\begin{aligned} & Pr(\beta_{mj}, \lambda_{mj}|y, x) = \\ & \frac{Pr(\beta_{mj}, \lambda_{mj}) \times L[y, x|\Im_j(\beta_m : x \in R_m)]}{\int_{\beta_{mj}}\int_{\lambda_{mj}} Pr(\beta_{mj}, \lambda_{mj}) \times L[y, x|\Im_j(\beta_m : x \in R_m)]d\beta_{mj}d\lambda_{mj}}. \end{aligned} \tag{40}$$

After a little arrangement, the posterior distribution can be defined as:

$$\begin{aligned} Pr(\beta_{mj}, \lambda_{mj}|y, x) = & \frac{(\lambda_{mj}n_1)^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{\lambda_{mj}}{2}(\beta_{mj}-\mu_1)^2\right] \\ & \times \frac{b_1^{a_1}\lambda_{mj}^{a_1-1}\exp\Big(-\lambda_{mj}b_1\Big)}{\Gamma(a_1)} \end{aligned} \tag{41}$$

where $\mu_1 m = (n_0\mu_0 + n_m\bar{y}_m)/(n_0 + n_m)$ is the posterior estimate of $\beta_m$, $n_1 = n_0 + n_m$ is the posterior sample size for which $\mu_1 m$ can be estimated, $a_1 = a_0 + n_m/2$ is the posterior sample size for which $\lambda_m$ can be estimated and $b_1 = b_0 + 1/2\sum_{i=1}^{n_m}(y_i - \bar{y}_m)^2 + \frac{n_0 n_m(\bar{y}_m-\mu_0)^2}{2(n_0+n_m)}$. The terminal node estimate $\bar{y}_m$ is defined as;

$$\bar{y}_m = (n_m)^{-1}\sum_{i=1}^{n_m}(y_i|x_i \in R_m) \tag{42}$$

is the maximum likelihood estimate of $E(y_i|x_i \in R_m)$. The estimate is unbiased and that is used in RF algorithms. The variance of the estimate followed as;

$$var(\bar{y}_m) = n_m^{-1}\sigma_m^2 \tag{43}$$

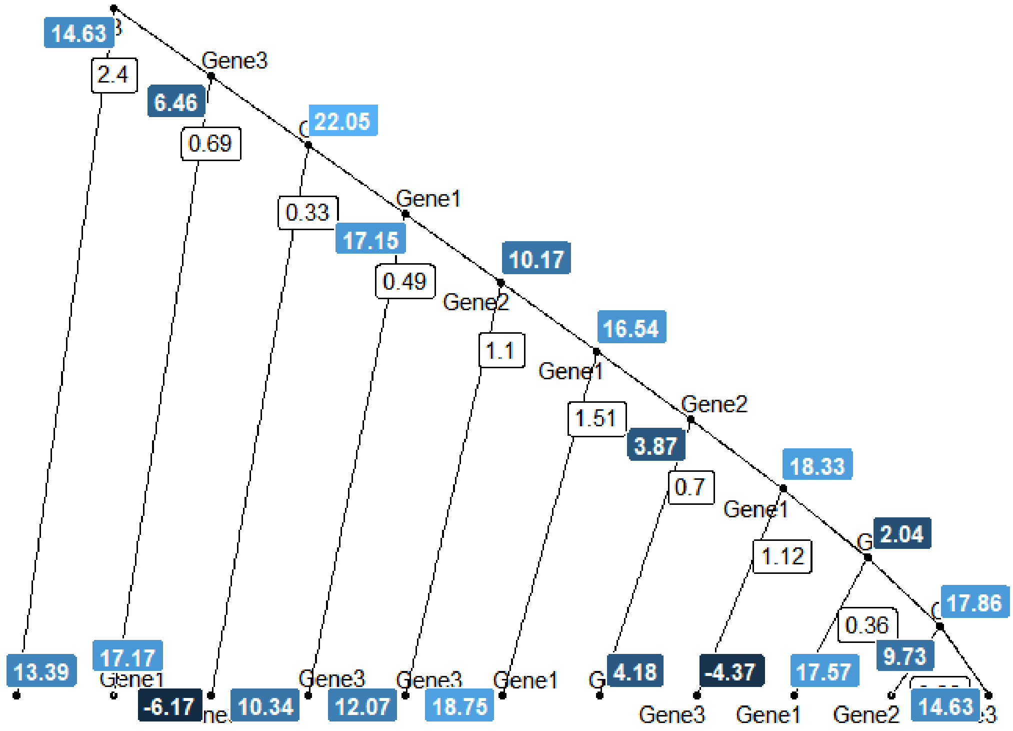

**Figure 4.** Single regression tree plot from forest of five trees using Bayesian Random Forest (BRF) hybrid Gibbs and MH procedure. The coloured box corresponds to the terminal node or final prediction, and the white box corresponds to the decision node or split point.

The corresponding empirical bayes estimate for equations (42) and (43) are;

$$\hat{\mu}_{[EB]m} = (\hat{n}_0\hat{\mu}_0 + n_m\bar{y}_m)/(\hat{n}_0 + n_m) \tag{44}$$

$$\hat{\sigma}^2_{[EB]m} = \frac{\hat{b}_0 + 1/2\sum_{i=1}^{n_m}(y_i - \bar{y}_m)^2 + \frac{\hat{n}_0 n_m(\bar{y}_m - \hat{\mu}_0)^2}{2(\hat{n}_0 + n_m)}}{\hat{a}_0 + n_m/2} \tag{45}$$

As a further update on the empirical Bayes procedure, the prior hyperparameters are estimated from a bootstrapped sample by following the procedure below:

1. Creating $B$ bootstrap samples $y_b$ from the initial sample $y_m$ in the terminal node $m$,
2. Estimating the hyperparameters (prior parameters) each time the samples are generated using the Maximum Likelihood (ML) method,
3. Updating the posterior estimates using the hyperparameters in step (2) above using equations (44) and (45),
4. Then obtaining the bootstrap empirical Bayesian estimates $\hat{\mu}_{BT}$ and $\hat{\sigma}^2_{BT}$ using;

$$\hat{\mu}_{BT} = B^{-1}\sum_{b=1}^{B}\hat{\mu}_{[EB]m_b} \tag{46}$$

$$\hat{\sigma}^2_{BT} = B^{-1}\sum_{b=1}^{B}\hat{\sigma}^2_{[EB]m_b} \tag{47}$$

**Theorem 2.** *The bootstrap prior estimate $\hat{\mu}_{BT}$ is a uniformly minimum variance unbiased estimator of $\beta_m = E(y_i|x_i \in R_m)$ under mild regularity conditions.*

**Proof.**

$$\hat{\mu}_{BT} = B^{-1}\sum_{b=1}^{B}\frac{\hat{n}_0\hat{\mu}_0 + n_m\bar{y}_m}{\hat{n}_0 + n_m} \tag{48}$$

$$\hat{\mu}_{BT} = B^{-1}\sum_{b=1}^{B}\left[\frac{\hat{n}_0\hat{\mu}_0}{\hat{n}_0 + n_m} + \frac{n_m\bar{y}_m}{\hat{n}_0 + n_m}\right] \tag{49}$$

Suppose we fix the prior parameters as; $\hat{n}_{0b} = B$ and $\hat{\mu}_{0b} = \bar{y}_{mb}$, where $\bar{(}y)_{mb}$ is the ML estimate based on a bootstrap sample selected from $y_b$. That is,

$$\bar{y}_{mb} = (n_m)^{-1}\sum_{i=1}^{n_m}(y_{bi}|x_i \in R_m)$$

Then,

$$\hat{\mu}_{BT} = B^{-1}\sum_{b=1}^{B}\left[\frac{B\bar{y}_{mb}}{B + n_m} + \frac{n_m\bar{y}_m}{B + n_m}\right] \tag{50}$$

$$\hat{\mu}_{BT} = \frac{\sum_{b=1}^{B}\bar{y}_{mb}}{B + n_m} + \frac{n_m\bar{y}_m}{B + n_m} \tag{51}$$

$$E[\hat{\mu}_{BT}] = E\left[\frac{\sum_{b=1}^{B}\bar{y}_{mb}}{B + n_m} + \frac{n_m\bar{y}_m}{B + n_m}\right] \tag{52}$$

$$E[\hat{\mu}_{BT}] = \frac{\sum_{b=1}^{B}E[\bar{y}_{mb}]}{B + n_m} + \frac{n_m E[\bar{y}_m]}{B + n_m} \tag{53}$$

Since $\bar{y}_m$ and $\bar{y}_{mb}$ are known unbiased estimates of $\beta_m$,
$\rightarrow$

$$E[\hat{\mu}_{BT}] = \frac{\sum_{b=1}^{B}\beta_m}{B + n_m} + \frac{n_m\beta_m}{B + n_m} \tag{54}$$

$$= \frac{1}{B + n_m}[B\beta_m + n_m\beta_m] \tag{55}$$

$$E[\hat{\mu}_{BT}] = \beta_m \tag{56}$$

Therefore, $\hat{\mu}_{BT}$ is unbiased for estimating $\beta_m$. Also, the MSE is the combination of the square of bias and variance of the estimate, then following from the above derivation the MSE is just the variance of the estimate. Thus,

$$var[\hat{\mu}_{BT}] = var\left[\frac{\sum_{b=1}^{B}\bar{y}_{mb}}{B + n_m} + \frac{n_m\bar{y}_m}{B + n_m}\right] \tag{57}$$

$$= \frac{\sum_{b=1}^{B}var[\bar{y}_{mb}]}{(B + n_m)^2} + \frac{n_m^2 var[\bar{y}_m]}{(B + n_m)^2} \tag{58}$$

$$= \frac{\sum_{b=1}^{B}[n_m^{-1}\sigma_m^2]}{(B + n_m)^2} + \frac{n_m^2[n_m^{-1}\sigma_m^2]}{(B + n_m)^2} \tag{59}$$

$$var[\hat{\mu}_{BT}] = \left[\frac{n_m^2 + B}{(B + n_m)^2}\right]n_m^{-1}\sigma_m^2 \tag{60}$$

Hence, it can be shown that the limiting form of $\left[\frac{n_m^2+B}{(B+n_m)^2}\right]$ is 0, by applying L'hospital rule:

$$\lim_{B\to\infty}\left[\frac{n_m^2+B}{(B+n_m)^2}\right] = \lim_{B\to\infty}\left[\frac{\frac{d(n_m^2+B)}{dB}}{\frac{d(B+n_m)^2}{dB}}\right] \tag{61}$$

$$\lim_{B\to\infty}\left[\frac{n_m^2+B}{(B+n_m)^2}\right] = \lim_{B\to\infty}\left[\frac{1}{2(B+n_m)}\right] \tag{62}$$

$$\lim_{B\to\infty}\left[\frac{n_m^2+B}{(B+n_m)^2}\right] = \frac{1}{\infty} \tag{63}$$

$$\lim_{B\to\infty}\left[\frac{n_m^2+B}{(B+n_m)^2}\right] = 0 \tag{64}$$

□

The derivation above implies that at sample size $n_m$, the $\lim_{B\to\infty} var[\hat{\mu}_{BT}] = 0$. This affirms that the experimenter can control the stability of the estimator by increasing the number of bootstrap samples $B$. In addition, $var[\hat{\mu}_{BT}] = \left[\frac{n_m^2+B}{(B+n_m)^2}\right] n_m^{-1}\sigma_m^2 < var[\hat{\mu}_{ML}] = n_m^{-1}\sigma_m^2$, by a factor $\left[\frac{n_m^2+B}{(B+n_m)^2}\right]$ that converges faster to zero with increasing $B$. Therefore, the frequentist estimator (ML) is less efficient than the estimator $\hat{\mu}_{BT}$, which is more efficient. Because they are both unbiased, this comparison is valid. Since this proposed estimator reduces the MSE in terms of bias and variance reduction, it is additionally more efficient within the Bayesian framework. By only lowering the variance, the conventional Bayesian estimator reduces the MSE.. Therefore, $\hat{\mu}_{BT}$ is a minimum variance unbiased estimator for estimating the population mean $\beta_m$. The proof established here serves as a baseline for using bootstrapped prior with the sum of trees model.

### *3.1. A New Weighted Splitting for Bayesian Random Forest in Sparse High-Dimensional Setting*

Apart from the probabilistic interpretation update on random forest regression achieved using Bayesian modelling, we also dealt with the variable selection principle used during splitting. Tree-based methods use a greedy approach to build trees. In a high-dimensional setting with a large number of covariates, modelling with all the covariates increases the computational time and thus subsampling variables randomly or using a deterministic approach is suitable for tree-based methods. There are two popular approaches for handling high-dimensional data;

1. **Greedy search:** Identifying the relevant subset of variables and fitting the desired model on them.
2. **Random search:** Randomly selecting subset (whether relevant or irrelevant) and fitting the desired model on them.

The two approaches are not 100% perfect in variable selection, greedy search fails to capture the interaction effect between variables and sometimes overfits while random search does not overfit if replicated a large number of times but tends to suffer the loss of efficiency when the variable space is populated with irrelevant variables. The RF regression algorithm randomly selects variables from the predictor space by selecting a fixed number $p/3$ irrespective of their predictive interaction with the response variable. This subsample size does not take into account the number of relevant predictors in the entire predictor space, thus the chance of selecting irrelevant features increases with increased $p$. Therefore, using the same data configuration, the predictive performance of RF reduces with increasing $p$.

The weakness of RF can be attributed to its random subset selection mechanism. Updating the subset selection with a data-driven approach such that predictors are ranked in the order of relative correlation with response $y$ will be fruitful. The motivation behind this idea follows from a greedy background, by trying to build a sum of tree models with only a relevant subset of predictors. However, this will affect the interaction modelling strength of RF which might further lead to a reduction in predictive power. In addition, we intend to update and not modify RF so as to maintain all its strength. Based on this fact, we developed a new framework that combines the strength of greedy search as well as random search by ranking the variables based on their initially computed importance.

Let $T_1, T_2, T_3, \ldots, T_p$ be $p$ independent $t$ statistics with cumulative distribution function $F(t)$. Here, $T_k$ corresponds to the $t$ statistic for each covariate $x_k$ after fitting a Bayesian simple linear regression model of the response $y$ on $x_k$. Specifically, $T_k$ can be defined as follows:

$$T_k = \frac{\hat{\theta}_k}{SD(\hat{\theta}_k)} \tag{65}$$

where $\hat{\theta}_k$ is the Bayesian estimated weight of $x_k$ in the simple linear regression model:

$$y = \theta_0 + \theta_k x_k + \epsilon \tag{66}$$

$\theta_0$ is the bias of estimating $y$ using $x_k$ and $\epsilon$ is the random noise that arises during the estimation of $y$ with the linear model; it is considered to be independent, identical, Gaussian-distributed noise with a mean of zero and a constant variance $\delta^2$. $SD(\hat{\theta}_k)$ is the posterior standard deviation of $\theta_k$. The t-statistics $T_k$ are then ranked in the increasing order of magnitude as; $T_{(1)} \leq T_{(2)} \leq T_{(3)} \leq \cdots \leq T_{(p)}$. The $T_{(k)}$ is the $kth$ order statistic $(k = 1, 2, \ldots, p)$. Then, the cumulative distribution function $(CDF)$ of the largest order statistic $T_{(p)}$ is given by;

$$F_p(t) = Pr\big(T_{(p)} \leq t\big) \tag{67}$$

$$F_p(t) = Pr\big(allT_{(k)} \leq t\big) = F^p(t) \tag{68}$$

Also, we can see that $Pr(T_{(k)} \geq allT_{(p-k)}) \equiv Pr(allT_{(p-k)} \leq T_{(k)})$; thus

$$F_{p-k}(t) = Pr\big(allT_{(p-k)} \leq T_{(k)}\big). \tag{69}$$

Equation (69) can be interpreted as the probability that at least $p - k$ of the $T_{(k)}$ are less than or equal to $t$. This also implies that all other $(p - k)$ variables are less relevant to response $y$ than $X_k$.

$$Pr(allT_{(p-k)} \leq T_{(k)}) = \sum_{k=(p-k)}^{p} \binom{p}{k} F^{(k)}(t)[1 - F(t)]^{p-k} \tag{70}$$

$$F^{(p-k)}(t) = \sum_{k=(p-k)}^{p} \binom{p}{k} F^{(k)}(t)[1 - F(t)]^{p-k}. \tag{71}$$

We now refer to $F^{(p-k)}(t)$ as weight $w_k$ which is the probability that each $x_{p-k}$ variable is less important to $y$ than $x_k$. A binary regression tree's splitting mechanism is then updated using this weight in the following ways:

$$Q^w_m(T) = (1 - w_k)\left[\sum_{i:x_k \in R_1(j,s)}^{n_{1m}} (y_i - \hat{\beta}_{1m})^2 + \sum_{i:x_k \in R_2(j,s)}^{n_{2m}} (y_i - \hat{\beta}_{2m})^2\right] \tag{72}$$

If a variable $x_k$ is important and subsequent splitting on it will have significance, the weighted deviation $Q^w_m(T)$ reduces to zero (since $w_k \to 1$). Due to the fact that variables with lower weights $w$ won't be further divided in the tree-building algorithm, this strategy helps to speed up the algorithm and improve the variable selection component of random forest regression. The procedure below summarizes BRF for a Gaussian response.

1. **Step 0:** Start with input data $D = [x; y]$
2. **Step 1:** Analyze each variable $x_k \in x$ individually by running a univariate analysis and save the bootstrap Bayesian $t$ statistic $t^{BP}$.

3. **Step 2:** Calculate the probability of maximal weight $w_k$ for each variable $x_k \in x$.
4. **Step 3:** For each of the $J$ trees, where $j = 1, 2, \ldots, J$:
5. **Step 4:** Compute the bootstrap prior predictive density weights $\omega_i$ from a Normal-Inverse $(NIG)$ distribution with parameters $\mu_M^{BP}, \sigma^2\Sigma_M^{BP}, a_{BP}, b_{BP}$.
6. **Step 5:** Generate a Bayesian weighted simple random sample $D^*$ of size $N$ with replacement from the training data $D$ using the weights $\omega_i$.
7. **Step 6:** Generate a Bayesian weighted simple random sample:
8. **Step 7:** Grow a weighted predictors CART tree $\Im_j$, by iteratively repeating the following steps for each terminal node $m$, until the minimum node size $n_{min}$ is reached:
   (a) Randomly select $mtry = \lfloor p/3 \rfloor$ variables without replacement from the $p$available variables.
   (b) Choose the best variable and split-point from the selected variables.
   (c) Divide the node into two daughter nodes.
   (d) Compute weighted splitting criterion $Q_m^w(T)$ and identify the node with the minimum deviance $Q_m^w(T)$.
9. **Step 8:** Print the ensemble of trees $\Im_j$ over $J$ iterations.
10. **Step 10:** To predict test data $x_{te}$, apply:

$$\hat{y}_{brf}^J = \frac{1}{J}\sum_{j=1}^{J} \Im_j(x_{te})$$

### 3.2. Oracle Properties of Bayesian Random Forest

In this section, we show that if the Bayesian bootstrap prior estimator $\hat{\mu}_{BT}$ is used for estimating $\beta_M$ and the weighted splitting approach is utilized, the Bayesian Random Forest (BRF) enjoys the oracle properties.

**Theorem 3.** *Suppose $\hat{\beta}_M = \hat{\mu}_{BT}$ and $F^{(p-k)}(t) \to 1$, then the Bayesian Random Forest (BRF) satisfy the following:*

*i. Identification of the right subset model $\mathcal{M}$ such that $P(\hat{\mathcal{M}} = \mathcal{M}) \to 1$.*
*ii. Achievement of the optimal estimation rate, $\sqrt{n}(\hat{\mathcal{M}} - \mathcal{M}) \xrightarrow{d} N(0, Var(\mathcal{M}))$*

**Proof.** From theorem (1), we know that the probability of selecting at least one relevant subset $R$ from the set $p$ using RF is

$$\begin{aligned} P(R_1 \cup R_2 \cup \ldots R_r) = \sum_{k=1}^{r} P(R_k) - \sum_{j,k=1;k>j}^{r} P(R_j \cap R_k) + \sum_{i,j,k=1;k>j>i}^{r} P(R_i \cap R_i \cap R_k) \\ - \cdots + (-1)^{r-1} P(R_1 \cap R_2 \cap \cdots \cap R_r). \end{aligned}$$

Now using the weighted splitting, there is assurance that the selected variable $x_k$ is relevant provided $F^{(p-k)}(t) \to 1$. This implies that the random selection of variables for splitting in BRF is a mutually exclusive process that is $P(R_i \cap R_j) = 0 \forall i \neq j$. Thus, the probability of selecting at least one relevant subset $R$ from the set $p$ using BRF is

$$\begin{aligned} P(R_1 \cup R_2 \cup \ldots R_r) &= \sum_{k=1}^{r} P(R_k) \\ &= \binom{r}{1}\left(\frac{1}{r}\right) \qquad (73) \\ &= 1 \end{aligned}$$

.

**Lemma 3.** *BRF variable selection is consistent if* $\lim_{p\to\infty} P(\hat{\mathcal{M}} = \mathcal{M}) = 1$.

**Corollary 2.** *From equation (73)* $\lim_{p\to\infty} P(\bigcup_{k=1}^{r\subset p} R_k) = 1$, *then BRF variable selection is consistent in HD with large p.*

Theorem (2) revealed that the Bayesian estimator $\hat{\mu}_{BT}$ is a uniformly minimum variance unbiased estimator for the parameter $\beta_M$ under mild regularity conditions. This implies

$$\begin{aligned} var(\hat{y}_{brf}) &= var(\hat{\mu}_{BT}) \\ var(\hat{y}_{brf}) &= \left[\frac{n_m^2 + B}{(B+n_m)^2}\right] n_m^{-1}\sigma_m^2 < var[\hat{y}_{rf}] = n_m^{-1}\sigma_m^2 \end{aligned} \tag{74}$$

**Remark 3.** *Equation (74) implies that BRF is more efficient than RF when the bootstrap size* $B \to \infty$.

□

## 4. Simulation and Results

In this section, we conducted an empirical evaluation of BRF in comparison to its major competitors using both simulation and real-life data. The analyses were performed through 10-fold cross-validations on the datasets. All the analyses were executed in the R statistical package. We utilized the newly built-in function $brf$ for BRF, $glmnet$ function [26] for LASSO, $gbm$ for Gradient Boosting [20], $rfsrc$ for Random Forest, $wbart$ for BART1 as described in [22], and $bartMachine$ for BART2 [41].

To implement the Bayesian Forest method (BF), we modified the $case.wt$ parameter of $rfsrc$ from [42], introducing random weights distributed exponentially with a rate parameter of 1. It's worth noting that we employed two different R packages for BART due to observed discrepancies in the results produced by these packages. Detailed information regarding the setup of tuning parameters can be found in Table 1.

**Table 1.** Tuning parameters set-up for the various methods used in data analysis.

| Method | Tuning Parameter Set-up |
|---|---|
| **LASSO** | $\lambda \in [0,1]$ is selected via 10 folds cross validation. Other settings are default as in $glmnet$. |
| **GBM** | Number of trees is fixed at 1000 and all other settings are default. |
| **RF** | $mtry$ settings are default $p/3$, number of trees is fixed at 1000. Other settings are default. |
| **BART1** | All settings are default. |
| **BART2** | All settings are default |
| **BF** | $mtry$ settings are default, number of trees is fixed at 1000. $case.wt \sim exp(1)$. Other settings are default. |
| **BRF** | $mtry$ settings are default $p/3$, number of trees is fixed at 1000, search type is random, split weight is obtained using $F^{(p-k)}(t)$. |

Two simulation scenarios were created based on the problem we intend to tackle in this paper. The simulation scenarios were adapted from the works of [22] and [13]. In each of the scenarios, six levels of low and high-dimensional settings were defined as $p = 50, 100, 500, 1000, 5000, 10000$ and used so as to mimic realistic gene expression datasets. The sample size corresponding to the number of patients $n$ which is usually far smaller than $p$ was fixed at 200 in all the scenarios. Here, the Root Mean Square Error $(RMSE)$ and Average Root Mean Square$(ARMSE)$ were used as performance measures over the 10-folds. Note $p = 50\&100$ were used to examine the behaviour of the methods in low dimensional data situations.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{n_{test}}}$$

$$ARMSE = \frac{\sum_{e=1}^{10} RMSE_e}{10}$$

***Scenario 1:*** Linear Case; Set $x_1, \ldots, x_p$ as multivariate standard normal $N(0,1)$ random variables with associated covariance structure define as; $\Sigma = blockdiag(\Sigma_1^1, \ldots, \Sigma_G^1) = I_G \otimes \Sigma^1$, where $\otimes$ is the kronecker product. Here we assume that the first five predictors $[x_1, \ldots, x_5]$ are relevant and the associated covariance structure is defined as

$$\Sigma^1 = \begin{cases} \rho & \text{if } i \neq j; \\ 1 & \text{if } i = j. \end{cases}$$

, such that the first five variables have pairwise correlation value $\rho = 0.9$ and likewise the other blocks of size five variables have the same correlation structure. The response is then simulated as $y = x_1 + 2x_2 + 3x_3 + 4x_4 + 5x_5 + \epsilon$, where $[x_6, \ldots, x_p]$ are the irrelevant predictor set. Note, with the covariance structure $\Sigma$ defined, the $p - 5$ variables are independent and identically distributed and $\epsilon \sim N(0,1)$.

***Scenario 2:*** Nonlinear Case; This follows the same structure as in scenario one except for the simulation of the response which is defined as $y = 10sin(x_1x_2) + 20(x_3 - 0.5)^2 + 10|x_4 - 0.5| + 5(x_5 - 0.5)^3 + \epsilon$ and $\rho = 0.2$.

### *4.1. Simulation Results*

Table 2 summarizes the 10-fold cross-validation simulation of a Gaussian response for the seven methods. As expected for scenario 1 with the linearity assumption, LASSO takes the lead followed by the new method BRF. Also, the ARMSE increases with an increase in $p$ for most of the methods except GBM which is unaffected by the increase in $p$. RF also performs much better than other ensemble methods like BF, BART1 and BART2 and most especially GBM. Although an increase in $p$ affects the performance of RF significantly, the situation is different for BRF as the increase in $p$ does not correspond to an increase in the ARMSE. BART2 performance tends to be better than BART1 for the low-dimensional case than the high-dimensional situation. BF performance is better than BART1 and BART2 but, BRF still takes the lead within the Bayesian class of models. The boxplot in Figure 5 corroborates the findings in Table 2 with the median RMSE of BRF and LASSO being the lowest over the different data dimensions.

**Table 2.** Average test Root Mean Square Error (ARMSE) over 10-fold cross-validation for scenario 1.

| Scenario 1: Linear | | | | | | |
|---|---|---|---|---|---|---|
| | $p$ | | | | | |
| **Method** | **50** | **100** | **500** | **1000** | **5000** | **10000** |
| BRF | 2.005 | 2.012 | 2.084 | 2.113 | 2.114 | 2.156 |
| GBM | 8.853 | 8.861 | 8.843 | 8.854 | 8.854 | 8.847 |
| LASSO | 0.950 | 0.949 | 0.952 | 0.958 | 0.967 | 0.977 |
| RF | 2.247 | 2.296 | 2.513 | 2.603 | 2.824 | 2.947 |
| BF | 2.568 | 2.627 | 2.855 | 3.022 | 3.185 | 3.590 |
| BART1 | 3.843 | 4.521 | 2.763 | 3.126 | 5.364 | 7.210 |
| BART2 | 2.596 | 3.113 | 3.007 | 3.621 | 5.658 | 8.395 |

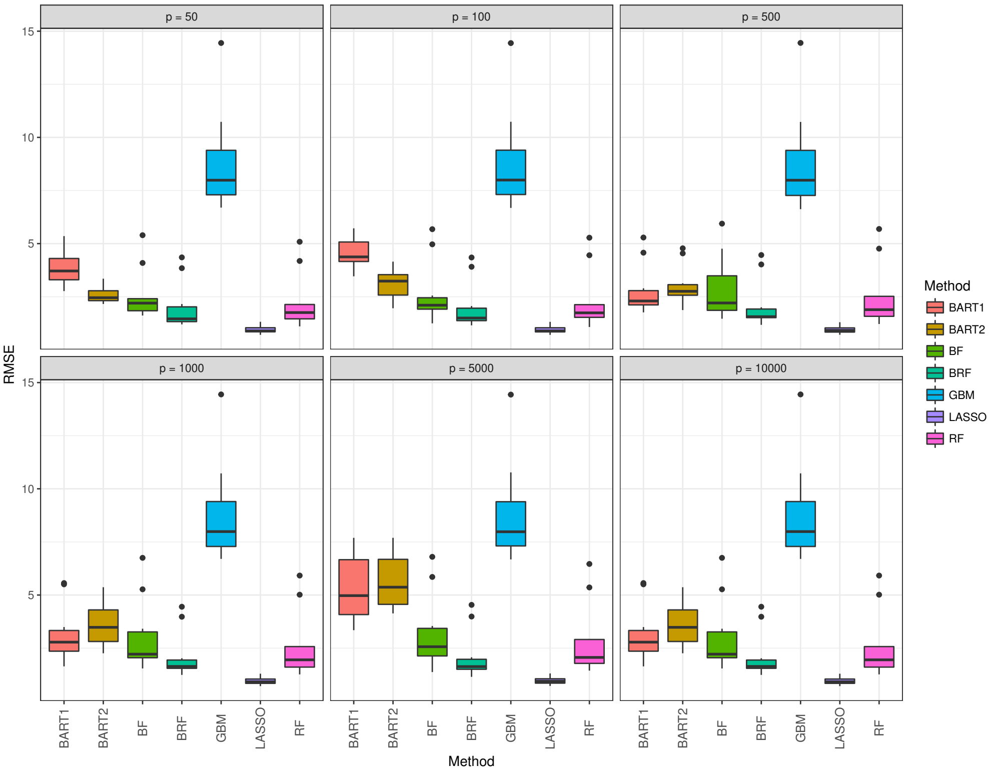

**Figure 5.** Boxplot of test 10-folds cross-validation RMSE of Scenario 1. The black middle line in each box represents the median. The dots represent outliers in RMSE results. The outliers in GBM is the highest.

The box and whisker plot for GBM (blue) was observed to be the highest in all data dimension situations. Table 3 summarizes the 10-fold cross-validation simulation for Gaussian response for the seven methods when the nonlinear model is assumed. The performance of all methods degrades drastically when compared to the linear case in Table 2. LASSO performs worse as expected in this situation. BART1 and BART2 performance are again better for the low dimensional situation when $p < n$, precisely for $p = 50, 100$. However, their performances depreciate faster as $p$ approaches 500 and in fact worse than LASSO as $p$ approaches 10000. GBM performance is again unaffected with the increase in $p$ but the performance is not different from LASSO. RF and likewise BF perform moderately better than BART1 and BART2 for $p > 1000$. BRF simultaneously achieves robustness to increase in $p$ as well as maintaining the lowest RMSE for low and high dimensional settings when compared with the six other competing methods. The boxplot in Figure 6 corroborates the findings in Table 3, with the median RMSE of BRF being the lowest for $p > 500$.

**Table 3.** Average test Root Mean Square Error (ARMSE) over 10-fold cross-validation for scenario 2.

| **Scenario 2: Nonlinear** | | | | | | |
|---|---|---|---|---|---|---|
| | $p$ | | | | | |
| **Method** | **50** | **100** | **500** | **1000** | **5000** | **10000** |
| BRF | 19.498 | 19.896 | 19.889 | 19.906 | 21.637 | 22.543 |
| GBM | 30.355 | 30.352 | 30.444 | 30.551 | 30.708 | 30.964 |
| LASSO | 32.404 | 32.453 | 32.585 | 33.526 | 34.626 | 35.230 |
| RF | 20.664 | 21.264 | 23.266 | 23.742 | 25.043 | 25.954 |
| BF | 23.288 | 22.993 | 26.062 | 26.095 | 29.121 | 28.210 |
| BART1 | 16.844 | 16.493 | 24.151 | 28.180 | 31.917 | 38.156 |
| BART2 | 14.037 | 18.071 | 22.193 | 27.450 | 33.522 | 37.932 |

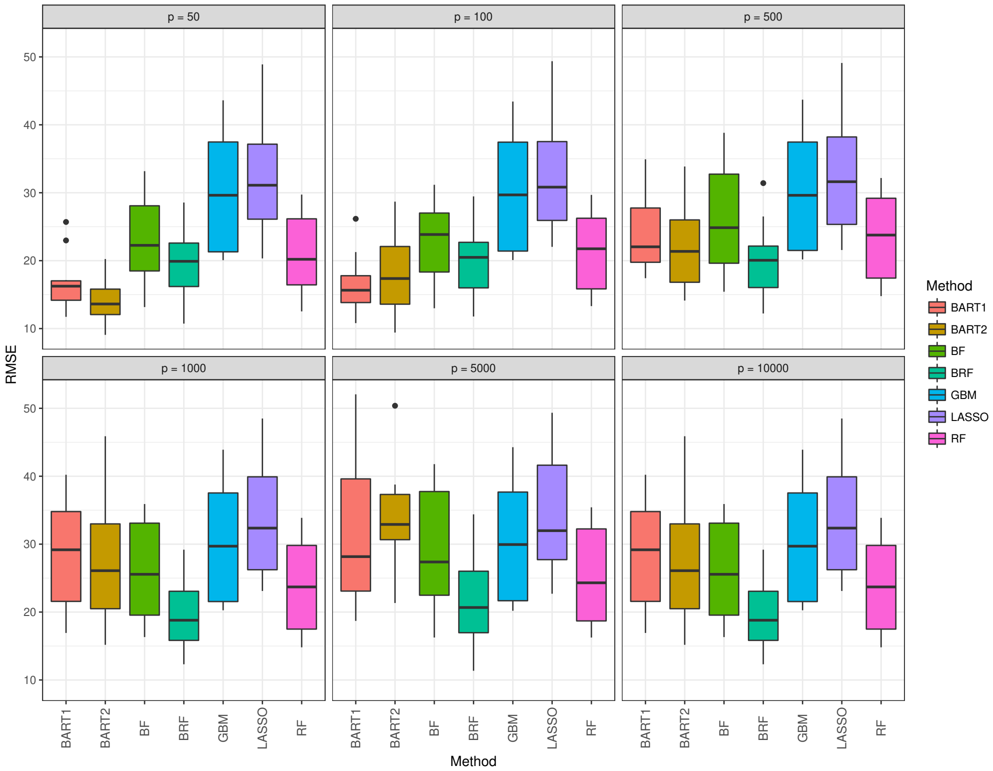

**Figure 6.** Boxplot of test 10 folds cross-validation RMSE of Scenario 2 for Gaussian response. The black middle line in each box represents the median.

### *4.2. Variable Selection*

The two scenario models (Linear and Non-linear) were investigated to determine the best method in terms of the selection of the five relevant variables imposed. Table 4 presents the results of the variable selection performance of BRF alongside competing methods. For the linear model, the average proportion of relevant variables identified using LASSO is 1 and constant over all the six datasets used. This result corroborates the findings in Table 2 where LASSO was found to be the best in terms of lowest ARMSE. The entire five relevant variables were correctly identified with LASSO under the linearity assumption. BRF competes favourably with LASSO with the identification of about 4/5 relevant variables up to $p = 5000$. BRF also consistently identified all the relevant variables in low dimensional conditions with $p = 50\&100$. The performances of RF, BF and GBM are very similar with GBM slightly above RF and BF. BART2 also consistently identified about 4/5 relevant variables up till $p = 1000$. However, the performance at $p = 5000\&10000$ is not presented due to computational difficulty while computing the probability of inclusion for $p > 1000$. The lowest performance was observed with BART1 over all the dimensions of datasets used.

For the non-linear condition, none of the methods could achieve 100% identification as the functional path is now rough but BRF is still the best for $p = 50$, and it converges to $2/5$ from $p = 1000$. LASSO performance is not consistent here and it also corroborates the high ARMSE observed in Table 3. BART2 competes with BRF at various levels of $p$ and in fact the highest for $p \leq 1000$. A similar worse performance was observed for BART1 under the non-linear condition.

**Table 4.** Average proportion of relevant variables selected in 10 folds cross validation.

|  | $p$ |  |  |  |  |  |
|---|---|---|---|---|---|---|
| **Method** | **50** | **100** | **500** | **1000** | **5000** | **10000** |
| **Linear** |  |  |  |  |  |  |
| BRF | 1.00 | 0.98 | 0.90 | 0.84 | 0.76 | 0.68 |
| RF | 0.96 | 0.84 | 0.80 | 0.78 | 0.68 | 0.64 |
| BF | 0.98 | 0.84 | 0.76 | 0.68 | 0.64 | 0.64 |
| GBM | 0.96 | 0.84 | 0.82 | 0.80 | 0.76 | 0.76 |
| LASSO | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| BART1 | 0.76 | 0.86 | 0.82 | 0.80 | 0.48 | 0.24 |
| BART2 | 0.88 | 0.88 | 0.84 | 0.80 | - | - |
| **Non-linear** |  |  |  |  |  |  |
| BRF | 0.66 | 0.58 | 0.44 | 0.40 | 0.40 | 0.40 |
| RF | 0.60 | 0.60 | 0.44 | 0.42 | 0.40 | 0.38 |
| BF | 0.58 | 0.58 | 0.40 | 0.36 | 0.36 | 0.26 |
| GBM | 0.58 | 0.56 | 0.40 | 0.40 | 0.38 | 0.34 |
| LASSO | 0.62 | 0.72 | 0.42 | 0.40 | 0.40 | 0.40 |
| BART1 | 0.54 | 0.56 | 0.40 | 0.30 | 0.14 | 0.10 |
| BART2 | 0.62 | 0.64 | 0.52 | 0.44 | - | - |

### *4.3. Predicting Tumour Size and Biomarker Score*

Three real-life cancer datasets on the prediction of tumour size and biomarker score. The two breast cancer datasets were used to predict the size of tumour before the patients underwent chemotherapy. The other dataset was used to predict the biomarker score of lung cancer for patients with smoking history. The dataset's detailed description can be found below:

1. ***Breast1 Cancer***: [43] obtained 22,283 gene expression profiles using Affymetrix Human Genome U133A Array on 61 patients prior to chemotherapy. The pre-chemotherapy size of tumours was recorded for both negative Estrogen Receptor (ER-) and positive Estrogen Receptor (ER+). A preliminary analysis carried out on the dataset using a Bayesian t-test revealed that only 7903 genes are relevant at some specific threshold.
2. ***Breast2 Cancer:*** [44] obtained 22,575 gene expression profiles using 60mer oligonucleotide array from 60 patients with ER-positive primary breast cancer and treated with tamoxifen monotherapy for 5 years. Data were generated from whole tissue sections of breast cancers. The pre-chemotherapy size of tumours was recorded for both negative Estrogen Receptor (ER-) and positive Estrogen Receptor (ER+). A preliminary analysis carried out on the dataset using a Bayesian t-test revealed that only 4808 genes are relevant at some specific threshold.
3. ***Lung Cancer***: [45] obtained 22,215 gene expression profiles using Affymetrix Suggested Protocol on 163 patients. The biomarker score to detect the presence or absence of lung cancer was recorded alongside the gene expression profile. A preliminary analysis carried out on the dataset using a Bayesian t-test revealed that only 7187 genes are relevant at some specific threshold.

The RMSE of the methods were obtained for the test dataset that arose from ten-fold cross-validation. Table 5 shows the summary of ARMSE for the test dataset over the ten-fold cross-validation. For Breast1 and Breast2, BRF was found to be the best with the lowest ARMSE. In terms of ranking, RF

doi:10.20944/preprints202310.1272.v1

was found to be in the second position in terms of performance when compared with other methods. For the prediction of biomarker score, the best is LASSO with the lowest ARMSE. On average, BRF has the lowest ARMSE over the three datasets. The Standard Error of Mean (SEM) estimates measure the relative spread of RMSE for each dataset. The SEM results show that the most stable method is BRF with least SEM over most datasets except Lung.

**Table 5.** Average test RMSE and (Standard error) over 10-fold cross-validation for regression cancer datasets.

| | | | | Method | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | **BRF** | **GBM** | **LASSO** | **RF** | **BF** | **BART1** | **BART2** |
| Breast1 | 1.014 | 1.131 | 1.283 | 1.117 | 1.120 | 1.128 | 1.123 |
| | (0.071) | (1.086) | (0.258) | (0.673) | (0.749) | (0.661) | (1.380) |
| Breast2 | 0.347 | 0.450 | 0.458 | 0.448 | 0.449 | 0.456 | 0.452 |
| | (0.048) | (0.352) | (0.246) | (0.298) | (0.576) | (0.139) | (0.239) |
| Lung | 1.099 | 5.243 | 0.825 | 2.287 | 2.420 | 1.589 | 1.934 |
| | (0.243) | (1.125) | (0.309) | (0.162) | (0.298) | (0.717) | (0.255) |

## 5. Discussion of Results

BRF achieves impressive results because it employs Bayesian estimation at the tree node parameter stage and combines a greedy and random search to select splitting variables. In contrast, RF fails since it randomly selects variables without considering their importance. Random search is adequate for low-dimensional cases, as seen in various simulation conditions. However, as the number of irrelevant variables increases, the performance of random search significantly deteriorates. For example, in a five-dimensional simulation with five relevant variables, the probabilities of selecting at least one relevant variable when $mtry = \lfloor \sqrt{p} \rfloor$ are as follows $0.546, 0.416, 0.202, 0.150, 0.06, 0.05, 0.04$ for different values of $p = 50, 100, 500, 1000, 5000, 10000$. This demonstrates that as the data dimension grows with a fixed sample size $n$, more irrelevant variables are selected, resulting in a poor model fit.

The new approach, BRF, directly addresses this issue by ensuring the use of only relevant variables, regardless of the dataset's dimension. This approach is akin to what GBM does, as it assesses the influence of each variable on the response. However, BRF surpasses GBM due to its application of Bayesian estimation methods and robust data-driven prior techniques. Moreover, it's clear that BRF's performance relies on correctly identifying variables during the greedy search. If irrelevant variables are ranked higher than relevant ones, it will affect performance, emphasizing the need for a robust procedure for preliminary variable ranking. While the bootstrap prior technique performed reasonably well in both linear and non-linear scenarios, the accuracy of BRF can also be improved by introducing a more effective subset selection procedure.

## 6. Conclusion

This paper investigated the strengths and flaws of Random Forest (RF) for modelling high-dimensional data. The major weakness of RF methods is that they are not governed by any statistical model, and thus, they cannot provide probabilistic results as in the Bayesian setting. Another critical issue with the RF methods occurs in high-dimensional data with a large number of predictors but a small number of relevant ones. The performance of RF tends to depreciate as the dimension of the data grows infinitely under this condition. These two issues motivated the development of Bayesian Random Forests (BRF) presented in this paper. The theoretical results revealed that BRF satisfies the oracle properties under mild regularity conditions. Furthermore, the various empirical results from the simulation and real-life data analysis established that BRF is more consistent and efficient than other competing methods for modelling non-linear functional relationships in low and high-dimensional situations. Also, BRF was found to be better than the competing Bayesian methods, especially in high-dimensional settings.

**Author Contributions:** "Conceptualization, O.R.O. and A.R.A.; methodology, O.R.O; software, O.R.O.; validation, O.R.O, and A.R.A; formal analysis, O.R.O; investigation, O.R.O; resources, A.R.A; data curation, O.R.O; writing—original draft preparation, O.R.O; writing—review and editing, O.R.O; visualization, A.R.A; supervision, O.R.O; project administration, O.R.O. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The authors confirm that the data supporting the findings of this study are available within the article.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gohil, S.H.; Iorgulescu, J.B.; Braun, D.A.; Keskin, D.B.; Livak, K.J. Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nature Reviews Clinical Oncology* **2021**, *18*, 244–256.
2. Quist, J.; Taylor, L.; Staaf, J.; Grigoriadis, A. Random forest modelling of high-dimensional mixed-type data for breast cancer classification. *Cancers* **2021**, *13*, 991.
3. Nederlof, I.; Horlings, H.M.; Curtis, C.; Kok, M. A high-dimensional window into the micro-environment of triple negative breast cancer. *Cancers* **2021**, *13*, 316.
4. Olaniran, O.R.; Olaniran, S.F.; Popoola, J.; Omekam, I.V. Bayesian Additive Regression Trees for Predicting Colon Cancer: Methodological Study (Validity Study). *Turkiye Klinikleri J Biostat.* **2022**, *14*, 103–109.
5. Olaniran, O.R.; Abdullah, M.A.A. Bayesian weighted random forest for classification of high-dimensional genomics data. *Kuwait Journal of Science* **2023**, *50*, 477–484.
6. Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning*; Springer, New York, 2009.
7. Olaniran, O.R. Shrinkage based variable selection techniques for the sparse Gaussian regression model: A Monte-Carlo simulation comparative study. In Proceedings of the AIP Conference Proceedings. AIP Publishing LLC, 2021, Vol. 2423, p. 070014.
8. Bühlmann, P.; Van De Geer, S. *Statistics for high-dimensional data: methods, theory and applications*; Springer Science & Business Media, 2011.
9. Gündüz, N.; Fokoue, E. Predictive performances of implicitly and explicitly robust classifiers on high dimensional data. *Communications Faculty of Sciences University of Ankara-Series A1 Mathematics And Statistics* **2017**, *66*, 14–36.
10. Vapnik, V. *The nature of statistical learning theory*; Springer science & business media, 2013.
11. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **1996**, *58*, 267–288.
12. Fan, J.; Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **2001**, *96*, 1348–1360.
13. Hernández, B.; Raftery, A.E.; Pennington, S.R.; Parnell, A.C. Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing* **2018**, *28*, 869–890.
14. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and regression trees*; 1984.
15. Hwang, K.; Lee, K.; Park, S. Variable selection methods for multi-class classification using signomial function. *Journal of the Operational Research Society* **2017**, *68*, 1117–1130.
16. Breiman, L. Bagging predictors. *Machine learning* **1996**, *24*, 123–140.
17. Efron, B.; Tibshirani, R.J. *An introduction to the bootstrap*; CRC press: Florida, 1994.
18. Breiman, L.; et al. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics* **1998**, *26*, 801–849.
19. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.
20. Friedman, J.H. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* **2001**, *29*, 1189–1232.
21. Hastie, T.; Friedman, J.; Tibshirani, R. *Overview of supervised learning*; Springer: New York City, 2010.
22. Chipman, H.A.; George, E.I.; McCulloch, R.E.; et al. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **2010**, *4*, 266–298.
23. Linero, A.R. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association* **2018**, pp. 1–11.
24. Breiman, L. Stacked regressions. *Machine learning* **1996**, *24*, 49–64.

25. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* **1997**, *55*, 119–139.
26. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **2010**, *33*, 1.
27. Olaniran, O.R.; Olaniran, S.F.; Popoola, J. Bayesian Regularized Neural Network for Forecasting Naira-USD Exchange Rate. In Proceedings of the International Conference on Soft Computing and Data Mining. Springer, 2022, pp. 213–222.
28. Chipman, H.A.; George, E.I.; McCulloch, R.E. Bayesian CART model search. *Journal of the American Statistical Association* **1998**, *93*, 935–948.
29. Taddy, M.A.; Gramacy, R.B.; Polson, N.G. Dynamic trees for learning and design. *Journal of the American Statistical Association* **2011**, *106*, 109–123.
30. Olaniran, O.R.; Abdullah, M.A.A.B.; Affendi, M.A. BayesRandomForest: An R implementation of Bayesian Random Forest for Regression Analysis of High-dimensional Data. *Romanian Statistical Review* **2018**, *66*, 95–102.
31. Johnson, N.L.; Kemp, A.W.; Kotz, S. *Univariate discrete distributions*; John Wiley & Sons: New York City, 2005.
32. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association* **2006**, *101*, 1418–1429.
33. Shi, G.; Lim, C.Y.; Maiti, T. High-dimensional Bayesian Variable Selection Methods: A Comparison Study. *Calcutta Statistical Association Bulletin* **2016**, *68*, 16–32.
34. Heinze, G.; Wallisch, C.; Dunkler, D. Variable selection–A review and recommendations for the practicing statistician. *Biometrical Journal* **2018**, *60*, 431–449.
35. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian data analysis*; CRC press: Florida, 2013.
36. Denison, D.G.; Holmes, C.C.; Mallick, B.K.; Smith, A.F. *Bayesian methods for nonlinear classification and regression*; John Wiley & Sons, 2002.
37. Olaniran, O.R.; Yahya, W.B. Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique. *Journal of Modern Applied Statistical Methods* **2017**, *16*, 34.
38. Olaniran, O.R.; Abdullah, M.A.A. Bayesian variable selection for multiclass classification using Bootstrap Prior Technique. *Austrian Journal of Statistics* **2019**, *48*, 63–72.
39. Olaniran, O.R.; Abdullah, M.A.A. Bayesian analysis of extended cox model with time-varying covariates using bootstrap prior. *Journal of Modern Applied Statistical Methods* **2020**, *18*, 7.
40. Laird, N.M.; Louis, T.A. Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **1987**, *82*, 739–750.
41. Bleich, J.; Kapelner, A.; George, E.I.; Jensen, S.T. Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics* **2014**, pp. 1750–1781.
42. Ishwaran, H.; Kogalur, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *The Annals of Applied Statistics* **2008**, pp. 841–860.
43. Iwamoto, T.; Bianchini, G.; Booser, D.; Qi, Y.; Coutant, C.; Ya-Hui Shiang, C.; et al. Gene pathways associated with prognosis and chemotherapy sensitivity in molecular subtypes of breast cancer. *Journal of the National Cancer Institute* **2010**, *103*, 264–272.
44. Ma, X.J.; Wang, Z.; Ryan, P.D.; Isakoff, S.J.; Barmettler, A.; Fuller, A.; Muir, B.; et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* **2004**, *5*, 607–616.
45. Gustafson, A.M.; Soldi, R.; Anderlind, C.; Scholand, M.B.; Qian, J.; Zhang, X.; Cooper, K.; et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Science Translational Medicine* **2010**, *2*, 1–25.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.