International Journal on Advanced Science Engineering Information Technology

Vol.12 (2022) No. 6
ISSN: 2088-5334

Bayesian Model Averaging (BMA) Based on Logistic Regression for Gene Selection and Classification of Animal Tumor Disease on Microarray Data

Heri Kuswanto [a,*], Ika Nur Laily Fitriana [a]

[a] *Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia*
Corresponding author: [*]heri_k@statistika.its.ac.id

***Abstract*— Tumor is one of the deadly diseases which is frequently to be found in animals. However, identifying whether an animal has a tumor still becomes a big challenge. Classification of tumor disease can be done through gene expression, which consists of hundreds of genes, but only a small number of samples is taken. This data structure is called microarray data having the characteristic of high-dimensional data. The choice of a single model can be a problem for high-dimensional data because it ignores model uncertainty. This research proposed to use Bayesian Model Averaging (BMA) to model the uncertainty model by averaging the posterior distribution of all best models, weighted by their posterior model probabilities. Selecting relevant genes to diagnose animal tumors is very important; hence, variable selection needs to be carried out. The selection of predictor variables is carried out by using the iterative BMA algorithm. The BMA results showed that from 335 gene expressions, 12 genes were selected to be relevant genes for classifying whether the animals have a tumor or normal. Moreover, from 2^{335} possible models formed, 12 of the best models are selected. The accuracy of BMA results is assessed using the Brier Score, resulting from a value indicating that the BMA model is good enough to classify animals, whether they have a tumor or not. This research has proven that BMA with logistic performance has very good predictability; hence, the method can be applied to classify other diseases.**

***Keywords*— Animal tumor; BMA; gene expression; microarray.**

Manuscript received 13 Oct. 2021; revised 25 Jan. 2022; accepted 20 May 2022. Date of publication 31 Dec. 2022.

I. INTRODUCTION

Animal tumor disease is one of the deadly diseases which typically attacks pets such as dogs and cats. The Indonesian Veterinary Medicine Association revealed that in 2013 the percentage of tumors in pets was 5% - 10%. Recently, the incidence of tumor disease in animals has been quite high, especially in dogs and cats. In general, diagnosing animals affected by tumor disease occurs when the animal is already in an advanced stage, and late handling of animals infected with tumors can be fatal. Therefore, animal tumor disease must be diagnosed as early as possible before tumor cells spread to other internal organs.

One of the current ways to classify animal diseases is by investigating gene expression. Microarray data is genetic information data in the form of gene expression. Microarrays can evaluate the expression of hundreds to thousands of genes and simultaneously monitor ongoing biological processes [1], [2]. Furthermore, the thousand gene expressions representing animal tissue will be classified as a tumor or healthy tissue. Microarray is part of high dimensional data because it has hundreds to thousands of features. One of the main challenges in analyzing microarray data is that the number of genes (predictor variables) exceeds the amount of tissue accessible. In addition, it is usually only a subset of genes relevant to differentiating into different classes [3], [4].

The previous study on gene expression data as microarray was carried out by Li and Yang [5], which applied the averaging model and ensemble model approaches to classify samples in microarray data. A more recent study by Yu et al. [6] used the Jackknife model averaging for predicting gene expression. Another study by Astuti [7] on identifying differences in microarray experiment gene expression with Bayesian mixture model averaging (BMMA) showed that the BMMA-normal model could adequately identify the ID group of Chickpea data genes in the upregulated, regulated, and down-regulated groups.

In this study, animal disease gene expression classification consists of hundreds of genes. However, the number of

samples is very small. Therefore, it needs to be resolved with an appropriate method. A single model approach can be a problem for high-dimensional microarray data due to a large number of genes while the number of samples is small. Therefore, to overcome these problems, the Bayesian Model Averaging (BMA) method is used. The Bayesian Model Averaging (BMA) approach offers the best alternative solution to this problem [8], [9]. Bayesian model averaging has been applied in other fields such as economic and financial [10]–[13], hydrology [14]–[17], engineering [18], behavioral research [19], environmental [20], agriculture [21], climate and meteorological [22]–[24] and many others.

BMA is a Bayesian approach that combines all possible models and averages the models by the posterior distribution of the selected best models [25]–[27]. Zhou, Liu, and Dannenberg [28] argued that the BMA method captures and measures complex relations between gene expression patterns and sample characteristics in microarray data. A study by Yeung, Bumgarner, and Raftery [3] reveals that applying BMA method for gene selection and microarray data classification led to high accuracy. A study by Annest et al. [29] applied iterative BMA to microarrays (Breast Cancer) data and found that iterative BMA can capture the uncertainty of models for gene selection with excellent performance.

In this study, the classification of animal tumor disease involves hundreds of genes (in this case, as predictor variables) with only a few samples; hence, the selection of variables needs to be done to obtain only relevant genes for classification. The predictor variable selection is done using the iterative Bayesian Model Averaging algorithm. The accuracy of classification is assessed by using Brier Score [30]. Using selected genes is expected to easily classify tumor and non-tumor diseases in animals using microarrays.

II. Material and Method

A. Dataset Description

The data used in this study are secondary data sourced from medicinal life science from one of the private universities in Japan. The data used is gene data with micro- RNA type, namely miRNA. This data consists of 29 animal samples consisting of 335 small RNA genes. The research variables used in this study are 335 predictor variables in the form of microRNA genes (miRNA) and one response variable in the form of animal disease categories that are affected by tumors or not affected by tumors.

A tumor is an abnormal growth of body cells. Gene expression is a series of processes using information from a gene to synthesize functional gene products. MicroRNA or miRNA is a small single bundle of ribonucleic acid (RNA) bundles (between 21 and 24 nucleotides in length) that inhibit the role (down-regulate) of target genes in the post-transcription stage of gene expression.

Gene expression comes from microarrays experiments. The microarray experiment is a data collection technique by using a platform that is the result of duplication of the original object identifier [31]. One technology for gene sequencing is the Next Generation Sequencing (NGS) platform, which provides genetic information in one run of the tool [32]. Data from microarrays have the following characteristics:

- The sample size that can be observed is minimal (few) due to limited funds, human resources, time and the availability of sample sizes.
- The characteristics of variables (genes) that can be observed are tremendous, reaching tens of thousands of characteristics (genes) in each experiment.

Based on the characteristics of microarray data, in microarray data analysis, it needs particular action because parametric statistics require a large enough sample size to meet the degrees of modeling freedom. If the assumptions are violated, the conclusions from the analysis results will be biased.

B. Bayesian Analysis

Bayesian analysis is a statistical method based on a posterior probability distribution model with a structure as a combination of two pieces of information, namely past data information (prior) and observational data (likelihood) [33].

The Bayesian analysis concept can be illustrated as follows; given an observation x which has a likelihood function $f(x|\theta)$, then the information about the parameter θ that is known before the observation is made is called prior θ, denoted as $p(\theta)$. Furthermore, the posterior probability distribution of θ, which is $p(\theta|x)$ can be determined based on the probability rule of the Bayes Theorem as follows:

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{f(x)} \quad (1)$$

where $f(x)$ is a normalized constant. Equation (1) can be written in a proportional form as:

$$p(\theta|x) \propto f(x|\theta)p(\theta) \quad (2)$$

posterior $\propto$ (likelihood function) x (prior).

Equation (2) shows that the posterior probability is proportional to the multiplication between the likelihood function and the prior probability of the model parameters. It means that the prior information will be updated using sample information in likelihood data to obtain the updated information used in decision making [33].

C. Bayesian Model Averaging (BMA)

Bayesian Model Averaging (BMA) is a Bayesian approach that combines all possible models that can be formed by averaging the posterior distribution of all the best models, weighted by the probability of the posterior model. The idea of BMA is to capture model uncertainty to obtain the best model [8], [25].

This research uses logistic regression to predict the class of tumor and non-tumor diseases in animals $Pr(Y = 1|D, M_k)$ which can be expressed as equation (3).

$$Pr(Y = 1|D, M_k) = ln\left[\frac{Pr(Y = 1|D, M_k)}{Pr(Y = 0|D, M_k)}\right] = b_0 + b_1x_1 + \ldots + b_px_p \quad (3)$$

where x_i represents the expression of the selected gene and b_i is a regression parameter, and $i = 1,2, \ldots p$. Suppose there are p predictor variables, then there are $k = 2^p$ models that might be formed assuming that there is no interaction between the predictor variables.

In the case of classification with two classes, let Y be the response variable (class), where $Y = 0$ or 1. Consider the dataset D as the one for which the classes or label are known. Then, the equation of BMA model is shown (4).

$$Pr(Y = 1|D) = \sum_{k=1}^{K} Pr(Y = 1|D, M_k) * Pr(M_k|D) \quad (4)$$

where the posterior probability of $Y = 1$ given by dataset D is the weighted average of the posterior probability of $Y = 1$ given by dataset D and model M_k multiplied by the posterior probability of model M_k given dataset D. The sum of the whole model M_k is the posterior predictive model $Pr(Y = 1|D)$.

Using the Bayes theorem, the posterior probability for the M_k model can be calculated using the following equation.

$$P(M_k|D) = \frac{P(M_k)P(D|M_k)}{P(D)} \quad (5)$$

where $P(D|M_k)$ and $P(M_k)$ are likelihood function and prior probability for M_k model, respectively. Prior probability $P(M_k)$ and $\pi_k(\beta_k)$ for model M_k determines the initial description of the model uncertainty. Based on information from dataset D, changes are made to the model uncertainty description based on the posterior probability model $P(M_k|D)$. The posterior mean Δ is given as follows in equation (6).

$$E(\Delta|D) = \sum_{k=0}^{K} Pr(M_k|D)E(\Delta|M_k, D) \quad (6)$$

$E(\Delta|D)$ shows the weighted expectation value Δ for each possible combination model (priors and models determine weight). Meanwhile, the variance $\Delta|D$ is as follows.

$$Var(\Delta|D) = \sum_{k=0}^{K} (var(\Delta|D, M_k) + [E(\Delta|M_k, D)]^2) Pr(M_k|D) - E(\Delta|D)^2 \quad (7)$$

As follows is the posterior probability that gene (x_i) is principal predictors.

$$Pr(b_i \neq 0|D) = \sum_{M_k \text{ which gen } i \text{ is principal}} Pr(M_k|D) \quad (8)$$

Generally, the posterior probability related with gene (x_i) is the total of posterior probabilities related with all selected M_k models that contain the gene (x_i).

In implementing BMA, determining the prior distributions is an important step. In this case, the prior distributions that must be determined are the prior distribution for the $P(M_k)$ and the prior distribution of the parameter $\pi(\boldsymbol{\beta_k})$. Because there is only little information about the probability of a model, thus it is assumed that all models have the same probability of being selected as the best model. So that the prior probability of the model $P(M_1), \ldots, P(M_k)$ is assumed to have a uniform distribution as follows.

$$P(M_k) = \frac{1}{k} \quad (9)$$

In the case of logistic regression, according to Raftery [22], the prior distribution of the $\boldsymbol{\beta_k}$ can be a multivariate normal distribution with mean $\widehat{\boldsymbol{\beta}}_{MLE}$ and variance I^{-1} which is the inverse of the expected Fisher information matrix for one observation data. This distribution can be thought of as the distribution of the prior parameter $\boldsymbol{\beta_k}$, which contains an amount equal to the amount of information about the parameter in one observation data [34]. Bayesian logistic regression can also be modelled using an adaptive MCMC [35].

Meanwhile, the process of calculating the integral on the marginal likelihood function in equation (6) oftenly leads to a non-analytical solution. To deal with this, approach is needed, namely the Bayesian Information Criteria (BIC) approach will be applied. The formula for calculating BIC is shown in equation (10).

$$BIC = -2 \log L(\widehat{\boldsymbol{\beta}}) + (p + 1) \log(n) \quad (10)$$

With the BIC approach, the posterior probability of the $\Pr(M_k|D)$ model in equation (5) can be expressed as

$$\Pr(M_k|D) = \frac{e^{-0,5*(BIC-mBIC)}}{\sum e^{-0,5*(BIC-mBIC)}} \quad (11)$$

where the maximum value of BIC is the BIC value of all models that have the highest value with the following formula

$$mBIC = maks\{BIC_k, k = 1,2, \ldots K\} \quad (12)$$

One of the challenges in analysis with the BMA method for microarray data is the number of models that the algorithm can explore. If there are some G genes, then a possible 2^G models is formed [29]. One of the methods proposed for selecting models in BMA is the Occam's Window method [28] which selects the models included in the BMA formula based on its posterior probability. The model accepted by this method (a model that can be included in BMA modeling) must fulfill equation (13).

$$\mathcal{A}' = M_k: \frac{max_l\{Pr(M_l|D)\}}{Pr(M_k|D)} \leq c \quad (13)$$

where $\mathcal{A}'$ is the posterior odds of the k-model with the limit value of c is 20. The limit value of the k-th model selection can be entered or not in BMA modeling. It is equivalent to the area of acceptance and rejection of the hypothesis with a significant level of $\alpha = 5\%$ when using test criteria through p-value. The value of 20 refers to the following Bayes Factor (BF) tabulation. Tabulation of BF values is shown in Table 1.

TABLE I
BAYES FACTOR TABULATION

2 log (B_{10})	(B_{10})	Evidence to Reject H_0
0 – 2	1 – 3	None
2 – 6	3 – 20	Positive
6 – 10	20 – 150	Strong
>10	>150	Very strong

Based on Table 1, the BF value of 20 indicates that the hypothesized data distribution model is positive according to the observed data. The maximum$_l$ formula ($P(M_l|\boldsymbol{x})$) in equation (7) is the l model in $\boldsymbol{M}$, which has a high posterior probability value. The posterior probability of each significant model parameter is determined by averaging the posterior probabilities of each parameter from *the* selected best models [18]. If a model has a value of $\mathcal{A}'$ greater than $c = 20$, the model will be eliminated. After this step is completed, the remaining group of models forms the set M_k to be used in equation (3).

D. Iterative BMA Algorithm

In general, the iterative BMA method works by iteratively applying the conventional BMA to a set of minimized predictor variables (w) known as the BMA window. Traditional BMA may be used to process the BMA window since it is small enough. Iterative BMA is achieved by ranking genes using the univariate gene selection approach, then using the conventional BMA algorithm successively to the sequenced genes. The between-group to within-group sum of squares (BSS/WSS) ratio was used to calculate the first gene sequencing [29]. The important variables are likely to be genes with a lot of variation across classes and a lot of variation within classes. BSS/WSS is a univariate gene selection technique in which genes with a high BSS/WSS ratio are excellent candidates for class prediction [3].

Let D_{ij} represent the level of gene j expression in sample i, D_{kj} symbolize the average level of gene j expression across all samples in class k and $\overline{Dj}$ imply the average level of gene j expression across all samples. The following formula is used to get the BSS/WSS ratio:

$$\frac{BSS(j)}{WSS(j)} = \frac{\Sigma_i \Sigma_k I(Y_i=k)(\bar{D}_{kj}-\bar{D}_j)^2}{\Sigma_i \Sigma_k I(Y_i=k)(D_{kj}-D_j)^2} \quad (14)$$

where $I(Y_i = k)$ is same with one if the sample i belongs to the group k, and zero if it is not. We compute the BSS/WSS ratio for each of the G genes in the first step of iterative BMA method and order the genes by their BSS/WSS ratio. The BMA window size refers to the number of variables (genes) used in every iteration of the conventional BMA algorithm. The iterative Bayesian Model Averaging algorithm is as follows:

a. Input: G genes and n samples data (D)
b. *Pre-Processing:* Using a univariate gene selection method, order all G genes. Let $X_1, X_2, \ldots, X_G$ be the ordered list of genes. The size of the BMA window is denoted by *w*.
c. Parameter: nbest and p, where p is the total number of genes to be processed such that w < p ≤ G.
   1. Initiation step is started with the *w* top ranked genes w (x_1, x_2, ..., x_w), and apply the conventional BMA algorithm. The result is a list of genes ordered from highest to lowest rank (w + 1) to p.
   2. Repeat this procedure until all p genes have been examined.
      - Using $Pr(b_i \neq 0|D) < c$, delete all *i* genes.
      - Determine the lowest minimum $P(b_i \neq 0|D)$, *minProbne0,* among the w genes in the current BMA window if all genes have $Pr(b_i \neq 0|D) \geq 1\%$. This step called adaptive threshold step.
      - Delete all genes $Pr(b_i \neq 0|D) < (minProbne0 + 1)\%$.
d. Output: selected genes and their probability posterior.

Genes having a high posterior probability $Pr(b_i \neq 0|D)$ are excellent candidates to be selected as the relevant gene. The gene with low posterior probability $Pr(b_i \neq 0|D)$ will be omitted. The threshold value used is c. However, Yeung, Bumgarner & Raftery [3] used 1% as the threshold value because 1% is a conservative threshold in which only genes that have a low posterior probability $Pr(b_i \neq 0|D)$ are omitted. So, in this study, used 1% for the threshold. A threshold of 1% usually results in good predictive performance [3].

E. Evaluation Performance

Brier Score is the score function to measure the accuracy of a probabilistic prediction [36]. In this study, prediction probability for each class, $\Pr(Y = k|D)$ are known. Let Y_i represent the response variable (class) of sample i for data with two classes (binary data), where $Y_i = 0$ or 1. $\Pr(Y_i = 1|D)$ is the probability of predicting that sample i belongs to class 1 or denoted as $\Pr(\hat{Y}_i)$. This is how the Brier Score is calculated:

$$BS = \sum_{i=1}^{n}(Y_i - Pr(\hat{Y}_i))^2 \quad (15)$$

Brier Score may be used to compare deterministic classification performance with probabilistic techniques like Bayesian Model Averaging (BMA). The Brier Score is a numerical number that goes from 0 to 1. The BMA model is more accurate the closer the number is near 0. In contrast, if the Brier Score is near to 1, the model will be inaccurate.

III. Result and Discussion

A. Characteristic of Data

Data on animal tumor microarrays consisted of 29 animal samples. The 29 animals are divided into two classes: animals with tumor tissue and do not have tumors (healthy). Figure 1 depicts the proportion of animals affected by tumor and non-tumor (healthy) diseases.

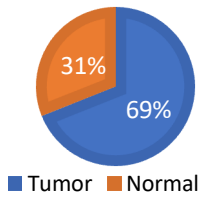

Fig. 1 Percentage of Animal Disease

Fig. 1 shows that the number of animals infected with tumors is doubled compared to animals that do not have tumors (healthy). The percentage of animals infected with tumors is 69% or equal to 20 animals included in the tumor class. These animals consist of three animals that have Mast Cell Tumor (MCT) tissue, five animals have MMelanoma (Malignant Melanoma) tissue, three animals have BMelanoma (Benign Melanoma) tissue, three animals have HCC (Hepatocellular Carcinoma) tissue, three animals have tissue BMGT (Benign Mammary Gland Tumor), and three animals have MMGT (Malignant Mammary Gland Tumor) disease.

Characteristics of animal gene expression or predictor variables for disease classification are shown in Fig. 2. Fig. 2 shows boxplots of the gene expression for each class of tumor-affected and non-tumor (healthy) animals. The box plot shows the differences in gene expression distribution (level) between animals with tumor disease and does not have tumor disease (healthy). Gene expression in animals infected with tumor disease is lower in value compared to healthy animals, and the healthy animal gene expression has a smaller range of values than the expression of animal genes infected with tumor disease.

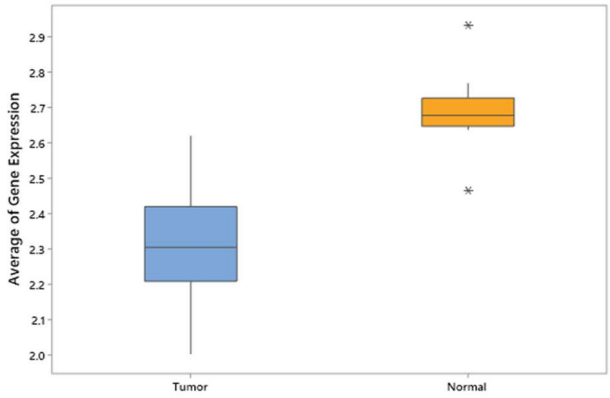

Fig. 2 Boxplot average of Gene Expression by Categories Disease

B. Gene Selection for Classification

This study uses 335 micro-RNA gene expressions as predictor variables to classify whether an animal is in the tumor and normal categories. The BMA approach considers all possible combinations of 335 variables. Thus, regardless of the interaction between variables, there are $2^{335} = 6.9992 \cdot 10^{100}$ models to predict the classification of animal tumor diseases.

The BMA implementation by Madigan and Raftery [25] is inefficient for high dimensional data, and hence, the iterative BMA algorithm is applied. The iterative BMA is started by sorting each gene expression according to the value of BSS/WSS. Furthermore, iteratively applies the BMA algorithm to the BMA window until the entire gene is applied. In one iteration in the BMA window, the model selection will be performed using the Occam Window selection method. Foreign genes will be removed from the BMA window and replaced by other genes not yet in the BMA window. This process will continue until all genes enter the BMA window. In the last iteration, there are 23 gene IDs left in the BMA window. The posterior probabilities of $Pr(b_i \neq 0|D)$, $E(b|D)$, and $SD(b|D)$ of the remaining predictor variables in the last iteration are shown in Table 2.

Table 2 column $Pr(b_i \neq 0|D)$ shows the posterior probability that the coefficient is not equal to zero. Whereas $E(b|D)$ shows the posterior mean of the coefficient or the value we expect in the BMA model, the posterior standard deviation shows the posterior SD, or standard deviation, giving a measure of the coefficient of variability.

Based on Table 2, there are 12 variables with posterior probability of predictor variables above 1%. This study uses 1% as the threshold value because 1% is a conservative threshold in which only genes that have a low posterior probability of $Pr(b_i \neq 0|D)$ are removed. Genes with a high posterior probability of $Pr(b_i \neq 0|D)$ are good candidates as relevant genes for predicting tumor disease in animals. Genes with a low posterior probability of $Pr(b_i \neq 0|D)$ will be removed. So that selected 12 relevant genes. The gene IDs are mir-23a, mir-23b, mir-491, mir-212, mir-424, mir-1468, mir-542, mir-450a, mir-450b, mir-503, mir-124-3- 1-2, and mir-8890 because it has a posterior probability of $Pr(b_i \neq 0|D)$ of more than 1%.

The higher the posterior probability value possessed by a gene indicates that the gene strongly influences the classification of animal tumor disease. Based on Table 2, the ID gene mir-23a is a gene with the highest posterior probability value, so the gene can be said to be the most crucial. The posterior probability of the variable is obtained by summing up the posterior probability of the model for each variable included in the model.

TABLE II
POSTERIOR PROBABILITY SELECTED GENES

Predictor	$Pr(b_i \neq 0 \mid D)$ (%)	$E(b\mid D)$	SD (b|D)
Intercept	100	70.16	65690
mir-23a	74.1	-0.1807	152.4
mir-23b	25.9	-0.7525	130.5
mir-491	4.8	-0.8332	1779
mir-212	4.8	-0.04464	1627
mir-424	4.8	-0.003397	615.7
mir-1468	4.8	-0.01904	46670
mir-542	4.8	$-2.67 \cdot 10^{-4}$	4.213
mir-450a	4.8	$-4.70 \cdot 10^{-5}$	1.382
mir-450b	4.8	$-2.09 \cdot 10^{-5}$	0.5206
mir-503	4.8	-0.00111	195.8
mir-124-3-1-2	4.8	-1.109	24890
mir-8890	4.8	0.1632	759.5
mir-219-1-2	0	0.0000	0.0000
mir-1838	0	0.0000	0.0000
mir-8859b	0	0.0000	0.0000
mir-26a-2-1	0	0.0000	0.0000
mir-138b	0	0.0000	0.0000
mir-326	0	0.0000	0.0000

The results of the iterative BMA show that there were 12 selected models out of $2^{335} = 6.9992 \cdot 10^{100}$ models formed. The twelve selected models are a combination of 12 relevant genes: ID mir-23a, mir-23b, mir-491, mir-212, mir-424, mir-1468, mir-542, mir-450a, mir-450b, mir-503, mir-124-3-1-2, and mir-8890. The results of selecting the best model are presented in Table 3.

Table 3 confirms that the mir-23a gene ID is a predictor in almost all models except the first model. It shows that the mir-23a ID gene results in a highly variable posterior probability because it is included in many models. The mir-23b gene ID has the second-highest possibility after the mir-23a gene id with a posterior probability of 25.9%.

Table 3 shows the model with the highest posterior model probability (PMP) of only 25.93% out of the total posterior probability, indicating that the model's uncertainty is quite high. Model 1 with PMP 0.259272 indicates that Model 1 contributes 25.93% of the total posterior probability. Likewise, model 2 contributes 25.93% of the total posterior probability.

In terms of contribution of each gene or predictor variable, the mir-23a gene ID contributes to eleven selected models so that it enormously influences the response variable. Therefore, the mir-23a gene ID has a significant posterior probability. ID mir-23b has the second-largest contribution compared to other genes even though it only appears in model 1. While other gene IDs only contribute to one model with low PMP, meaning that these genes' influence is quite small. The predictor variable or gene ID's coefficient has a consistent negative sign on all models. Based on the value of Bayesian Information Criterion (BIC), the first model and the second model are the models that have the smallest BIC, indicating that the first and the second models fit better than other models. BIC shows the goodness of fit of a model. The smaller the BIC value, the better the model formed.

TABLE III
SELECTED MODEL USING ITERATIVE BMA

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Intercept	47.77	118.95	50.46	54.01	51.32	57.94	55.99	56.65	58.89	59.42	59.76	72.55
mir-23a	-	-0.42	-0.11	-0.11	-0.11	-0.12	-0.144	-0.145	-0.146	-0.147	-0.15	-0.29
mir-23b	-0.29	-	-	-	-	-	-	-	-	-	-	-
mir-491	-	-	-17.31	-	-	-	-	-	-	-	-	-
mir-212	-	-	-	-	-	-0.93	-	-	-	-	-	
mir-424	-	-	-	-	-	-	-	-	-0.07	-	-	-
mir-1468	-	-	-	-	-0.39	-	-	-	-	-	-	-
mir-542	-	-	-	-	-	-	-	-	-	-	-0.006	-
mir-450a	-	-	-	-	-	-	-0.00098	-	-	-	-	-
mir-450b	-	-	-	-	-	-	-	-0.00043	-	-	-	-
mir-503	-	-	-	-	-	-	-	-	-	-0.23067	-	-
mir-124-3-1-2	-	-	-	-23.03	-	-	-	-	-	-	-	-
mir-8890	-	-	-	-	-	-	-	-	-	-	-	3.39
n	1	1	2	2	2	2	2	2	2	2	2	2
BIC	-90.92	-90.92	-87.54	-87.54	-87.54	-87.54	-87.54	-87.54	-87.54	-87.54	-87.54	-87.54
PMP	0.259	0.259	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048

Based on the posterior model's probability in Table 3, the BMA model to predict or classify tumor disease in animals is as follows.

$$\Pr(Y=1|D) = 0.2593M_1 + 0.2593M_2 + 0.0481M_3 + 0.0481M_4 + 0.0481M_5 + 0.0481M_6 + 0.0481M_7 + 0.0481M_8 + 0.0481M_9 + 0.0481M_{10} + 0.0481M_{11} + 0.0481M_{12}$$

where M_1, M_2, until M_{12} are logistic models. The logistic models have the following model equations.

$$M_1 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 47,77 - 0,29x_{67}(\text{mir-23b})$$

$$M_2 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 118,95 - 0,42x_{68}(\text{mir-23a})$$

$$M_3 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 50,46 - 0,11x_{68}(\text{mir-23a}) - 17,31x_{132}(\text{mir-491})$$

$$M_4 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 54,01 - 0,11x_{68}(\text{mir-23a}) - 23,0307x_{160}(\text{mir-124-3-1-2})$$

$$M_5 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 51,32 - 0,11x_{68}(\text{mir-23a}) - 0,39x_{123}(\text{mir-1468})$$

$$M_6 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 57,94 - 0,12x_{68}(\text{mir-23a}) - 0,93x_{32}(\text{mir-212})$$

$$M_7 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 54,99 - 0,144x_{68}(\text{mir-23a}) - 0,000098x_{335}(\text{mir-450a})$$

$$M_8 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 55,65 - 0,145x_{68}(\text{mir-23a}) - 0,00043x_{334}(\text{mir}-450b)$$

$$M_9 = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 58,89 - 0,146x_{68}(\text{mir-23a}) - 0,07x_{80}(\text{mir-424})$$

$$M_{10} = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 59,42 - 0,147x_{68}(\text{mir-23a}) - 0,23x_{184}(\text{mir-503})$$

$$M_{11} = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 59,76 - 0,15x_{68}(\text{mir-23a}) - 0,006x_{295}(\text{mir-542})$$

$$M_{12} = \ln\left[\frac{\Pr(Y=1)}{\Pr(Y=0)}\right] = 72,55 - 0,29x_{68}(\text{mir-23a}) - 3,39x_{180}(\text{mir-136})$$

The BMA model will obtain the predictive probability for an animal to be classified as an animal with a tumor or an animal not infected by a tumor (normal). The higher the probability value, the greater the likelihood that an animal is included in animals that have tumor disease. The minimum opportunity limit to determine the class of animals based on the percentage of the number of animals that have tumors and not tumors that exist in Fig. 1. If the predictive probability is more than 0.69, the animals are classified into animal with tumor disease. The visualization of genes and selected models using the BMA iterative algorithm is shown in Fig. 3.

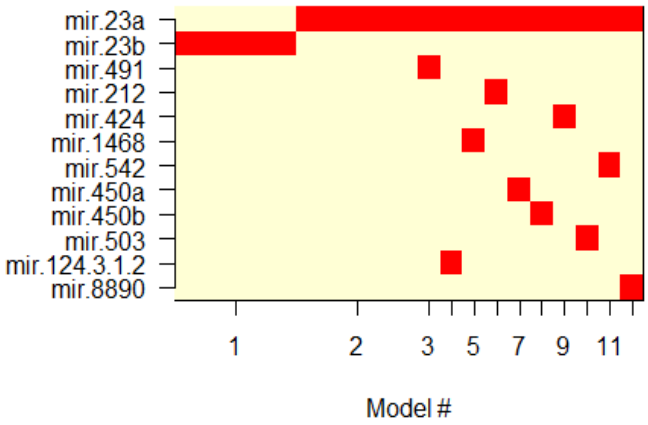

Fig. 3 Selected Genes and Models with Iterative BMA

In Fig. 3, the selected variable by BMA is shown on the vertical axis, and the selected BMA model is displayed on the horizontal axis. The variables (genes) are sorted from the highest to the lowest posterior probability from top to bottom. Models are sorted in order based on the largest to smallest posterior model (PMP) probability from left to right. The heatmap shows that the mirna-23a gene ID as a predictor variable for models 2 to 12 models mir-23b gene ID is only a predictor variable in model 1.Mir-491 gene ID enters into model 6.Mir-212 gene ID as a variable a predictor in model 11.Mir-424 gene ID as a predictor variable in model 3.Mir-1468 gene ID as a predictor variable in model 5. Mir-542 gene ID as a predictor variable in model 8. Mir-450a gene ID is

included in model 9. Mir-450b gene ID was included as a predictor variable in model 7. Mir-124-3-1-2 gene ID was included in model 4. Mir-503 gene ID was a predictor variable in model 10. While the mir-8890 gene ID was a variable predictor in model 12.

C. Performance of Classification

The results of classification accuracy are calculated using the Brier Score value. By using the 12 selected genes, the resulting Brier Score value is 0.00231799. This value is close to 0, so it can be said that the resulting BMA model is quite good. The smaller the Brier Score value, the better the BMA results in classifying animals into tumor or non-tumor categories (normal).

IV. Conclusion

This study used an iterative BMA algorithm to classify animal disease based on gene expression microarray data. The BMA results show that of 335 gene expressions, 12 relevant genes were selected to classify animals included in the tumor or normal class. The coefficient of the predictor variable or gene ID has a consistent sign on all models, which is negative. Besides, from 2^{335} possible models that were formed, selected 12 best models. The accuracy of the BMA results is measured using the Brier Score value. The resulting Brier Score is 0.00231799. This value is small enough and close to 0 so that it can be said that the BMA model is good enough to classify animals included in the tumor or normal categories.

References

[1] P. T. Ramadhani, U. Novia Wisesty, and A. Aditsania, "Deteksi Kanker berdasarkan Klasifikasi Data Microarray menggunakan Functional Link Neural Network dengan Seleksi Fitur Genetic Algorithm," *Indones. J. Comput.*, vol. 2, no. 2, pp. 11–22, Nov. 2017, doi: 10.21108/INDOJC.2017.2.2.173.

[2] M. S. Rao *et al.*, "Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies," *Front. Genet.*, vol. 9, 2019, doi: 10.3389/fgene.2018.00636.

[3] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, May 2005, doi: 10.1093/BIOINFORMATICS/BTI319.

[4] Ž. Avsec *et al.*, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nat. Methods 2021 1810*, vol. 18, no. 10, pp. 1196–1203, Oct. 2021, doi: 10.1038/s41592-021-01252-x.

[5] W. Li and Y. Yang, "How Many Genes are Needed For A Discriminant Microarray Data Analysis," *Bioinformatics*, vol. 9, pp. 2429–2437, 2002.

[6] X. Yu, L. Xiao, P. Zeng, and S. Huang, "Jackknife Model Averaging Prediction Methods for Complex Phenotypes with Gene Expression Levels by Integrating External Pathway Information," *Comput. Math. Methods Med.*, vol. 2019, 2019, doi: 10.1155/2019/2807470.

[7] A. B. Astuti, N. Iriawan, Irhamah, and H. Kuswanto, "Bayesian Mixture Model Averaging Untuk Mengidentifikasi Perbedaan Ekspresi Gen Percobaan Microarray," *Appl. Math. Sci.*, vol. 8, no. 145–148, pp. 7277–7287, 2017, doi: 10.12988/AMS.2014.49760.

[8] M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers, "A Conceptual Introduction to Bayesian Model Averaging," *Adv. Methods Pract. Psychol. Sci.*, vol. 3, no. 2, pp. 200–215, Jun. 2020, doi: 10.1177/2515245919898657.

[9] D. Kaplan, "On the Quantification of Model Uncertainty: A Bayesian Perspective," *Psychometrika*, vol. 86, no. 1, pp. 215–238, Mar. 2021, doi: 10.1007/S11336-021-09754-5/TABLES/5.

[10] M. F. J. Steel, "Model Averaging and Its Use in Economics," *J. Econ. Lit.*, vol. 58, no. 3, pp. 644–719, Sep. 2020, doi: 10.1257/JEL.20191385.

[11] Y. Ouyang, H. Cai, X. Yu, and Z. Li, "Capitalization of social infrastructure into China's urban and rural housing values: Empirical evidence from Bayesian Model Averaging," *Econ. Model.*, vol. 107, p. 105706, Feb. 2022, doi: 10.1016/J.ECONMOD.2021.105706.

[12] M. Camarero, S. Moliner, and C. Tamarit, "Japan's FDI drivers in a time of financial uncertainty. New evidence based on Bayesian Model Averaging," *Japan World Econ.*, vol. 57, p. 101058, Mar. 2021, doi: 10.1016/J.JAPWOR.2021.101058.

[13] B. K. Bierut and P. Dybka, "Increase versus transformation of exports through technological and institutional innovation: Evidence from Bayesian model averaging," *Econ. Model.*, vol. 99, p. 105501, Jun. 2021, doi: 10.1016/J.ECONMOD.2021.105501.

[14] J. Xu, F. Anctil, and M. A. Boucher, "Hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems," *J. Hydrol.*, vol. 578, p. 124002, Nov. 2019, doi: 10.1016/J.JHYDROL.2019.124002.

[15] S. Samadi, M. Pourreza-Bilondi, C. A. M. E. Wilson, and D. B. Hitchcock, "Bayesian Model Averaging With Fixed and Flexible Priors: Theory, Concepts, and Calibration Experiments for Rainfall-Runoff Modeling," *J. Adv. Model. Earth Syst.*, vol. 12, no. 7, p. e2019MS001924, Jul. 2020, doi: 10.1029/2019MS001924.

[16] Y. Hao, J. Baik, H. Tran, and M. Choi, "Quantification of the effect of hydrological drivers on actual evapotranspiration using the Bayesian model averaging approach for various landscapes over Northeast Asia," *J. Hydrol.*, p. 127543, Jan. 2022, doi: 10.1016/J.JHYDROL.2022.127543.

[17] P. Darbandsari and P. Coulibaly, "Introducing entropy-based Bayesian model averaging for streamflow forecast," *J. Hydrol.*, vol. 591, p. 125577, Dec. 2020, doi: 10.1016/J.JHYDROL.2020.125577.

[18] A. Rema and A. K. Swamy, "Use of Bayesian Model Averaging to Estimate Model Uncertainty for Predicting Strain in a Four-Layered Flexible Pavement," *ASCE-ASME J. Risk Uncertain. Eng. Syst. Part A Civ. Eng.*, vol. 7, no. 1, p. 04021002, Jan. 2021, doi: 10.1061/AJRUA6.0001123.

[19] S. Depaoli, K. Lai, and Y. Yang, "Bayesian Model Averaging as an Alternative to Model Selection for Multilevel Models," *Multivariate Behav. Res.*, vol. 56, no. 6, pp. 920–940, 2020, doi: 10.1080/00273171.2020.1778439.

[20] M. Gharekhani, A. A. Nadiri, R. Khatibi, S. Sadeghfam, and A. Asghari Moghaddam, "A study of uncertainties in groundwater vulnerability modelling using Bayesian model averaging (BMA)," *J. Environ. Manage.*, vol. 303, p. 114168, Feb. 2022, doi: 10.1016/J.JENVMAN.2021.114168.

[21] Y. Gao *et al.*, "Evaluation of crop model prediction and uncertainty using Bayesian parameter estimation and Bayesian model averaging," *Agric. For. Meteorol.*, vol. 311, p. 108686, Dec. 2021, doi: 10.1016/J.AGRFORMET.2021.108686.

[22] G. Zhang *et al.*, "Solar radiation estimation in different climates with meteorological variables using Bayesian model averaging and new soft computing models," *Energy Reports*, vol. 7, pp. 8973–8996, Nov. 2021, doi: 10.1016/J.EGYR.2021.10.117.

[23] F. Panahi, M. Ehteram, A. N. Ahmed, Y. F. Huang, A. Mosavi, and A. El-Shafie, "Streamflow prediction with large climate indices using several hybrid multilayer perceptrons and copula Bayesian model averaging," *Ecol. Indic.*, vol. 133, p. 108285, Dec. 2021, doi: 10.1016/J.ECOLIND.2021.108285.

[24] Y. Hao, J. Baik, and M. Choi, "Combining generalized complementary relationship models with the Bayesian Model Averaging method to estimate actual evapotranspiration over China," *Agric. For. Meteorol.*, vol. 279, p. 107759, Dec. 2019, doi: 10.1016/J.AGRFORMET.2019.107759.

[25] D. Madigan and A. E. Raftery, "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *J. Am. Stat. Assoc.*, vol. 89, no. 428, p. 1535, Dec. 1994, doi: 10.2307/2291017.

[26] M. Höge, A. Guthke, and W. Nowak, "Bayesian Model Weighting: The Many Faces of Model Averaging," *Water*, vol. 12, no. 2, p. 309, Jan. 2020, doi: 10.3390/W12020309.

[27] D. Fouskakis and A. I. Ntzoufras, "Bayesian Model Averaging Using Power-Expected-Posterior Priors," *Econometrics*, vol. 8, no. 2, p. 17, May 2020, doi: 10.3390/ECONOMETRICS8020017.

[28] X. K. Zhou, F. Liu, and A. J. Dannenberg, "A Bayesian model averaging approach for observational gene expression studies," *Ann.*

Appl. Stat., vol. 6, no. 2, pp. 497–520, Jun. 2012, doi: 10.1214/11-AOAS526.

[29] A. Annest, R. E. Bumgarner, A. E. Raftery, and K. Y. Yee, "Iterative bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data," *BMC Bioinformatics*, vol. 10, no. 1, pp. 1–17, Feb. 2009, doi: 10.1186/1471-2105-10-72/TABLES/9.

[30] K. Rufibach, "Use of Brier score to assess binary predictions," *J. Clin. Epidemiol.*, vol. 63, no. 8, pp. 938–939, Aug. 2010, doi: 10.1016/J.JCLINEPI.2009.11.009.

[31] Y. J. Guan, J. Y. Ma, and W. Song, "Identification of circRNA-miRNA-mRNA regulatory network in gastric cancer by analysis of microarray data," *Cancer Cell Int.*, vol. 19, no. 1, pp. 1–9, Jul. 2019, doi: 10.1186/S12935-019-0905-Z/FIGURES/7.

[32] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, Feb. 2009, doi: 10.1016/J.YGENO.2008.10.003.

[33] J. K. Kruschke, "Bayesian Analysis Reporting Guidelines," *Nat. Hum. Behav. 2021 510*, vol. 5, no. 10, pp. 1282–1291, Aug. 2021, doi: 10.1038/s41562-021-01177-7.

[34] A. E. Raftery, "Bayes Factors and BIC: Comment on 'A Critique of the Bayesian Information Criterion for Model Selection,'" *Sociol. Methods Res.*, vol. 27, no. 3, pp. 411–427, Feb. 1999, doi: 10.1177/0049124199027003005.

[35] K. Y. Y. Wan and J. E. Griffin, "An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models," *Stat. Comput.*, vol. 31, no. 1, pp. 1–11, Jan. 2021, doi: 10.1007/S11222-020-09974-2/TABLES/4.

[36] Z. Javanshiri, M. Fathi, and S. A. Mohammadi, "Comparison of the BMA and EMOS statistical methods for probabilistic quantitative precipitation forecasting," *Meteorol. Appl.*, vol. 28, no. 1, p. e1974, Jan. 2021, doi: 10.1002/MET.1974.