data

MDPI

Article

Consistency and Stability in Feature Selection for High-Dimensional Microarray Survival Data in Diffuse Large B-Cell Lymphoma Cancer

Kazeem A. Dauda [1,*] and Rasheed K. Lamidi [2]

[1] Department of Mathematics, University of Bergen, 5007 Bergen, Norway
[2] Department of Mathematics and Statistics, Kwara State University, Malete, P.M.B. 1530, Ilorin 23431, Kwara State, Nigeria; rasheedlamidi44@gmail.com
* Correspondence: kazeem.dauda@uib.no or qdauda70@gmail.com

Abstract: High-dimensional survival data, such as microarray datasets, present significant challenges in variable selection and model performance due to their complexity and dimensionality. Identifying important genes and understanding how these genes influence the survival of patients with cancer are of great interest and a major challenge to biomedical scientists, healthcare practitioners, and oncologists. Therefore, this study combined the strengths of two complementary feature selection methodologies: a filtering (correlation-based) approach and a wrapper method based on Iterative Bayesian Model Averaging (IBMA). This new approach, termed Correlation-Based IBMA, offers a highly efficient and effective means of selecting the most important and influential genes for predicting the survival of patients with cancer. The efficiency and consistency of the method were demonstrated using diffuse large B-cell lymphoma cancer data. The results revealed that the 15 most important genes out of 3835 gene features were consistently selected at a threshold p-value of 0.001, with genes with posterior probabilities below 1% being removed. The influence of these 15 genes on patient survival was assessed using the Cox Proportional Hazards (Cox-PH) Model. The results further revealed that eight genes were highly associated with patient survival at a 0.05 level of significance. Finally, these findings underscore the importance of integrating feature selection with robust modeling approaches to enhance accuracy and interpretability in high-dimensional survival data analysis.

Keywords: iterative Bayesian model averaging (IBMA); posterior probability; wrapper; parametric; filtering; semi-parametric

check for updates

Academic Editor: Rüdiger Pryss

Received: 8 January 2025
Revised: 5 February 2025
Accepted: 8 February 2025
Published: 18 February 2025

Citation: Dauda, K.A.; Lamidi, R.K. Consistency and Stability in Feature Selection for High-Dimensional Microarray Survival Data in Diffuse Large B-Cell Lymphoma Cancer. *Data* **2025**, *10*, 26. https://doi.org/10.3390/data10020026

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Diffuse large B-cell lymphoma (DLBCL) is one of the most prevalent forms of non-Hodgkin lymphoma and is characterized by its aggressive clinical behavior and diverse biological landscape [1]. DLBCL affects mature B-cells and is typified by rapid growth, necessitating prompt diagnosis and treatment [2]. Despite advancements in therapeutic protocols, the prognosis for patients with DLBCL remains variable, influenced by both the initial presentation and subsequent responses to therapy [3]. A key challenge in managing DLBCL lies in the heterogeneity of the disease at the genetic, epigenetic, and phenotypic levels. Such diversity necessitates sophisticated analytical approaches to unravel the underlying mechanisms of tumor progression and treatment response [4].

Recent years have registered explosive growth in high-throughput genomic technologies, culminating in the generation of large datasets that capture the complexities of DLBCL

at a molecular level [5,6]. The integration of these data into clinical applications can facilitate the discovery of predictive biomarkers, elucidate treatment pathways, and ultimately improve patient outcomes [6].

In order to extract meaningful information from high-dimensional data while managing the inherent risks of overfitting and model complexity, researchers have turned to various statistical methodologies [7,8]. Among these, filtering and wrapper methods are prominent, particularly in the context of variable selection [9]. Filtering techniques, such as correlation-based methods, enable researchers to assess the relationship between individual features and outcomes independently before fitting a model. On the other hand, wrapper methods, such as Iterative Bayesian Model Averaging (IBMA) [10–12], consider multiple models and their respective probabilities, utilizing model performance to inform variable selection [13]. Both methodologies play pivotal roles in optimizing feature spaces, thereby enhancing model interpretability and efficacy [14]. However, the IBMA method alone is computationally time-consuming and may include some redundant genes when applied to large-scale microarray survival datasets.

Once the relevant predictors have been identified through filtering and wrapper methods, the next step involves incorporating these variables into survival models. Survival analysis is a pivotal component of cancer research as it provides insights into time-to-event data, which are crucial for understanding patient prognoses. Both parametric and semi-parametric survival models have been widely employed in the analysis of cancer data, with each offering distinct advantages [15].

Feature selection is a critical step in data preprocessing, particularly for high-dimensional datasets, where irrelevant or redundant features can adversely affect model performance [16]. Numerous studies have proposed hybrid filter–wrapper approaches, which combine the strengths of both filters (that select features based on statistical criteria) and wrappers (that assess feature subsets using predictive models) [17,18]. These methods aim to improve feature selection efficacy, model accuracy, and computational efficiency.

One notable method is the Hybrid Genetic-Particle Swarm Feature Selection (HGP-FS) hybrid filter–wrapper, which integrates genetic algorithms with particle swarm optimization. This approach has been shown to effectively enhance feature subset selection, leading to improved classification accuracy and fewer irrelevant features in high-dimensional datasets [16]. Similarly, a combination of Ant Colony Optimization (ACO) within a wrapper–filter framework has demonstrated superior performance compared to conventional feature selection techniques, particularly in real-world datasets like facial emotion recognition and microarray data [19].

In the domain of sentiment analysis, hybrid filter–wrapper methods have proven to be particularly useful. By utilizing fewer features, these methods can achieve higher classification accuracy than state-of-the-art algorithms [9]. Furthermore, in time-series forecasting tasks, hybrid approaches combining Partial Mutual Information with firefly algorithms have significantly improved short-term load forecasting accuracy. This is achieved by reducing redundant features without sacrificing forecasting precision [20].

The Large-Margin Hybrid Feature Selection Algorithm (LMFS) stands out by offering improvements not only in classification performance but also in model interpretability. The LMFS reduces both the computational time and complexity of the classifier, making it more efficient compared to traditional filter or wrapper methods [21].

Further advancements in hybrid feature selection are observed in tri-objective evolutionary algorithms, such as the filter–wrapper-based nondominated Sorting Genetic Algorithm-II. This algorithm has provided competitive results across various datasets, outperforming existing feature selection techniques by balancing multiple objectives [22] as a novel hybrid dimension reduction technique designed to enhance the selection of

biomarker genes and improve the prediction of heart failure status in patients. The study highlighted the effectiveness, adaptability, and broad applicability of hybrid dimension reduction algorithms (HDRAs) within the framework of deep neural networks [23]. Lastly, hybrid methods that employ differential evolution have been shown to enhance classification performance while reducing the size of the feature subset and computational time, surpassing both traditional and more recent evolutionary feature selection techniques [24].

All these methods were primarily developed for classification and regression cases; however, most of them cannot be directly applied in high-dimensional survival analysis settings. Additionally, DLBCL datasets often involve thousands of gene features. Along with their heterogeneous nature, gene–gene interactions and pathways play a critical role in determining prognostic outcomes. For instance, interactions between BCL2 and MYC (double-hit lymphomas) are associated with significantly poorer prognoses in DLBCL patients [25]. This complexity makes them more challenging to handle using the existing methods, necessitating further filtering of noisy genes [4,26].

High-dimensional microarray survival data present a major challenge in cancer research (such as DLBCL), particularly when attempting to predict patient outcomes and understand the underlying biological mechanisms of diseases such as cancer [27]. The vast number of genes (features) compared to the limited number of samples increases the risk of overfitting and reduces model interpretability [28]. Traditional methods often struggle to effectively select the most relevant variables from high-dimensional data, which impairs survival models' accuracy and reliability. The need for sophisticated feature selection approaches is critical to ensure that the models used for predicting survival are both robust and informative [29,30].

This study seeks to address the challenges posed by high-dimensional microarray survival data by integrating two advanced feature selection methods: Iterative Bayesian Model Averaging (a wrapper method) and filtering techniques based on correlation. By combining these approaches, the research aims to improve the accuracy and interpretability of variable selection for high-dimensional datasets. These selected features will then be incorporated into both parametric and semi-parametric survival models, providing a comprehensive framework for predicting patient outcomes in cancer research. The improved survival models could significantly enhance personalized medicine by identifying key genetic markers, guiding treatment strategies, and offering better prognostic tools for clinicians. This hybrid approach offers a balance between model complexity and performance, leading to more precise survival predictions in clinical settings.

The paper is structured as follows: Section 2 describes the methodology, Section 3 outlines the parametric survival model, Section 4 illustrates the algorithm process flowchart, Section 5 presents the results, and Section 6 concludes the paper.

2. Methodology

In this section, we provide a comprehensive and detailed explanation of the two adopted methods that form the foundation of our approach. The first method, Iterative Bayesian Model Averaging, is a sophisticated wrapper technique designed to enhance model performance by systematically combining predictions from multiple models. The second method involves filtering techniques grounded in correlation analysis, which aim to identify and prioritize key variables based on their statistical relationships. Together, these methods offer a robust framework for achieving accurate and reliable results in our analysis.

2.1. Filtering Procedure

Filtering correlation is a feature selection method that evaluates how each feature (gene) correlates with the target variable. In this context, the target variable is the cumu-

lative hazard function ($\hat{\Lambda}(t)$) [31], which is estimated using the procedure outlined in Algorithm 1. This method operates under the assumption that highly correlated genes are more relevant for predicting the target outcome ($\hat{\Lambda}(t)$). For a gene g_i and the target outcome $\hat{\Lambda}(t)_i$ (cumulative hazard function), the correlation r is calculated using Pearson's correlation coefficient as follows:

$$r(g_i, \hat{\Lambda}(t)_i) = \frac{\sum_{i=1}^{n}(g_i - \overline{g_i})\left(\hat{\Lambda}(t)_i - \overline{\hat{\Lambda}(t)}\right)}{\sqrt{\sum_{i=1}^{n}(g_i - \overline{g_i})^2} \cdot \sqrt{\left(\hat{\Lambda}(t)_i - \overline{\hat{\Lambda}(t)}\right)^2}} \tag{1}$$

where g_i is the individual expression level of the gene for sample i, $\hat{\Lambda}(t)_i$ is the target value for sample i, $\overline{g_i}$ and $\overline{\hat{\Lambda}(t)}$ are the mean values of g_i and $\hat{\Lambda}(t)_i$, respectively, and n is the total number of samples. The primary goal of Algorithm 1 is to calculate the cumulative hazard function denoted as ($\hat{\Lambda}(t)$). Algorithm 1 employs the Nelson–Aalen method [31], a nonparametric estimator for the cumulative hazard function. The method is particularly suited for handling censored data, where not all individuals experience the event of interest during the observation period. The Nelson–Aalen estimator ensures that the risk is calculated based on individuals still "at risk" just before each observed event time.

Algorithm 1: Cumulative Hazard Function Null Model Computation

Input: Set of event times $T = \{t_1, t_2, \ldots, t_n\}$, status indicator $S = \{s_1, s_2, \ldots, s_n\}$
Output: Cumulative hazard estimates $\hat{\Lambda}(t)$ for all $t \in T$

Step 1: Fit the cumulative hazard model
Compute the Nelson–Aalen estimator for the cumulative hazard:

$$\hat{\Lambda}(t) = \sum_{t_i \le t} \frac{\delta_i}{Y_i},$$

where δ_i is the event indicator (1 for events, 0 otherwise), and Y_i is the number of individuals at risk at time t_i. **Step 2: Define jump points**
Create a set of jump points:

$$J = \{0, t_1, t_2, \ldots, t_k, \max(T)\},$$

where $t_1, t_2, \ldots, t_k$ are the observed event times, t_k represent k^{th} event time, and k is an index referring to the position of an event time.
Step 3: Extend survival probabilities
Define the extended survival probabilities $\hat{S}(t)$:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t = 0, \\ \hat{S}(t_i) & \text{if } t_i \in J, \\ \hat{S}(t_k) & \text{if } t = \max(T). \end{cases}$$

Step 4: Transform survival to hazard
Compute the cumulative hazard from the survival probabilities:

$$\hat{\Lambda}(t) = -\log\left(\hat{S}(t)\right).$$

Step 5: Initialize estimates
Create an empty array $\hat{\Lambda}_{\text{est}}$ with the same length as T.
Step 6: Assign cumulative hazard to intervals
foreach *interval* $[t_{j-1}, t_j]$ *for* $j = 2, \ldots, |J|$ **do**
  foreach $t \in [t_{j-1}, t_j]$ **do**
    Assign: $\hat{\Lambda}_{\text{est}}(t) = \hat{\Lambda}(t_{j-1})$.
  end
end
Step 7: Return results
Return the cumulative hazard estimates $\hat{\Lambda}_{\text{est}}$ for all $t \in T$.

In our context, the cumulative hazard function serves as the target outcome, and we aim to determine the correlation between this outcome and gene expression levels. The cumulative hazard function is preliminarily denoted $\hat{\Lambda}(t)$ in Equation (1), representing the target outcome in Algorithm 2.

Algorithm 2: Correlation Filtering Algorithm

Input: Gene expression data $\mathbf{X} = \{g_1, g_2, \ldots, g_n\}$, Target outcome $(\hat{\Lambda}(t))$, Threshold $\zeta = 0.001$

Output: A subset of top correlated genes

```
for i = 1 to n do
    Compute the correlation r_i between g_i and the target outcome y;
Rank genes g_i based on the absolute value of their correlation |r_i|;
Apply threshold ζ: Select genes where |r_i| > ζ;
Select the top genes with the highest absolute correlation values;
return Subset of top correlated genes;
```

Algorithm 2 employs the correlation filtering method for feature selection, evaluating features (genes) based on their degree of correlation with the target variable. A gene is selected for further analysis if the p-value associated with the correlation coefficient between the gene and the target outcome exceeds the predefined threshold ζ.

This selection criterion is based on statistical significance as the p-value indicates the likelihood of observing the given correlation by chance. If the p-value is greater than T, the correlation is deemed statistically insignificant, suggesting that the gene may not have a meaningful relationship with the target outcome. Consequently, such genes are excluded from the feature set to enhance the model's focus on relevant predictors. Conversely, genes with p-values below the threshold are retained as they exhibit statistically significant correlations with the target outcome, making them potentially valuable for predictive analysis.

2.2. Iterative BMA for Survival Analysis

We present our work on adapting the Iterative BMA approach to survival analysis, including many algorithmic changes. Initially, we utilize the Cox Proportional Hazards Model [32,33] to rank each individual gene in the preprocessing stage rather than using the BSS/WSS approach [34] to rank-order the genes. Because of its wide range of applications and ability to handle censored data, Cox regression is a commonly used method in the field of survival analysis. It measures a subject's hazard rate at time T using a semi-parametric approach, where the equation is presented in Equation (2)

$$h_i(t) = h_o(t)e^{x_i^T b} \tag{2}$$

where β is a fixed vector of mtry $\leq$ p coefficients, $h_o(t)$is the baseline hazard function, and $h_i(t)$ is the hazard for participant i at time t. The partial likelihood is utilized in the estimation of β.

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^T \beta}}{\sum_{j \epsilon R_i} e^{x_j^T \beta}} \tag{3}$$

where R_i is the set of indices, j, with $y_j \geq t_i$ (i.e., those still at risk at time t_i), and $j_{(i)}$ is the index of the observation for which an event occurred at time t_i.

Once the regression parameters in the Cox model are estimated by maximizing the partial likelihood, the genes can be ranked in descending order of their log–likelihood [35], which is presented in Algorithm 3.

Algorithm 3: Iterative BMA Algorithm for Survival Analysis

Pre-sorting step: rank the genes in descending order of their log–likelihood in Equation (3) using the Cox Proportional Hazards Model in Equation (2). The genes with a larger log–likelihood are assigned a higher ranking.
Apply the traditional BMA algorithm to the maxNvar (default is maxNvar = 25) top log-ranked genes.
Remove genes with low posterior probabilities in the predictive model; [35] used 1% as the threshold and eliminated genes with posterior probabilities less than 1%.

2.3. Posterior Model Probability in Bayesian Model Averaging (BMA)

Once the top-ranked genes are selected, *Bayesian Model Averaging (BMA)* is applied. BMA combines several models by computing a weighted average based on the posterior probabilities of the models. The *posterior probability* of a model ν_k is provided by Bayes' theorem, which is presented in Equation (4)

$$P(\nu_k \mid \text{data}) = \frac{P(\text{data} \mid \nu_k)P(\nu_k)}{\sum_j P(\text{data} \mid \nu_j)P(\nu_j)}, \tag{4}$$

where $P(\text{data} \mid \nu_k)$ is the likelihood of the data given model ν_k, $P(\nu_k)$ is the prior probability of model ν_k, and the denominator normalizes the probabilities across all possible models.

In survival analysis, $P(\text{data} \mid \nu_k)$ is the likelihood derived from the Cox model or other survival models. This equation enables the IBMA algorithm to account for model uncertainty by averaging across models based on their fit to the data.

2.4. Gene Posterior Probability

During the iterative process, genes with low posterior probabilities are removed from the predictive model. The posterior probability of a gene g within a model ν_k is computed as

$$P(g \mid \text{data}, \nu_k) = \frac{P(\text{data} \mid g, \nu_k)P(g \mid \nu_k)}{\sum_i P(\text{data} \mid g_i, \nu_k)P(g_i \mid \nu_k)}. \tag{5}$$

Equation (5) updates the prior probability of gene g based on its contribution to model ν_k. Genes with posterior probabilities below a certain threshold (e.g., 1%) are deemed to contribute little to the prediction of survival and are eliminated. This iterative removal step refines the model by focusing on the most relevant genes.

As the Bayesian Model Averaging (BMA) algorithm progresses, genes with a high posterior probability are kept, while genes with a low posterior probability are removed. These iterative steps continue until all genes have been evaluated. This procedure is described in Algorithm 3.

3. Parametric Survival Models

Parametric survival models analyze time-to-event data by assuming that survival times follow a specific probability distribution. These models enable precise estimation and interpretation of survival probabilities, hazard rates, and other survival characteristics [36,37]. Commonly applied distributions include exponential, Weibull, log–normal, and log–logistic, each tailored to distinct survival patterns.

3.1. Core Components of Parametric Survival Models

The survival function, $S(t)$, quantifies the probability of surviving beyond a specific time t. It is mathematically expressed as

$$S(t) = P(T > t), \tag{6}$$

and serves as the foundation for understanding the likelihood of survival over time. The hazard function, $h(t)$, describes the instantaneous failure rate at time t, conditioned on prior survival. It is defined as

$$h(t) = \frac{f(t)}{S(t)}, \tag{7}$$

where $f(t)$ represents the probability density function of survival times. This function provides insights into the risk of event occurrence at any given time. The cumulative hazard function, $H(t)$, represents the accumulated risk of failure over time and is defined as

$$H(t) = \int_0^t h(u)\, du = -\ln(S(t)). \tag{8}$$

This measure links the hazard and survival functions, highlighting the overall risk of failure across time.

3.2. Choice of Parametric Models

The selection of a parametric distribution depends on the observed survival data's characteristics. The exponential model assumes a constant hazard rate, making it appropriate for scenarios where the failure risk remains steady over time. The Weibull model extends the exponential model by enabling the hazard rate to either increase or decrease, offering a flexible approach to modeling varying patterns of risk [38]. The log–normal and log–logistic models are suited for data with skewed survival times or situations where the hazard rate declines after peaking. These models effectively capture variability in survival distributions [39]. Choosing the appropriate distribution enables a more accurate and nuanced understanding of survival behavior and associated risk factors.

4. Flowchart of the Correlation-Based IBMA Algorithm

Hybridized Method: Correlation-Based IBMA

This section outlines a structured framework to tackle the challenges of high-dimensional microarray survival data, where the number of predictors (genes) often far exceeds the sample size. The imbalance introduces issues such as overfitting, multicollinearity, and instability in model estimates, requiring a systematic approach to dimensionality reduction and survival modeling.

The process begins with a filtering step that employs correlation analysis to identify genes significantly associated with survival outcomes. Correlations between gene expression levels and the cumulative hazard function are computed using a 99% cutpoint. Filtering methods like this have been shown to enhance interpretability and improve model performance in microarray studies [40].

Next, a wrapper method using an Iterative Bayesian Model Averaging (IBMA) approach is applied to refine the feature set further. This method evaluates subsets of predictors based on their contributions to survival prediction, integrating over all possible models to account for model uncertainty. Bayesian wrappers are effective in reducing dimensionality while maintaining predictive accuracy in high-dimensional datasets [41].

A critical decision point follows to determine whether the number of selected predictors is less than or equal to the sample size. This ensures that the downstream survival

models are feasible. If the number of genes remains too large, the process loops back to the wrapper for further refinement. This iterative approach, aligned with the recommendations in [42], addresses overfitting risks and ensures stable parameter estimates.

Once dimensionality is reduced, two survival models are employed: the Cox Proportional Hazards (Cox-PH) Model and parametric models. The Cox-PH Model is semi-parametric, requiring no assumptions regarding baseline hazard distributions, making it widely applicable [32]. In contrast, parametric models assume specific distributions for survival times, offering more precise estimates when the assumptions hold [43].

Finally, the framework compares the performance of the Cox-PH and parametric models using metrics such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) [44]. Lower AIC/BIC values indicate superior model performance. This systematic comparison aligns with evidence that model selection based on these criteria improves robustness and interpretability [45,46].

By integrating filtering, Bayesian feature selection, and survival modeling, this framework addresses the challenges of high-dimensional data, providing a robust approach to identifying biomarkers associated with survival outcomes. The iterative nature ensures that dimensionality is appropriately managed, resulting in models that are both accurate and interpretable. However, the choice in p-value thresholds is often arbitrary. If the threshold is too strict, it may exclude genuinely important features. Conversely, if it is too lenient, it might include irrelevant features, leading to noisy results. To mitigate these limitations, our model utilizes 99% power, ensuring that the results are robust and less dependent on arbitrary p-value thresholds. On the other hand, reducing maxNvar (see Algorithm 3) below 30 is necessary to prevent convergence errors due to matrix singularity and instability. In this study, a value of 25 maxNvar was chosen for the DLBCL dataset as it balances accuracy and stability, as was similarly used by Annest et al. [35]. The flowchart for our proposed Correlation-Based IBMA algorithm is presented in Figure 1. The code for reproducing the results of the Correlation-Based IBMA algorithm is available at https://github.com/Dydx1989/Correlation-Based-IBMA (accessed on 14 February 2025).

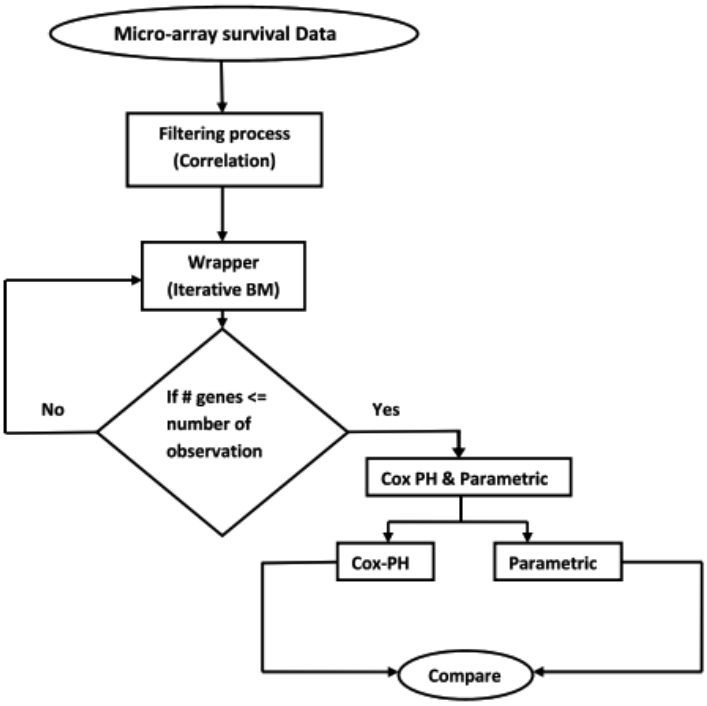


Figure 1. Flowchart of the Correlation-Based IBMA algorithm.

5. Results

5.1. Real-Life Studies

The dataset (DLBCL) undergoing R-CHOP treatment was obtained from the R repository GEOquery, a repository for microarray data. These samples were collected from surgically resected specimens in colorectal cancer patients [47], containing survival data of n = 181 patients who have diffuse large B-cell lymphoma, and d = 3835 different gene expression measurements are included.

5.2. Procedure 1 Results: Filtering Method (Correlation-Based)

Table 1 presents a comparison of genes associated with the cumulative hazard function based on varying alpha levels used in a filtering correlation technique.

Table 1. Comparison between genes associated with cumulative hazard function based on varying alpha levels.

α	0.001	0.01	0.02	0.03	0.04	0.1	0.2	0.3	0.4	0.5
Gene selected	27	89	158	208	247	522	907	1287	1647	2032

As the alpha value increases, the number of selected genes also increases. Specifically, at an alpha level of 0.001, 27 genes were selected. When the alpha level was raised to 0.01, the selection increased to 89 genes. At an alpha of 0.02, 158 genes were identified, and, at 0.03, the number rose to 208 genes. This upward trend continues with further increases in alpha, resulting in the selection of 247 genes at 0.04, 522 genes at 0.1, 907 genes at 0.2, 1287 genes at 0.3, 1647 genes at 0.4, and 2032 genes at 0.5.

The line plot in Figure 2 reveals the performance of the model at various levels of alpha. It is evident that, as the alpha value increases, the number of genes selected increases too, and vice versa. This affirms that the model is at its best when the alpha value is small. In this case, the model performs best, i.e., selected the minimum number (27) of genes, at α = 0.001.



Figure 2. Line plot showing the filtering method's performance and gene selection across various alpha levels.

The most common approach to presenting gene expression data is to present a heatmap, which may also be paired with cluster dendrograms. The heatmap's goal is to discover genes that are over- or underexpressed, as well as biological signatures that are connected with a certain state (i.e., illness condition). The data in the heatmap in Figure 3 are of the 3835 genes in the dataset, where each row represents a sample and each

column represents a gene. The color and intensity of the boxes represent changes in gene expression: shades of blue and green indicate elevated expression (i.e., highly significant to the sample), shades of purple indicate decreased expression (i.e., zero significance to the sample), and blue indicates unchanged expression (i.e., genes with scale measures of zero). A correlated heatmap in Figure 4 was created to illustrate the connections between these chosen variables (genes). Another approach is to present the genomic data in a correlated heatmap. The color and intensity of the boxes represent the degree of correlation regarding gene expression data: shades of blue and green indicate elevated correlation (i.e., highly significant to the sample), shades of purple indicate slightly weak correlation (i.e., little or no significance to the sample), and blue indicates unchanged correlation (i.e., genes with high correlation). The summary statistics in Table 2 provide a detailed overview of the results for the 27 selected genes based on the filtering methods.

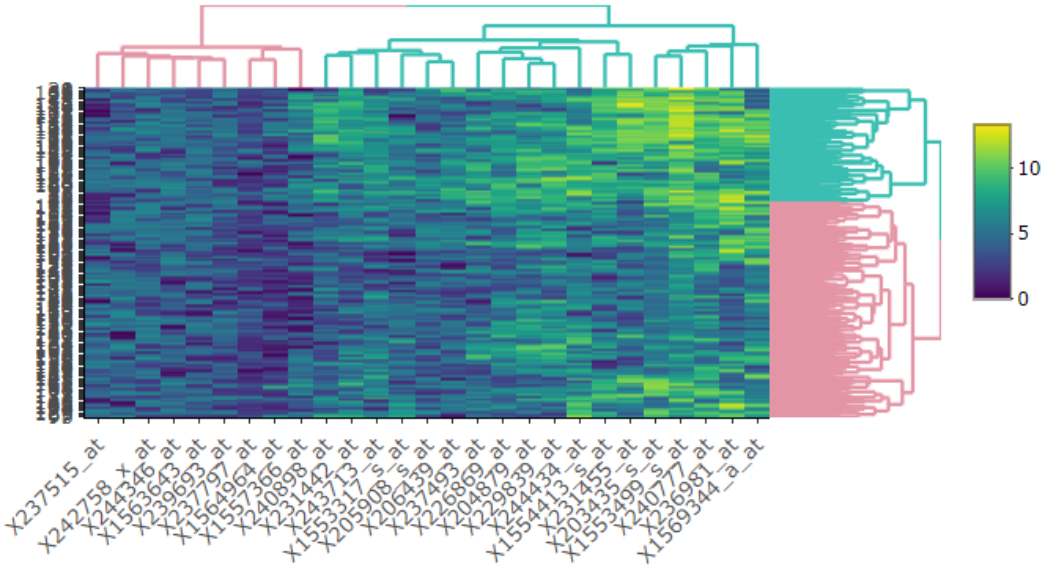


Figure 3. The variable importance plot of the 27 extracted gene features from DLBCL data using the filtering method.

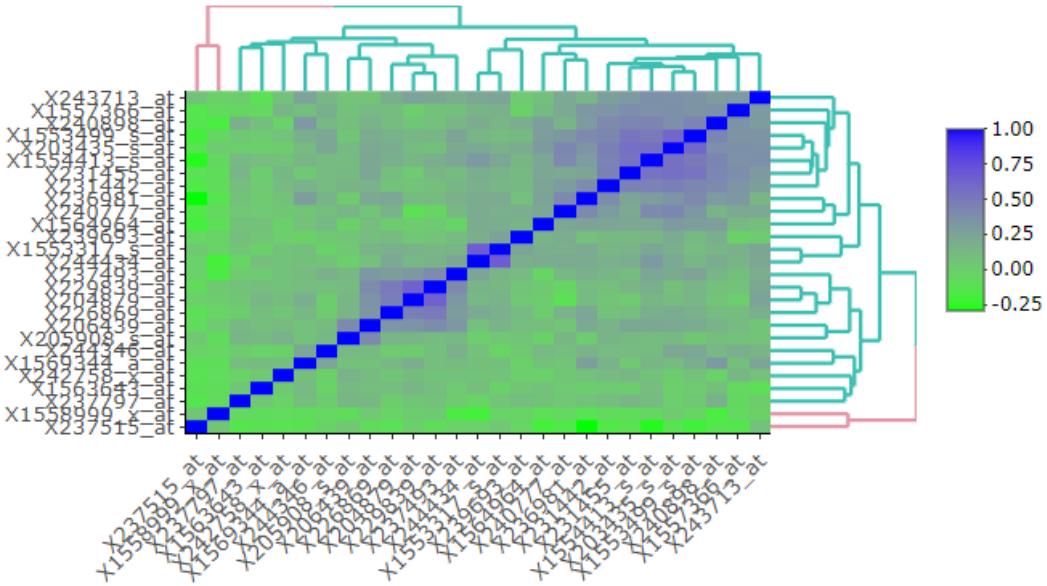


Figure 4. The correlated heatmap plot of the 27 extracted gene features from DLBCL data using the filtering method.

Table 2. Performance metrics for each of the 27 selected genes via filtering method.

Gene Name	Correlation	Statistic	Std. Error	*p*-Value	Confidence Interval
X240898_at	0.3237640	4.578262	0.090331525	8.749838×10^{-6}	[0.1867240, 0.4484446]
X229839_at	0.3173771	4.477716	0.069626215	1.340571×10^{-5}	[0.1798446, 0.4427396]
X1553317_s_at	0.2956320	4.140353	0.064923470	5.334807×10^{-5}	[0.1565236, 0.4232396]
X240777_at	0.2843484	3.968123	0.044254417	1.047141×10^{-4}	[0.1444830, 0.4130742]
X237797_at	0.2835054	3.955328	0.085518890	1.100011×10^{-4}	[0.1435851, 0.4123134]
X1569344_a_at	0.2823278	3.937473	0.054291651	1.178045×10^{-4}	[0.1423312, 0.4112504]
X244434_at	0.2812136	3.920596	0.021532875	1.256621×10^{-4}	[0.1411452, 0.4102443]
X1553499_s_at	0.2788330	3.884597	0.014299265	1.441196×10^{-4}	[0.1386127, 0.4080937]
X243713_at	0.2724357	3.788235	0.078522591	2.070259×10^{-4}	[0.1318159, 0.4023070]
X231442_at	0.2679058	3.720332	0.071828843	2.661317×10^{-4}	[0.1270112, 0.3982033]
X226869_at	0.2594559	3.594373	0.014263358	4.202434×10^{-4}	[0.1180662, 0.3905344]
X205908_s_at	0.2592865	3.591856	0.010320057	4.240451×10^{-4}	[0.1178871, 0.3903804]
X237493_at	0.2578275	3.570201	0.049895640	4.581279×10^{-4}	[0.1163450, 0.3890544]
X236981_at	0.2538053	3.510638	0.003152094	5.656506×10^{-4}	[0.1120973, 0.3853959]
X1554413_s_at	0.2530921	3.500097	0.013363995	5.869903×10^{-4}	[0.1113447, 0.3847467]
X237515_at	−0.2523931	−3.489771	0.072919927	6.086257×10^{-4}	[−0.3841104, −0.1106071]
X1563643_at	0.2514562	3.475941	0.085027354	6.387770×10^{-4}	[0.1096189, 0.3832573]
X1557366_at	0.2513477	3.474340	0.031633671	6.423559×10^{-4}	[0.1095045, 0.3831585]
X239693_at	0.2513261	3.474022	0.046496654	6.430702×10^{-4}	[0.1094817, 0.3831388]
X1564964_at	0.2485053	3.432449	0.006858373	7.429787×10^{-4}	[0.1065081, 0.3805688]
X206439_at	0.2459578	3.394983	0.080566498	8.452993×10^{-4}	[0.1038247, 0.3782460]
X203435_s_at	0.2454648	3.387742	0.006054556	8.665362×10^{-4}	[0.1033057, 0.3777964]
X231455_at	0.2453459	3.385995	0.005362892	8.717351×10^{-4}	[0.1031805, 0.3776878]
X204879_at	0.2449448	3.380106	0.012096357	8.894732×10^{-4}	[0.1027583, 0.3773219]
X244346_at	0.2443666	3.371620	0.066907309	9.156259×10^{-4}	[0.1021497, 0.3767943]
X242758_x_at	0.2437652	3.362797	0.090331525	9.435776×10^{-4}	[0.1015169, 0.3762455]
X1558999_x_at	−0.2436319	−3.360842	0.014299265	9.498769×10^{-4}	[−0.3761239, −0.1013766]

5.3. Procedure 2 Results: Wrapper Method (IBMA)

Table 3 shows the results of applying the Iterative Bayesian Model Averaging (IBMA) technique, a wrapper method, to reduce dimensionality by selecting key genes. Across various numbers of models (ranging from five to forty), the method consistently identified 15 genes. This stability in gene selection indicates that increasing the number of models did not lead to the selection of more genes as IBMA continually settled on the same 15 variables.

Table 3. Results of dimensionality reduction using the Iterative Bayesian Model Averaging (IBMA) technique across varying models.

Number of Models	5	10	15	20	25	30	35	40
Genes selected	15	15	15	15	15	15	15	15

Table 4 summarizes the number of features (genes) selected from a high-dimensional survival dataset of diffuse large B-cell lymphoma (DLBCL). The dataset contains 3835 covariates, and the analysis was performed on a sample of 181 individuals.

Using the filtering method, 27 genes were selected from the original set of 3835 covariates. This method reduces the large number of variables by applying specific criteria to identify a subset of relevant features for further analysis.

After applying the filtering process, the Iterative Bayesian Model Averaging (Iterative BMA) method identified 15 genes. This technique employs Bayesian principles to systematically select those features that are most influential in predicting survival outcomes, emphasizing simplicity and significance in the selection process. Table 5 provides the

results of 15 selected genes via the Iterative Bayesian Model Averaging (IBMA) approach, summarizing the key metrics of each gene's effect on the outcome.

Table 4. Number of selected features (genes) from the DLBCL high-dimensional survival data.

Method	Number of Genes Selected	Total Number of Covariates
Filtering	27	3835
Iterative BMA	15	3835
Sample Size	181	

Table 5. Number of selected 15 genes via Iterative BMA.

Selected Gene	Coef. Posterior	Posterior Probability %
*X*1558999_*x*_*at*	0.2455597	100
*X*229839_*at*	−0.1786791	100
*X*237797_*at*	−0.2836618	100
*X*244346_*at*	0.1580137	100
*X*1553317_*s*_*at*	−0.1855569	99
*X*237515_*at*	0.1580137	95.9
*X*1563643_*at*	−0.1518949	90.3
*X*242758_*x*_*a*	−0.1339099	80.2
*X*243713_*at*	−0.0886892	66.4
*X*1569344_*a*_*at*	−0.0302844	31.6
*X*237493_*at*	−0.0290551	26.0
*X*240777_*at*	−0.0214521	24.8
*X*1557366_*at*	−0.0183163	20.4
*X*205908_*s*_*at*	−0.0033933	3.9
*X*244434_*at*	−0.0006353	1.4

Table 5 outlines the details of the 15 genes selected through Iterative Bayesian Model Averaging (Iterative BMA) from a high-dimensional survival dataset. For each gene, the posterior coefficient and posterior probability percentage are provided.

The posterior coefficient represents the estimated influence of each gene on the survival outcome. Positive coefficients, such as for *X*1558999_*x*_*at* (0.2455597), indicate a positive relationship with survival, whereas negative coefficients, such as for *X*229839_*at* (−0.1786791), reflect a negative relationship. The magnitude of the coefficient signifies the strength of this influence.

The posterior probability percentage indicates the likelihood of a gene being included in the predictive model. Genes like *X*1558999_*x*_*at*, *X*229839_*at*, and *X*237797_*at* have posterior probabilities of 100%, showing strong evidence of their relevance in predicting survival outcomes. In contrast, genes such as *X*244434_*at* (1.4%) and *X*205908_*s*_*at* (3.9%) have low probabilities, suggesting limited importance in the model.

The variable importance of the 27 selected genes, as presented in Figure 5, specifically underscores their contribution to predicting the survival outcomes of patients with diffuse large B-cell lymphoma (DLBCL).

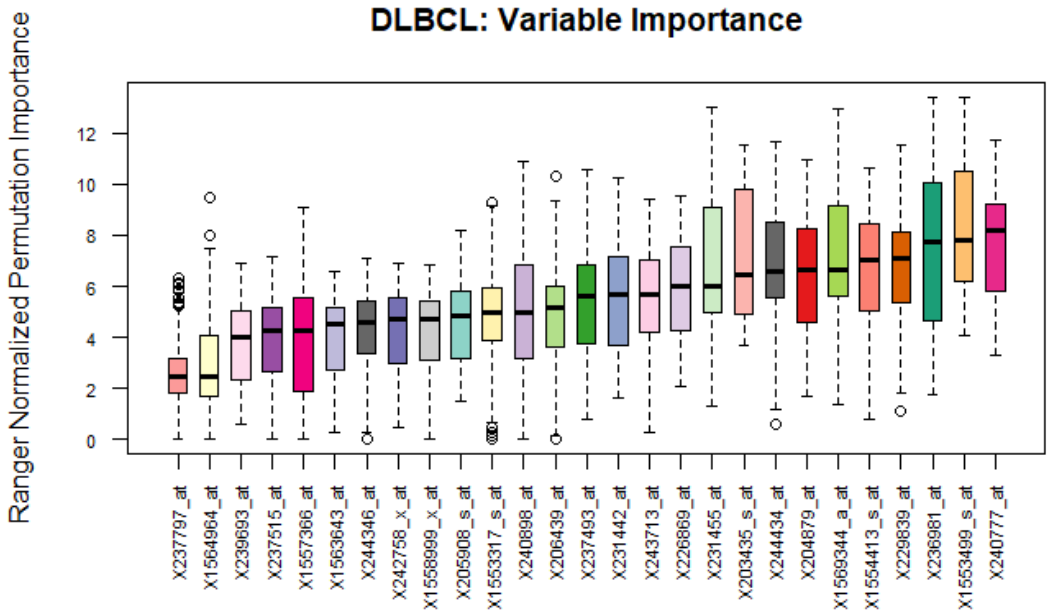


Figure 5. The variable importance plot for the 27 extracted gene features from DLBCL data.

5.4. Comparison Between the Existing (IBMA) and Proposed (Correlation-Based IBMA) Methods

In this section, we investigate and present the predictive power of the proposed method (Correlation-Based IBMA) over the existing method (IBMA).

The results in Table 6 compare the performance of these methods using three key metrics: accuracy, True Positive Rate (TPR), and True Negative Rate (TNN). The Correlation-Based IBMA method shows a notable improvement across all the performance metrics, demonstrating its predictive power and effectiveness over the existing IBMA method.

Table 6. Performance comparison of the existing method (IBMA) and Correlation-Based IBMA.

Methods	Accuracy (%)	TPR (%)	TNN (%)
IBMA	87.27	90.32	83.33
Correlation-Based IBMA	**94.55**	**94.44**	**94.74**

Bold values indicate the best performance in each column.

5.5. Comparison with the Parametric Methods

In assessing the influence of the selected genes on patient survival in DLBCL, we begin by considering parametric survival models [48]. While these models are valuable in scenarios where distributional assumptions hold, their application may not be suitable for many real-world datasets. Therefore, we propose an alternative approach by using a semi-parametric model, specifically the Cox Proportional Hazards (Cox-PH) Model [32].

These chosen models are suitable for DLBCL survival data as they reflect the influence of gene biomarkers on patient survival. The results from these models will guide clinical decision-making by identifying biomarkers that influence DLBCL cancer, enabling the provision of appropriate treatment for patients.

Table 7 summarizes the comparison of different parametric models applied to the dataset using AIC and BIC as performance measures. Lower values of both AIC and BIC indicate better model fit.

Based on this comparison in Table 7, the Weibull model has the smallest AIC (70.1692) and BIC (124.5437), indicating that it provides the best fit to the data. In contrast, the exponential model has the highest AIC (159.9177) and BIC (211.0937) values, making it the least suitable model.

Table 7. Comparison within the parametric methods.

Model	AIC	BIC
Weibull	70.1692	124.5437
Exponential	159.9177	211.0937
Log–normal	133.0065	187.3810
Log–logistic	110.5485	164.9230

Table 8 below presents the comparison between a parametric model (Weibull) and a semi-parametric model (Cox-PH) using AIC and BIC to assess model performance. Lower values indicate a better model fit.

Table 8. Comparison between the parametric method and semi-parametric method.

Model	AIC	BIC
Parametric (Weibull)	70.1692	124.5437
Semi-parametric (Cox-PH)	1450.9680	1498.9450

The parametric Weibull model shows far lower AIC (70.1692) and BIC (124.5437) values compared to the semi-parametric Cox-PH Model, indicating that it provides a better fit to the data. Table 2 summarizes the correlation coefficients, P-values, test statistics, and confidence intervals for 27 genes. The correlation values range from moderate-positive (e.g., $r = 0.3238$ for *X*240898_*at*) to negative (e.g., $r = -0.2524$ for *X*237515_*at*), reflecting varying strengths and directions of associations between the gene expressions. The majority of the genes display P-values below the threshold of 0.001, highlighting statistically significant associations. For instance, *X*240898_*at* demonstrates strong significance ($P = 8.75 \times 10^{-6}$), with a confidence interval of $[0.1867, 0.4484]$, underscoring the robustness of this result. The confidence intervals further affirm the reliability of these findings. Negative correlations, such as *X*1558999_*x*_*at* ($r = -0.2436$), have confidence intervals that exclude zero ($[-0.3761, -0.1014]$), strengthening the evidence of inverse associations. These results emphasize the effectiveness of Iterative BMA in identifying significant genes and characterizing their correlations, offering valuable insights with potential biological and clinical implications.

The tables presenting the results from the Iterative Bayesian Model Averaging (BMA), Cox Proportional Hazards Model, and Weibull model provide insights into the significance of covariates (genes) across different statistical approaches, which can be interpreted through their *p*-values.

In the Iterative BMA table, genes are selected based on their posterior probabilities. A high posterior probability, such as 100% for genes like *X*1558999_*x*_*at*, *X*229839_*at*, and *X*237797_*at*, suggests that these genes are highly likely to be relevant for survival prediction. However, while posterior probabilities indicate the likelihood of relevance, they do not directly reflect statistical significance. Therefore, a gene with a high posterior probability might not necessarily exhibit a low *p*-value in the survival models (Cox and Weibull).

The *p*-values in the Cox model are crucial for assessing the statistical significance of each gene's effect on survival. For example, genes such as *X*1558999_*x*_*at* and *X*229839_*at* have *p*-values below 0.05, indicating that these genes have a significant impact on survival. In contrast, genes like *X*1569344_*a*_*at* and *X*244434_*at* show higher *p*-values (e.g., 0.171501 and 0.796420), suggesting that they do not significantly influence survival. When comparing the posterior probabilities from the Iterative BMA table with the *p*-values from the Cox model, it is evident that, although some genes are selected by BMA with

high posterior probabilities, they may not all have significant *p*-values in the Cox model (e.g., *X*1569344_*a*_*at*).

The Weibull model also uses *p*-values to determine the significance of each gene's association with survival. Low *p*-values, such as for *X*1558999_*x*_*at*, *X*229839_*at*, and *X*237797_*at*, suggest that these genes have a strong and statistically significant relationship with survival. Genes with higher *p*-values (e.g., *X*1569344_*a*_*at* and *X*244434_*at*) indicate that their effect on survival is weaker or insignificant. Comparing the *p*-values from both the Weibull and Cox models, we observe that certain genes, like *X*1558999_*x*_*at* and *X*229839_*at*, are significant in both models, highlighting their importance. However, some genes that were significant in the Weibull model (e.g., *X*237797_*at*) may show weaker significance in the Cox model, as demonstrated with *X*237797_*at* having a *p*-value of 0.000324 in the Cox model.

Genes such as *X*1558999_*x*_*at*, *X*229839_*at*, and *X*237797_*at* appear in both the Iterative BMA selection and exhibit low *p*-values in both the Cox and Weibull models. This consistency suggests that these genes are strongly associated with survival and are robust across different statistical approaches.

On the other hand, genes like *X*1569344_*a*_*at* and *X*244434_*at*, despite being selected in the Iterative BMA process, show higher *p*-values in the survival models, indicating that, while they may have been considered relevant during feature selection, their impact on survival is not statistically significant.

The analysis from Table 5 revealed that the following genes, with 100% posterior probabilities—*X*1558999_*x*_*at*, *X*229839_*at*, and *X*237797_*at*—which were consistently selected by the Correlation-Based IBMA algorithm, were also found to be statistically significant in both the Cox-PH and Weibull models' extended results presented in Tables 9 and 10. Thus, these genes demonstrated strong predictive potential and maintained their importance across different statistical approaches, reinforcing their relevance for predicting the survival of patients with DLBCL cancer.

Table 9. Extended results of the Cox-PH Model based on the most important genes selected by Correlation-Based IBMA.

Variable	Coef.	Exp(Coef.)	SE(Coef.)	z	$Pr(>\|z\|)$
*X*1558999_*x*_*at*	0.14290	1.15361	0.04934	2.896	**0.003776**
*X*229839_*at*	−0.12788	0.87995	0.04506	−2.838	**0.004540**
*X*1553317_*s*_*at*	−0.14269	0.86703	0.05751	−2.481	**0.013093**
*X*240777_*at*	−0.08048	0.92267	0.03849	−2.091	**0.036534**
*X*237797_*at*	−0.22207	0.80086	0.06177	−3.595	**0.000324**
*X*1569344_*a*_*at*	−0.04794	0.95319	0.03506	−1.367	0.171501
*X*244434_*at*	−0.01388	0.98621	0.05381	−0.258	0.796420
*X*242758_*x*_*at*	−0.14690	0.86338	0.05369	−2.736	**0.006219**
*X*243713_*at*	−0.06560	0.93651	0.04399	−1.491	0.135934
*X*1557366_*at*	−0.02867	0.97174	0.03960	−0.724	0.469128
*X*237515_*at*	0.06234	1.06433	0.04464	1.397	0.162518
*X*205908_*s*_*at*	−.06816	0.93411	0.05448	−1.251	0.210914
*X*237493_*at*	−0.03084	0.96963	0.04571	−0.675	0.499840
*X*244346_*at*	−0.15775	0.85406	0.05714	−2.761	**0.005765**
*X*1563643_*at*	−0.14897	0.86159	0.04811	−3.096	**0.001960**

Bold values indicate significant genes.

Although some genes, such as *X*1569344_*a*_*at* and *X*244434_*at*, were selected by the Correlation-Based IBMA algorithm but did not exhibit statistically significant *p*-values in the survival models, they still contribute to the power of the Correlation-Based IBMA algorithm by narrowing down the pool of candidate genes. Even if their direct effects are not immediately evident, these genes may play a role in a broader biological context,

further highlighting the value of the Correlation-Based IBMA algorithm in identifying potentially important genes.

Table 10. Extended results of the Weibull model based on the most important genes selected by Correlation-Based IBMA.

Variable	Value	Std. Error	z	*p*-Value
(Intercept)	-2.64×10^{0}	3.54×10^{-1}	-7.46	$\mathbf{8.4 \times 10^{-14}}$
*X*1558999_*x*_*at*	-5.47×10^{-2}	2.40×10^{-2}	-2.28	$\mathbf{2.26 \times 10^{-2}}$
*X*229839_*at*	5.65×10^{-2}	2.18×10^{-2}	2.59	$\mathbf{9.6 \times 10^{-3}}$
*X*1553317_*s*_*at*	6.88×10^{-2}	2.90×10^{-2}	2.37	$\mathbf{1.76 \times 10^{-2}}$
*X*240777_*at*	3.14×10^{-2}	1.87×10^{-2}	1.67	9.43×10^{-2}
*X*237797_*at*	8.45×10^{-2}	2.97×10^{-2}	2.85	$\mathbf{4.4 \times 10^{-3}}$
*X*1569344_*a*_*at*	2.15×10^{-2}	1.79×10^{-2}	1.20	2.292×10^{-1}
*X*244434_*at*	1.81×10^{-5}	2.66×10^{-2}	0.00	9.995×10^{-1}
*X*242758_*x*_*at*	3.77×10^{-2}	2.60×10^{-2}	1.45	1.474×10^{-1}
*X*243713_*at*	2.98×10^{-2}	2.25×10^{-2}	1.32	1.864×10^{-1}
*X*1557366_*at*	1.85×10^{-2}	1.99×10^{-2}	0.93	3.526×10^{-1}
*X*237515_*at*	-4.82×10^{-2}	2.30×10^{-2}	-2.10	$\mathbf{3.58 \times 10^{-2}}$
*X*205908_*s*_*at*	3.52×10^{-2}	2.67×10^{-2}	1.32	1.875×10^{-1}
*X*237493_*at*	1.54×10^{-2}	2.27×10^{-2}	0.68	4.960×10^{-1}
*X*244346_*at*	5.83×10^{-2}	2.70×10^{-2}	2.16	$\mathbf{3.10 \times 10^{-2}}$
*X*1563643_*at*	7.34×10^{-2}	2.46×10^{-2}	2.98	$\mathbf{2.9 \times 10^{-3}}$
Log(scale)	-6.80×10^{-1}	6.18×10^{-2}	-11.00	$< 2 \times 10^{-16}$

Bold values indicate significant genes.

6. Conclusions

This study highlights the effectiveness of the proposed Correlation-Based IBMA algorithm in enhancing predictive accuracy within high-dimensional microarray datasets. Through the use of filtering and wrapping techniques, Correlation-Based IBMA efficiently isolates a small subset of significant predictor genes, simplifying complex datasets and making it a cost-effective and practical tool for clinical diagnostics.

The filtering process ensures that only those predictors that meet predefined selection criteria are included, thereby improving model precision by excluding less impactful or redundant variables based on correlation-based statistical measures. On the other hand, the Iterative Bayesian Model Averaging (IBMA) method, derived from machine learning techniques used for classification tasks, leverages BMA's multivariate capabilities to address model uncertainty. This integration makes it particularly effective for extracting predictive genes from high-dimensional biological datasets.

Interestingly, the proposed Correlation-Based IBMA algorithm demonstrates greater stability in gene selection compared to regularization methods such as COX-LASSO. By repeatedly applying the algorithm with the same threshold, the gene selection process yields consistent and reliable results. This stability underscores the robustness and reliability of the Correlation-Based IBMA algorithm as a method for gene selection in comparative studies, offering notable advantages in accuracy and reproducibility over regularization techniques.

Moreover, we compared the results of parametric Weibull and semi-parametric Cox-PH Models using the 15 selected genes. The results revealed that the parametric Weibull model outperformed the semi-parametric Cox-PH Model, as evidenced by its lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values. Although the Weibull model demonstrated a better fit to the DLBCL data, the same number of significant genes were identified by both models, with only one difference in the specific significant genes (see Tables 9 and 10).

Finally, the performance of the Correlation-Based IBMA approach was demonstrated using diffuse large B-cell lymphoma (DLBCL) cancer data. However, it is important to validate the findings across other types of cancers and datasets. Testing the Correlation-Based IBMA method on a variety of cancer types (e.g., breast cancer and lung cancer) and datasets (e.g., RNA-Seq and clinical trial data) would help to ensure its generalizability and robustness in predicting patient survival across diverse conditions.

Author Contributions: Conceptualization, K.A.D.; methodology, K.A.D. and R.K.L.; writing—original draft preparation, K.A.D.; writing R.K.L. and K.A.D.; software, K.A.D. and R.K.L.; formal analysis, R.K.L. and K.A.D.; validation, R.K.L. All authors have read and agreed to the published version of the manuscript.

Funding: Kazeem A. Dauda was supported by the Trond Mohn Foundation (project HyperEvol under grant agreement no. TMS2021TMT09), through the Centre for Antimicrobial Resistance in Western Norway (CAMRIA) (TMS2020TMT11).

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used are available in published sources.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Correlation-Based IBMA	Correlation-Based Approach and Iterative Bayesian Model Averaging (IBMA)
Cox-PH	Cox Proportional Hazard
DLBCL	Diffuse Large B-Cell Lymphoma
BMA	Bayesian Model Averaging
HGP-FS	Hybrid Genetic-Particle Swarm Feature Selection
ACO	Ant Colony Optimization
LMFS	Large-Margin Hybrid Feature Selection Algorithm
HDRA	Hybrid Dimension Reduction Algorithm
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion

References

1. Chan, J.Y.; Somasundaram, N.; Grigoropoulos, N.; Lim, F.; Poon, M.L.; Jeyasekharan, A.; Yeoh, K.W.; Tan, D.; Lenz, G.; Ong, C.K.; et al. Evolving therapeutic landscape of diffuse large B-cell lymphoma: challenges and aspirations. *Discov. Oncol.* **2023**, *14*, 132. [CrossRef] [PubMed]
2. Avalos, A.M.; Meyer-Wentrup, F.; Ploegh, H.L. B-cell receptor signaling in lymphoid malignancies and autoimmunity. *Adv. Immunol.* **2014**, *123*, 1–49. [PubMed]
3. Shi, Y.; Xu, Y.; Shen, H.; Jin, J.; Tong, H.; Xie, W. Advances in biology, diagnosis and treatment of DLBCL. *Ann. Hematol.* **2024**, *103*, 3315–3334. [CrossRef]
4. Harkins, R.A.; Chang, A.; Patel, S.P.; Lee, M.J.; Goldstein, J.S.; Merdan, S.; Flowers, C.R.; Koff, J.L. Remaining challenges in predicting patient outcomes for diffuse large B-cell lymphoma. *Expert Rev. Hematol.* **2019**, *12*, 959–973. [CrossRef]
5. He, R.; Oliveira, J.L.; Hoyer, J.D.; Viswanatha, D.S. Molecular hematopathology. In *Hematopathology*; Elsevier: Amsterdam, The Netherlands, 2018; pp. 712–760.
6. Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G.; et al. Next-generation sequencing technology: Current trends and advancements. *Biology* **2023**, *12*, 997. [CrossRef] [PubMed]
7. Rahnenführer, J.; De Bin, R.; Benner, A.; Ambrogi, F.; Lusa, L.; Boulesteix, A.L.; Migliavacca, E.; Binder, H.; Michiels, S.; Sauerbrei, W.; et al. Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges. *BMC Med.* **2023**, *21*, 182. [CrossRef] [PubMed]
8. Dauda, K.A.; Adeniyi, E.J.; Lamidi, R.K.; Wahab, O.T. Exploring Flexible Penalization of Bayesian Survival Analysis Using Beta Process Prior for Baseline Hazard. *Computation* **2025**, *13*, 21. [CrossRef]

9. Ansari, G.; Ahmad, T.; Doja, M.N. Hybrid filter–wrapper feature selection method for sentiment classification. *Arab. J. Sci. Eng.* **2019**, *44*, 9191–9208. [CrossRef]
10. Annest, A.; Yeung, W.K.Y. *iterativeBMAsurv: The Iterative Bayesian Model Averaging (BMA) Algorithm for Survival Analysis*, R package version 1.60.0; Bioconductor: Seattle, WA, USA, 2023. [CrossRef]
11. Raftery, A.; Hoeting, J.; Volinsky, C.; Painter, I.; Yeung, K.Y. *BMA: Bayesian Model Averaging*, R package version 3.18.19; CRAN: Vienna, Austria, 2024 .
12. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian Model Averaging: A Tutorial (with Comments by M. Clyde, D. Draper, and E.I. George, and a Rejoinder by the Authors). *Stat. Sci.* **1999**, *14*, 382–417. [CrossRef]
13. Atmakuru, A.; Di Fatta, G.; Nicosia, G.; Badii, A. Improved Filter-Based Feature Selection Using Correlation and Clustering Techniques. In Proceedings of the International Conference on Machine Learning, Optimization, and Data Science, Grasmere, UK, 22–26 September 2023; pp. 379–389.
14. Borah, K.; Das, H.S.; Seth, S.; Mallick, K.; Rahaman, Z.; Mallik, S. A review on advancements in feature selection and feature extraction for high-dimensional NGS data analysis. *Funct. Integr. Genom.* **2024**, *24*, 139. [CrossRef] [PubMed]
15. Parvaiz, A.; Nasir, E.S.; Fraz, M.M. From Pixels to Prognosis: A Survey on AI-Driven Cancer Patient Survival Prediction Using Digital Histology Images. *J. Imaging Inform. Med.* **2024**, *37*, 1728–1751. [CrossRef]
16. Moslehi, F.; Haeri, A. A novel hybrid wrapper–filter approach based on genetic algorithm particle swarm optimization for feature subset selection. *J. Ambient Intell. Humaniz. Comput.* **2020**, *11*, 1105–1127. [CrossRef]
17. Abiodun, E.O.; Alabdulatif, A.; Abiodun, O.I.; Alawida, M.; Alabdulatif, A.; Alkhawaldeh, R.S. A systematic review of emerging feature selection optimization methods for optimal text classification: The present state and prospective opportunities. *Neural Comput. Appl.* **2021**, *33*, 15091–15118. [CrossRef]
18. Shukla, A.K.; Singh, P.; Vardhan, M. A New Hybrid Feature Subset Selection Framework Based on Binary Genetic Algorithm and Information Theory. *Int. J. Comput. Intell. Appl.* **2019**, *18*, 1950020. [CrossRef]
19. Ghosh, M.; Guha, R.; Sarkar, R.; Abraham, A. A wrapper-filter feature selection technique based on ant colony optimization. *Neural Comput. Appl.* **2020**, *32*, 7839–7857. [CrossRef]
20. Hu, Z.; Bao, Y.; Xiong, T.; Chiong, R. Hybrid filter–wrapper feature selection for short-term load forecasting. *Eng. Appl. Artif. Intell.* **2015**, *40*, 17–27. [CrossRef]
21. Zhang, J.; Xiong, Y.; Min, S. A new hybrid filter/wrapper algorithm for feature selection in classification. *Anal. Chim. Acta* **2019**, *1080*, 43–54. [CrossRef]
22. Hammami, M.; Bechikh, S.; Hung, C.C.; Ben Said, L. A multi-objective hybrid filter-wrapper evolutionary approach for feature selection. *Memetic Comput.* **2019**, *11*, 193–208. [CrossRef]
23. Dauda, K.A.; Olorede, K.O.; Aderoju, S.A. A novel hybrid dimension reduction technique for efficient selection of bio-marker genes and prediction of heart failure status of patients. *Sci. Afr.* **2021**, *12*, e00778. [CrossRef]
24. Hancer, E. Differential evolution for feature selection: A fuzzy wrapper–filter approach. *Soft Comput.* **2019**, *23*, 5233–5248. [CrossRef]
25. Johnson, N.A.; Slack, G.W.; Savage, K.J.; Connors, J.M.; Ben-Neriah, S.; Rogic, S.; Scott, D.W.; Tan, K.L.; Steidl, C.; Sehn, L.H.; et al. Concurrent Expression of MYC and BCL2 in Diffuse Large B-Cell Lymphoma Treated With Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone. *J. Clin. Oncol.* **2012**, *30*, 3452–3459. [CrossRef] [PubMed]
26. Dauda, K.A.; Olorede, K.O.; Banjoko, A.W.; Yahya, W.B.; Ayipo, Y.O. Genetic Diagnosis, Classification, and Risk Prediction in Cancer Using Next-Generation Sequencing in Oncology. In *Computational Approaches in Biomaterials and Biomedical Engineering Applications*; CRC Press: Boca Raton, FL, USA, 2024; pp. 107–122.
27. McDermott, J.E.; Wang, J.; Mitchell, H.; Webb-Robertson, B.J.; Hafen, R.; Ramey, J.; Rodland, K.D. Challenges in biomarker discovery: Combining expert insights with statistical analysis of complex omics data. *Expert Opin. Med Diagn.* **2013**, *7*, 37–51. [CrossRef]
28. Banegas-Luna, A.J.; Peña-García, J.; Iftene, A.; Guadagni, F.; Ferroni, P.; Scarpato, N.; Zanzotto, F.M.; Bueno-Crespo, A.; Pérez-Sánchez, H. Towards the interpretability of machine learning predictions for medical applications targeting personalised therapies: A cancer case survey. *Int. J. Mol. Sci.* **2021**, *22*, 4394. [CrossRef] [PubMed]
29. Xu, X.; Qi, Z.; Han, X.; Wang, Y.; Yu, M.; Geng, Z. Combined-task deep network based on LassoNet feature selection for predicting the comorbidities of acute coronary syndrome. *Comput. Biol. Med.* **2024**, *170*, 107992. [CrossRef] [PubMed]
30. Dauda, K.A.; Yahya, W.B.; Banjoko, A.W. Survival analysis with multivariate adaptive regression splines using Cox-Snell residual. *Ann. Comput. Sci. Ser.* **2015**, *13*, 25–41.
31. Aalen, O. Nonparametric Inference for a Family of Counting Processes. *Ann. Stat.* **1978**, *6*, 701–726. Available online: http://www.jstor.org/stable/2958850 (accessed on 14 February 2025). [CrossRef]
32. Cox, D.R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B* **1972**, *34*, 187–202. [CrossRef]
33. Dauda, K.A.; Pradhan, B.; Shankar, B.U.; Mitra, S. Decision tree for modeling survival data with competing risks. *Biocybern. Biomed. Eng.* **2019**, *39*, 697–708. [CrossRef]

34. Dudoit, S.; Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol.* **2002**, *3*, research0036.1. [CrossRef]
35. Annest, A.; Bumgarner, R.E.; Raftery, A.E.; Yeung, K.Y. Iterative bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinform.* **2009**, *10*, 72. [CrossRef]
36. Dauda, K.A. Optimal tuning of random survival forest hyperparameter with an application to liver disease. *Malays. J. Med. Sci. MJMS* **2022**, *29*, 67. [CrossRef] [PubMed]
37. Haredasht, F.N.; Dauda, K.A.; Vens, C. Exploiting censored information in self-training for time-to-event prediction. *IEEE Access* **2023**, *11*, 96831–96840. [CrossRef]
38. Klein, J.P.; Moeschberger, M.L. *Survival Analysis: Techniques for Censored and Truncated Data*; Springer: New York, NY, USA, 2003.
39. Hosmer, D.W., Jr.; Lemeshow, S.; May, S. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*; Wiley: Hoboken, NJ, USA, 2008.
40. Ein-Dor, L.; Kela, I.; Getz, G.; Givol, D.; Domany, E. Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics* **2005**, *21*, 171–178. [CrossRef]
41. Raftery, A.E.; Madigan, D.; Hoeting, J.A. Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **1997**, *92*, 179–191. [CrossRef]
42. Harrell, F.E., Jr.; Lee, K.L.; Mark, D.B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **1996**, *15*, 361–387. [CrossRef]
43. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*; John Wiley & Sons: Hoboken, NJ, USA, 2002.
44. Chakrabarti, A.; Ghosh, J.K. AIC, BIC and recent advances in model selection. *Philos. Stat.* **2011**, *7*, 583–605.
45. Hafner, L.; Walsh, L. Application of multi-method-multi-model inference to radiation related solid cancer excess risks models for astronaut risk assessment. *Z. Für Med. Phys.* **2024**, *34*, 83–91. [CrossRef] [PubMed]
46. Dauda, K.A.; Lamidi, R.K.; Dauda, A.A.; Yahya, W.B. A New Generalized Gamma-Weibull Distribution with Applications to Time-to-event Data. *bioRxiv* **2023**. [CrossRef]
47. Lenz, G.; Wright, G.; Dave, S.; Xiao, W.; Powell, J.; Zhao, H.; Xu, W.; Tan, B.; Goldschmidt, N.; Iqbal, J.; et al. Stromal gene signatures in large-B-cell lymphomas. *N. Engl. J. Med.* **2008**, *359*, 2313–2323. [CrossRef]
48. Lawless, J.F. Parametric Models in Survival Analysis. In *Wiley StatsRef: Statistics Reference Online*; Balakrishnan, N., Colton, T., Everitt, B., Piegorsch, W., Ruggeri, F., Teugels, J.L., Eds.; Wiley: Hoboken, NJ, USA, 2014. [CrossRef]

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.