*International Journal of Engineering & Technology*, 7 (4.30) (2018) 170-175

**International Journal of Engineering & Technology**

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***

# Empirical Bayesian Binary Classification Forests Using Bootstrap Prior

**Oyebayo Ridwan Olaniran[12]*, Mohd Asrul Affendi Bin Abdullah[2], Khuneswari A/P Gopal Pillay[2], Saidat Fehintola Olaniran[3]**

*[1] Department of Statistics, University of Ilorin, Ilorin, PMB 1515, Nigeria*
*[2] Department of Mathematics and Statistics, Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Pagoh Educational Hub, 84600 Pagoh, Malaysia*
*[3] Department of Statistics and Mathematical Sciences, College of Pure and Applied Sciences, Kwara State University, Malete, PMB 1530, Ilorin, Nigeria*
**Corresponding author E-mail:rid4stat@yahoo.com*

**Abstract**

In this paper, we present a new method called Empirical Bayesian Random Forest (EBRF) for binary classification problem. The prior ingredient for the method was obtained using the bootstrap prior technique. EBRF addresses explicitly low accuracy problem in Random Forest (RF) classifier when the number of relevant input variables is relatively lower compared to the total number of input variables. The improvement was achieved by replacing the arbitrary subsample variable size with empirical Bayesian estimate. An illustration of the proposed, and existing methods was performed using five high-dimensional microarray datasets that emanated from colon, breast, lymphoma and Central Nervous System (CNS) cancer tumours. Results from the data analysis revealed that EBRF provides reasonably higher accuracy, sensitivity, specificity and Area Under Receiver Operating Characteristics Curve (AUC) than RF in most of the datasets used.

***Keywords***: *Binary Classification; Empirical Bayes; High-Dimensional, Random Forest.*

## 1. Introduction

Recent advancement in technology has made collection of big datasets referred to as high-dimensional data in statistical parlance possible [1]. High-dimensional data popularly addressed as *"large $p$ small $n$"* syndrome often arises in most areas of research especially in genomic studies [2-3]. Several techniques for handling high-dimensional data have been proposed in different areas of research. The methodologies of the methods differ from each other, but the universal standpoint of the methods is to find a way to analyze high-dimensional data better. [4] identified the needs for developing robust methods for high-dimensional data. Classical methods like ordinary least squares, logistic regression etc. often breaks down due to ill-conditioned design matrix when $p \gg n$. [2] described two major approaches for analysing high-dimensional data namely: modification of $n > p$ procedures to accommodate high-dimensional data or developing a new strategy. Modification of procedures involves moving from complex models to simple model by selecting relevant subsets of the $p$ variables.

Single classifiers such as Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), or Naïve Bayes have been used for handling high-dimensional data [5-7]. However, in the recent time, ensembles algorithms have been shown to be a better alternative to single classifier especially when multi-modals variable are grouped [3]. [3] among others claimed that Random Forest (RF) produced the highest accuracies in most scientific applications.

Random Forest (RF) developed by [8] is an ensemble statistical learning method designed to improve the predictive accuracy of decision trees. It is one of the most popular ensemble algorithms which have been applied to different fields. It is widely applicable because of its distribution free assumption, modelling of non-linear effects, computational speed and direct applicability to high-dimensional datasets [9]. It also gives room for model interpretation via variable importance measures which make it better than other black box models like Support Vector Machine (SVM) and Artificial Neural Network (ANN) [2]. Random forests algorithm involves selecting subsets of training datasets as well as subsets of variable space to build classification tree (CART, [10]).

Among the above strengths of RF lies the weakness which is the determination of relevant input variables to be used at each splitting step [11-12]. Random subset selection of variables leads to hypergeometric probability model. The hypergeometric probability reduces with decrease in the number of relevant variables in the predictor space. This reduction further results to decrease in RF accuracy.

There have been many improvements on RF in the past years, especially in the area of random subset of input variables. [13] made one of the earliest improvement on RF after the original paper in 2001. He questioned the use of Gini Index (GI) criterion proposed by [10] for selecting the variables at the splitting stage. He replaced GI with Gain Ratio (GR). The improvement only worked in some specific situations like low dimensional data.

Following the drawback observed from the work of [13], [14] proposed Meta Random Forest. The brain behind their algorithm is to use random forest themselves as base classifiers for making ensembles. Meta random forests are developed using bagging and

**Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.**

boosting approaches. The performances of those two new models were tested and compared with the original random forest algorithm. Among the observed models, bagged random forest produced the best results. [15] improved RF by using a combination of an attribute evaluator method and an instance filter method. Their approach involves preliminary variable selection before applying RF. They believe that by selecting the subset variables before applying RF will increase the chance of choosing relevant informative variables. The framework of their approach could be regarded to fall under the filter approach which is prone to false negative or positive issue [16]. False positive is not as crucial as a false negative in this scenario as second stage subset selection is embedded in RF. However, false negative is an important issue, that's a case where the relevant variable(s) have been dropped in the preliminary feature selection stage. Although, [15] used three different filter methods namely; Correlation-based feature selection (CFS), Symmetrical Uncertainty (SU) and GR. Their performance analysis results showed equal strength for the three purposes. Also, their approach violates the fundamental principle of RF as subset selection is already embedded in it.

Apart from the non-probabilistic modification of RF, Bayesian approaches have also been proposed. One of the first Bayesian approaches to RF is Bayesian Additive Regression Trees (BART) [17-18]. Bayesian methods are the emerging solution to most real-life problems because they model uncertainty in parameters [1]. As RF follows from ensembles of the CART, BART is an ensemble of Bayesian CART [19]. BART is similar in spirit to boosting but motivated by RF. BART provides appealing results with low dimensional data but fails in handling high-dimensional data. [1] illustrated the computational inefficiency of BART, because a full Bayesian probability modelling scheme is used for building each tree. Also, BART is more of modification of boosting than RF as trees priors are specified such that trees with low information about the classification of each class are boosted.

The full modelling of decision trees structure captured by BART gives room for slow computation. [1] addressed the issue by providing a full probabilistic model for the sum of trees rather than an individual tree. Their approach is motivated by Bayesian model averaging and thus referred to Bayesian Additive Regression Trees using Bayesian Model Averaging (BART-BMA). Their method is faster than BART but the accuracy observed in a drug discovery example is not different from RF and BART. This implies BART-BMA only improves the algorithmic time as well as interpretability.

A simplified approach of Bayesian random forest called Bayesian Forest (BF) was proposed by [20]. Their approach focuses on modification of training samples selection rather than input variable selection. The traditional RF uses bootstrapping procedure by [21], while BF uses the Bayesian bootstrap of [22]. They showed that BF is not better than RF except in improving the interpretability of RF. They further extended BF to Empirical Bayesian Forest (EBF). EBF was motivated by building hierarchical modelling stages of RF. Empirical Bayes is a well-known framework for fast approximate Bayesian inference [23-24]. They only use EBF to produce an approximate estimate for BF in situations where full BF could not be achieved. They conclude that EBF is not better than BF and both are equally not better than RF in most applications reviewed.

The various Bayesian modifications, as well as non-probabilistic approaches (frequentist), fail to handle RF flaws especially in the area of high-dimensional data. Therefore, in this paper, we present an improved random forest classifier for binary class data with an update on the splitting stage and samples selection. Specifically, we replaced the hypergeometric probability weights with an Empirical Bayes weight. The Bayes weight is motivated by the posterior density of hypergeometric probability of selecting any relevant variable. The Bayesian inference for the approach was driven by hybridizing bootstrap prior technique of [24] with empirical Bayes approach.

## 2. Random Forest (RF)

Given a training dataset $D = [y_i, x_{i1,}\ x_{i2,}\ldots,\ x_{ip},\ i = 1,2,\ldots,n]$, where $y_i$ is a binary outcome that assumes $k = 0,1$ values and $x_i$ is the vector of variables. Random Forests algorithm automatically decides on the splitting variables and splitting point by partitioning the response into $R_1, R_2, \ldots, R_M$ regions, the closest form of model that RF assumes is;

$$y = \sum_{m=1}^{M} \beta_m I(x \in R_m) \tag{1}$$

where $\beta_m$ is a constant in region $m$. Estimating $\beta_m$ requires the computation of an impurity function. For classification case, the commonly used impurity functions are Misclassification Error Rate (MER), Gini Index, and deviance [8]. Random Forests (RF) then update built trees (1) in two steps; (i) bootstrapping the training dataset $B$ times to obtain a total of $B$ trees (ii) Subsampling $q < p$ variables without replacement at each split step in each $b$ tree. Thus if we denote (1) by $\Im(\hat{\beta}_m : x \in R_m)$, RF model is;

$$\hat{y} = \sum_{b=1}^{B} \Im_b(\hat{\beta}_m : x \in R_m) \tag{2}$$

RF has two tuning parameters, the number of trees $B$ and number of subsampled variables $q$. [8] suggested using at least $B = 200$ and $q = \sqrt{p}$.

## 3. Empirical Bayesian Random Forest (EBRF)

[8] established that RF is highly sensitive to $q$. He suggested using cross-validation to choose $q$ but at the expense of computation time. Also, arbitrarily increasing $q$ increases the adjacent trees correlation thereby reducing the accuracy of RF. Likewise, reducing $q$ increases accuracy but introduces bias. In the face of this dilemma, we introduce here a data dependent Bayesian approach called Empirical Bayes [24-27; 6] for estimating $q$. The likelihood of selecting randomly any $r$ relevant variables out of a total random subset $q$ is given by;

$$P(r|R,q,p) = \frac{\binom{R}{r}\binom{p-R}{q-r}}{\binom{p}{q}}, \max(0, q - \mathrm{p} + \mathrm{R}) \le \mathrm{l} \le \min(q, \mathrm{R}) \tag{3}$$

where $R$ is the number of relevant variables and also the parameter of interest, $r$ is the sample realization of $R$. [28] defined a discrete $ABC(N,a,b)$ conjugate distribution as a particular case of Polya or beta-binomial distribution [29]. Thus, for a hypergeometric likelihood with $R$ target outcomes the $ABC(N,a,b)$ conjugate distribution of $R - r$ is given by;

$$P(R|N,a,b) = \frac{\binom{a+R}{a}\binom{b+N-R}{b}}{\binom{a+b+N+1}{a+b+1}}, R = 0,1,.\ .\ .,N$$

Therefore, the posterior distribution of $R$ is thus;

$$P(R|\,N,a,b;r,q,p) = \frac{\frac{\binom{R}{r}\binom{p-R}{q-r}}{\binom{p}{q}} \times \left[\frac{\binom{a+R}{a}\binom{b+N-R}{b}}{\binom{a+b+N+1}{a+b+1}}\right]}{\sum_{R=r}^{N-q+r} \frac{\binom{R}{r}\binom{p-R}{q-r}}{\binom{p}{q}} \times \left[\frac{\binom{a+R}{a}\binom{b+N-R}{b}}{\binom{a+b+N+1}{a+b+1}}\right]}$$

$\rightarrow$

$$P(R|\,N,a,b;r,q,p) = \frac{\binom{a+R}{a+r}\binom{b+N-R}{b+q-r}}{\binom{a+b+N+1}{a+b+1}}, r \le R \le N - q + r \tag{4}$$

Consequently, the posterior for $R - r$,

$$P(R-r|\,N,a,b;r,q,p)=\frac{\binom{a+r+R-r}{a+r}\binom{b+q-r+N-q-R+r}{b+q-r}}{\binom{a+r+b+n-r+N-n+1}{a+r+b+n-r+1}},0\le R-r\le N-q \quad (5)$$

The Bayesian estimate $\hat{R}_\Theta$ of $R$ is given by the posterior mean of (4);

$$\hat{R}_\Theta=\frac{q(a+R)}{a+b+N+1} \quad (6)$$

where $\hat{R}_\Theta$ is the posterior estimate of relevant variables. Moving from (6), the empirical Bayesian approach here implies $q=p$, and redefine (6) as;

$$\hat{R}_\Theta=\frac{p(a+R)}{a+b+N+1} \quad (7)$$

From (7) we then obtain $\hat{q}_\Theta=\hat{R}_\Theta$, so that $\hat{q}_\Theta$ contains relevant variables. To complete the prior specification of parameters, $R$ prior parameter of the relevant variable is obtained by fitting a hypergeometric distribution to the data then estimate $R$. The parameters $(N,a,b)$ were fixed at $\left(p,\frac{p}{2},\frac{p}{2}\right)$. The prior specification fixed the initial probability of relevant variable as half of the entire variable space. The posterior relevant probability is then denoted as $\delta$. The remaining steps of RF then follow with $q=\hat{q}_\Theta$. After selecting $\hat{q}_\Theta$ variables, the impurity functions can then be obtained using Gini index $\vartheta$;

$$\vartheta=(1-\delta)\sum_{k\in[0,1]}\hat{p}_{mk}(1-\hat{p}_{mk})$$

where $\hat{p}_{mk}$ is the estimated $k$ class probability at each node $m$. The variable with weight $\delta\to 1$, will correspond to variable with minimal unweighted Gini index and therefore useful for further splitting step. If on the other hand $\delta\to 0$, implies the variable is not useful and consequently expected to yield a maximal unweighted Gini index. In this case, the proposed weighted Gini index $\vartheta$ returns the unweighted Gini index so that the variable is dropped at the splitting stage. The idea behind this is to control the mixture behaviour of hypergeometric distribution [30]. The dominant category determines the estimates of categories probability. RF fails to balance this gap by specifying $l=\sqrt{p}$, for example, if $p=2000$; $l\approx 45$, which implies taking a random sample of 45 variables to be used in each split. The hypergeometric probability of selecting any relevant variable out of say five relevant feature is approximately 0.11. This implies that at each splitting step, there is about 89% chance of selecting irrelevant feature. This high probability can be attributed to fewer number of relevant variables. Thus RF assumes that the entire predictor or input space $p$ is reasonably populated with relevant features. Also, one might think that increasing the subsample size $l$ will increase the chance of selecting relevant variable. It is indeed true, but it will increase the correlation between adjacent trees which is the primary objective of developing RF. This is the dilemma at the forefront of RF which we tackled in this research.

## 4. Datasets

Five microarray cancer datasets were used to compare the performance of RF and EBRF. The datasets covered colon cancer, breast cancer, Lymphoma cancer and CNS cancer.

***1.) Colon cancer data:*** [31] first analyzed the data to identify biomarkers for colon cancer in 62 subjects based on 2000 genes expression profiles. Two distinct groups are identified; 40 tumorous samples and 22 normal samples.

***2.) CNS data:*** The Central Nervous System Embryonal Tumour were analyzed by [32] to identify biomarkers for CNS tumour in 34 subjects based on 7128 genes expression profiles. Two distinct groups are identified; 25 classic (C) samples and 9 desmoplastic (D) samples.

***3.) Breast Cancer data:*** [33] first analyzed the data to identify biomarkers for breast cancer in 49 subjects based on 7129 genes expression profiles. Two distinct groups are identified; 25 negative Estrogen Receptor ($ER^-$) and 24 positive Estrogen Receptor ($ER^+$).

***4.) Breast Cancer data:*** [34] analyzed the data to identify biomarkers for breast cancer in 168 patients based on 2905 genes expression profiles. Two distinct groups are identified; Good: - 111 patients with no event after five years of diagnosis and Poor: - 57 patients with early metastasis.

***5.) Lymphoma Cancer data:*** [35] analyzed the data to identify biomarkers for lymphoma cancer in 77 subjects based on 6817 genes expression profiles. Two distinct groups are identified; 58 Diffuse Large B-cell Lymphoma (DLBCL) and 19 Follicular Lymphoma (FL).

## 5. Performance Comparison

The performance criteria used to compare the two methods are sensitivity, specificity, accuracy, balance accuracy and Area under Receiver operating characteristics curve (AUC). The metrics were computed using the confusion matrix shown in Table 1.

**Table 1:** Confusion matrix

| True Class | Predicted Class | | Total |
|---|---|---|---|
| | 0 | 1 | |
| 0 | TN | FP | N |
| 1 | FN | TP | P |
| Total | N* | P* | T |

*0: Normal, 1: Tumour*

where TN represents True Nesgative, FP is the False Positive, FN represents False Negative, and TP is the True Positive. Also, N* is the total predicted negative and P* represents total predicted positive. Similarly, N is the total actual negative while P is the total actual positive. T represents the total number of observation equivalent to;

$$T=TN+FP+TP+FN$$

Here negative means normal cells while positive means tumour cells. The class specific and overall classification metrics used can be defined as follows [7, 36]

**Accuracy (ACC)**: $\%ACC=100\times\left(\frac{TN+TP}{T}\right)$

**Sensitivity**:$\%Sensitivity=100\times\left(\frac{TP}{P}\right)$

**Specificity**:$\%Specificity=100\times\left(\frac{TN}{N}\right)$

**Balance Accuracy (BACC)**: $\%BACC=\frac{\%Sensitivity+\%Specificity}{2}$

The Area under the Receiver Operating Characteristics Curve (AUC) can be computed based on Mann-Whitney U statistic. If we let $U$ be the Mann-Whitney U statistic and $n_0,n_1$ is the number of 0: (Normal cells) and 1: (Tumour cells). Thus the AUC can be defined as;

$\%AUC=100\times\left(\frac{U}{n_0n_1}\right)$.

## 6. Results and Discussion

In this section, we illustrate the application of Empirical Bayesian Random Forest (EBRF) on five published real datasets. Table 1 presents the data set which is a subset of 22 datasets from package "datamicroarray" in R statistical package [37]. The performance metrics were computed from 10-folds train/test cross-validation. The metrics used are accuracy, balance accuracy, sensitivity, specificity, and area under receiver operating characteristics curve (AUC) [36]. For each of the five datasets, 10 independent train/test splits were generated by randomly selecting $9/10$ of the data as a training set and the remaining $1/10$ as a test set. Thus, $10 \times 10 = 100$ test/train splits were created. Based on each training set, each method was then used to predict the corresponding test set and evaluated by its predictive performance. A similar approach was used in [5; 37-40].

**Table 2**: The five datasets used in the analysis

| Cancer type | Author | $n$ | $p$ |
|---|---|---|---|
| Colon Cancer | Alon, (1999) | 62 | 2000 |
| Breast Cancer | Gravier, (2010) | 168 | 2905 |
| Breast Cancer | West, (2001) | 49 | 7129 |
| Lymphoma Cancer | Shipp, (2002) | 77 | 6817 |
| CNS Cancer | Pomeroy, (2002) | 60 | 7128 |

**Table 3**: Subsample variable size results $\hat{R}_\Theta$ using EBRF and $\sqrt{p}$ for RF

| Cancer type | Author | EBRF | RF |
|---|---|---|---|
| Colon Cancer | Alon, (1999) | 155 | 44 |
| Breast Cancer | Gravier, (2010) | 204 | 53 |
| Breast Cancer | West, (2001) | 87 | 84 |
| Lymphoma Cancer | Shipp, (2002) | 1059 | 82 |
| CNS Cancer | Pomeroy, (2002) | 48 | 84 |

Table 3 showed the subsample variable size using EBRF and RF. It could be observed that RF tends to select minimal variable size no matter the total number of variables available. The values were later used to obtain the probability that at least one relevant variable is contained in the variable selected $P(q \geq 1)$. Thus;

$$P(q \geq 1) = 1 - \left[\frac{\binom{q}{0}\binom{p-q}{q}}{\binom{p}{q}}\right]$$

$$P(q \geq 1) = 1 - \left[\frac{\binom{p-q}{q}}{\binom{p}{q}}\right]$$

$$\lim_{p\to\infty} P(q \geq 1) = 1 - e^{-1}$$

$$\lim_{p\to\infty} P(q \geq 1) \approx 0.63$$

The above derivation implies that irrespective of $p$, RF can only guarantee 63% of relevant subset present in the entire predictor space. Thus, for [30] dataset, RF subset size empirical estimate guarantees selection of about 62.8% of relevant variables. Similar, percentile values were observed for the other datasets. This implies about 37% of the relevant variable would not be selected, thus leading to sub-optimal model fitting and reduction in accuracy or performance. On the other hand, EBRF estimate varies for the different datasets. The $P(q \geq 1)$ returned using EBRF is almost equal to 1 in most datasets except for West and Pomeroy datasets. This implies, EBRF approach guarantees about 100% selection of relevant variables. Fig. 1 presents the detailed comparison of $P(q \geq 1)$ for the various datasets.

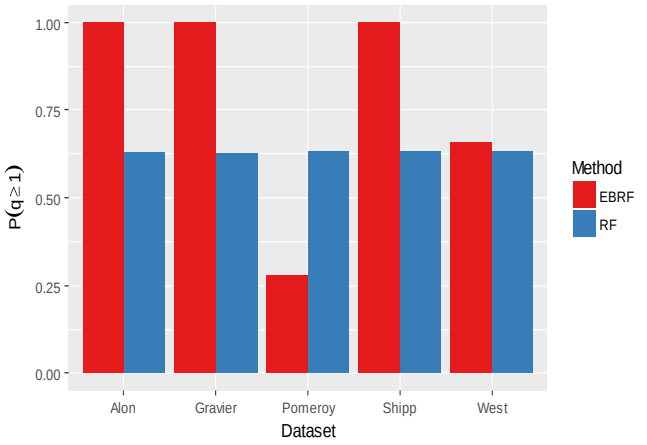


**Fig. 1:** Probability of selecting at least one relevant variable for the five datasets using RF and EBRF

**Table 4:** Percentage (%) Accuracy (ACC), Balance Accuracy (BACC), Sensitivity (Sens), Specificity (Specs) and Area under receiver operating characteristics curve (AUC)

| Dataset | Method | ACC | BACC | Sens | Specs | AUC |
|---|---|---|---|---|---|---|
| Alon, (1999) | EBRF | 87.2 | 85.9 | 81.7 | 90.0 | 87.5 |
| | RF | 82.1 | 79.7 | 71.7 | 87.5 | 87.1 |
| Gravier, (2010) | EBRF | 81.7 | 77.2 | 91.2 | 63.3 | 86.7 |
| | RF | 76.4 | 68.8 | 92.0 | 46.0 | 83.5 |
| West, (2001) | EBRF | 83.5 | 84.8 | 85.1 | 85.1 | 87.3 |
| | RF | 64.0 | 67.5 | 70.0 | 65.0 | 75.3 |
| Shipp, (2002) | EBRF | 91.3 | 85.9 | 96.6 | 75.0 | 95.8 |
| | RF | 89.8 | 79.2 | 98.3 | 60.0 | 95.0 |
| Pomeroy, (2002) | EBRF | 70.0 | 62.5 | 35.0 | 90.0 | 78.0 |
| | RF | 61.5 | 49.6 | 10.0 | 89.2 | 65.6 |
| Average | EBRF | 82.7 | 79.3 | 77.9 | 80.7 | 83.3 |
| | RF | 62.3 | 57.5 | 57.0 | 58.0 | 63.5 |

Table 4 showed the performance of the two methods across five datasets used. On average, the update on RF largely improved the overall predictive performance. The large update can be attributed to an optimal number of subsampled variables used by EBRF. EBRF also enhanced the class-specific performance which is an essential metric in medicine. On average, EBRF is as sensitive as specific. That's it can correctly identify the presence of disease at least 78% of the time as well as correctly identify the absence of a disease at least 80% of time. Also, the false alarm rate (1 – specificity) for EBRF is approximately half of RF. Fig. 2 shows the ROC curves for the various datasets. The thick lines represent EBRF while the dotted lines represent RF. For Alon, (1999), Gravier, (2010) and West, (2001) datasets, the AUC of EBRF is higher than EF. This implies better predictive performance for EBRF at threshold points ranging from 0 – 1. However, for Shipp, (2002) and Pomeroy, (2002), the AUC is relatively the same at varying threshold points.

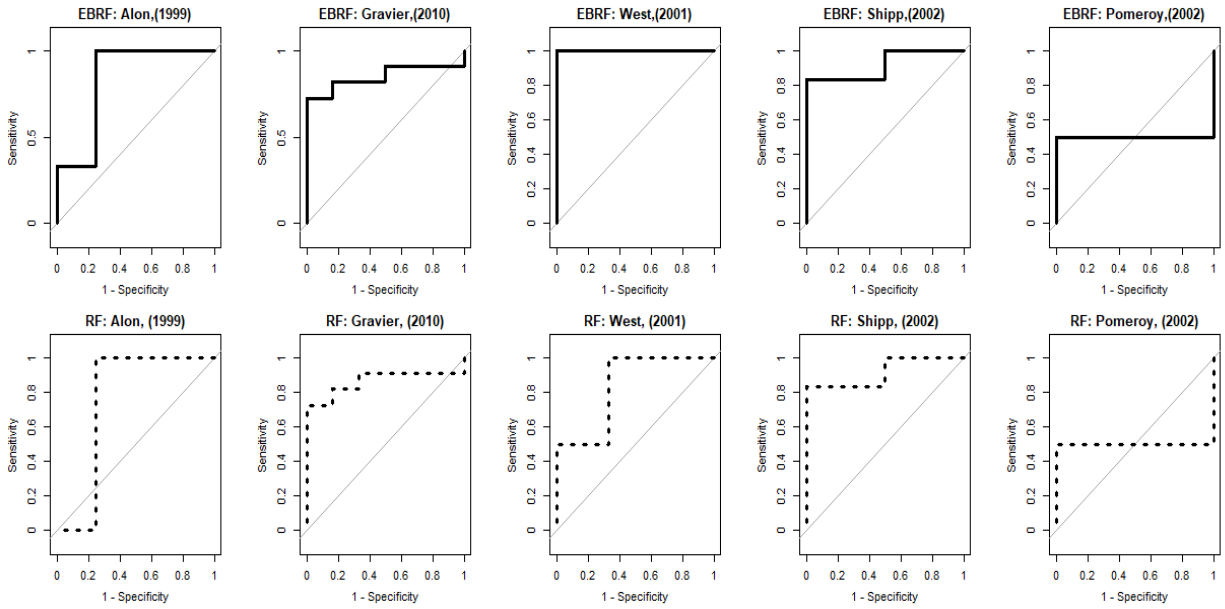


**Fig. 2:** Receiver operating characteristics curves (ROC) for the five datasets using EBRF and RF. The thick lines represent AUC for EBRF while the dotted lines represent AUC for RF. For Alon, (1999), Gravier, (2010) and West, (2001) datasets, the AUC of EBRF is relatively higher than EF. The EBRF better performance can be deduced from the approximately right-angled triangle formed with the datasets ROC curves. This further implies better predictive performance for EBRF at varying threshold points ranging from 0 – 1. However, for Shipp, (2002) and Pomeroy, (2002), the AUC is relatively the same at varying threshold points.

## 7. Conclusion

In this paper, an attempt is made to review the Random Forest (RF) algorithm by updating the subsampling variables selection method used. We replaced the arbitrary subsample variable size by an optimal empirical Bayes estimate. The results from the performance analysis using the new method revealed its high predictive performance strength. In almost all the datasets used, the new method largely improved the overall and class-specific accuracy of predicting the disease outcome. Also, a relatively lower false alarm rate was achieved with the new approach in all the datasets used.

## Acknowledgement

The authors would like to thank the Ministry of Higher Education (MOHE) and Office of Research, Innovation, Commercialization and Consultancy Office (ORICC), Universiti Tun Hussien Onn Malaysia (UTHM) for financially supporting this research under the postgraduate research grant (GPPS) [Vot, U607].

## References

[1] Hernández B, Raftery AE, Pennington SR & Parnell AC (2018), Bayesian additive regression trees using Bayesian model averaging. *Statistics and Computing* 28(4), 869-890.

[2] Hastie T, Tibshirani R & Wainwright M (2015), *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

[3] Sarica, A., Cerasa, A., & Quattrone, A. (2017). Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. *Frontiers in Aging Neuroscience* 9, 329.

[4] Gündüz N & Fokoue E (2017), Predictive performances of implicitly and explicitly robust classifiers on high dimensional data. *Communications faculty of sciences university of ankara-series a1 mathematics and statistics* 66(2), 14-36.

[5] Banjoko AW, Yahya WB, Garba MK, Olaniran OR, Dauda KA & Olorede KO (2015), Efficient Support Vector Machine Classification of Diffuse Large B-Cell Lymphoma And Follicular Lymphoma mRNA Tissue Samples. *Annals. Computer Science Series* 13(2), 69-79.

[6] Olaniran, OR, Olaniran SF, Yahya WB, Banjoko AW, Garba MK, Amusa LB & Gatta NF (2016), Improved Bayesian Feature Selection and Classification Methods Using Bootstrap Prior Techniques. *Annals. Computer Science Series* 14(2), 46 – 52.

[7] Olaniran OR & Abdullah MAA (2017), Gene Selection for Colon Cancer Classification using Bayesian Model Averaging of Linear and Quadratic Discriminants, *Journal of Science and Technology: Special Issue on the Application of Science and Technology* 9(3), 140-144.

[8] Breiman L (2001) Random forests. *Machine Learning*, 45, 5–32.

[9] Kapelner A & Bleich J (2015), Prediction with missing data via Bayesian additive regression trees. *Canadian Journal of Statistics* 43(2), 224-239.

[10] Breiman L, Friedman J, Stone CJ & Olshen RA (1984), *Classification and regression trees*. CRC press.

[11] Genuer R, Poggi JM & Tuleau C (2008), Random Forests: some methodological insights. *arXiv preprint arXiv:0811.3619*.

[12] Huang BF & Boutros PC (2016), The parameter sensitivity of random forests. *BMC bioinformatics* 17(1), 331.

[13] Robnik-Šikonja M (2004, September), Improving random forests. In *European conference on machine learning* (pp. 359-370). Springer, Berlin, Heidelberg.

[14] Boinee P, De Angelis A & Foresti GL (2005), Meta random forests. *International Journal of Computationnal Intelligence* 2(3), 138-147.

[15] Chaudhary A, Kolhe S & Kamal R (2016), An improved random forest classifier for multi-class classification. *Information Processing in Agriculture* 3(4), 215-222.

[16] Hwang K, Lee K & Park S (2017), Variable selection methods for multi-class classification using signomial function. *Journal of the Operational Research Society* 68(9), 1117-1130.

[17] Chipman HA, George EI & McCulloch RE (2010), BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1), 266-298.

[18] Pratola MT (2016) Efficient Metropolis-Hastings proposal mechanisms for Bayesian regression tree models. *Bayesian analysis* 11(3), 885-911.

[19] Chipman HA, George EI & McCulloch RE (1998), Bayesian CART model search. *Journal of the American Statistical Association* 93(443), 935-948.

[20] Taddy M, Chen CS, Yu J & Wyle M (2015), Bayesian and empirical Bayesian forests. *arXiv preprint arXiv:1502.02312*.

[21] Efron B (1979), Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1-26.

[22] Rubin, DB (1981), The Bayesian bootstrap. *The annals of statistics* 9(1), 130-134.
[23] Efron B (2012), *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge University Press.
[24] Olaniran OR & Yahya WB (2017), Bayesian Hypothesis Testing of Two Normal Samples using Bootstrap Prior Technique. *Journal of Modern Applied Statistical Methods* 16(2), 618-638.
[25] Wang S, Zhang J & Lawson AB (2016), A Bayesian normal mixture accelerated failure time spatial model and its application to prostate cancer. *Statistical methods in medical research* 25(2), 793-806.
[26] Yahya WB, Olaniran OR & Ige SO (2014), On Bayesian Conjugate Normal Linear Regression and Ordinary Least Square Regression Methods: A monte Carlo Study. *Ilorin Journal of Science* 1(1), 216-227.
[27] Olaniran OR & Abdullah MAA (2018), Bayesian Analysis of Extended Cox Model with Time-Varying Covariates using Bootstrap Prior. *Journal of Modern Applied Statistical Methods*, Accepted. In press.
[28] Peskun P (2016), Some Relationships and Properties of the Hypergeometric Distribution. *arXiv preprint arXiv:1610.07554*.
[29] Dyer D & Pierce RL (1993), On the choice of the prior distribution in hypergeometric sampling. *Communications in Statistics-Theory and Methods* 22(8), 2125-2146.
[30] Olaniran OR & Abdullah MAA (2018), BayesRandomForest: An R Implementation of Bayesian Random Forest for Regression Analysis of High-Dimensional Data. *Romanian Statistical Review* 66(1), 95-102.
[31] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D & Levine AJ (1999), Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences* 96(12), 6745-6750.
[32] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME & Allen JC (2002), Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415(6870), 436-442.
[33] West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R & Nevins JR (2001), Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences* 98(20), 11462-11467.
[34] Gravier E, Pierron G, Vincent-Salomon A, Gruel N, Raynal V, Savignoni A & Fourquet A (2010), A prognostic DNA signature for T1T2 node-negative breast cancer patients. *Genes, chromosomes and cancer* 49(12), 1125-1134.
[35] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC & Ray TS (2002), Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine* 8(1), 68-74.
[36] Fawcett T (2006), An introduction to ROC analysis. *Pattern recognition letters* 27(8), 861-874.
[37] Ramey JA (2016). datamicroarray: Collection of Data Sets for Classification.https://github.com/ramhiser/datamicroarray,http://ramhiser.com.
[38] Yahya WB, Olaniran OR, Garba MK, Oloyede I, Banjoko AW, Dauda KA & Olorede KO (2016), A Test Procedure for Ordered Hypothesis of Population Proportions Against a Control. *Turkiye Klinikleri Journal of Biostatistics* 8(1).
[39] Jamil SAM, Abdullah MAA, Kek SL, Olaniran OR & Amran SE (2017, September), Simulation of parametric model towards the fixed covariate of right censored lung cancer data. In *Journal of Physics: Conference Series* 890(1), p. 012172.
[40] Adeleke AO, Samsudin NA, Mustapha A & Nawi NM (2017) Comparative analysis of text classification algorithms for automated labelling of Quranic verses. *Int. J. Advanc. Sci. Eng. Info. Tech* 7, 1419-1427.