JOURNAL OF STATISTICAL MODELING & ANALYTICS (JOSMA)
(ISSN: 2180-3102)
UNIVERSITI MALAYA

# Modeling Colon Cancer Survival using a Proportional Hazard Mixture Cure Model with Principal Component Covariates

Haruna Suleiman[1,2], Noraslinda Mohamed Ismail[1*], Shariffah Suhaila Syed Jamaludin[1]

[1] *Department of Mathematics, Faculty of Science, University Teknologi Malaysia (UTM), 81310 Johor Bahru, Malaysia*

[2] *Department of Statistics, School of Applied Sciences, Nuhu Bamalli Polytechnic, Zaria, Kaduna State, 810282, Nigeria*

*Corresponding author: noraslinda@utm.my*

Received 10 February 2025
Accepted 22 May 2025
Published 29 May 2025

RESEARCH ARTICLE

**Abstract**

This study explores how gene expression data can help predict the survival times of colon cancer patients. Since the dataset is high-dimensional, Principal Component Analysis (PCA) reduces complexity while retaining essential information. Based on eigenvalue one criteria, proportion of variance accounted for, and scree plot analysis, 60 principal components (PCs) are selected as covariates. These are then used in a Proportional Hazard Mixture Cure Model, applying both Cox and Weibull as baseline models to differentiate between cured and uncured patients over a five-year follow-up period. Maximum Likelihood Estimation (MLE) is applied to estimate the model parameters. The results show that the Cox model provides more reliable estimates, indicated by lower AIC values, higher hazard rates, and statistically significant p-values (<0.05). On the other hand, the Weibull model finds no significant covariates (p-values >0.05), with only the intercept being significant. Furthermore, the Weibull model estimates a 100% cure rate, while the Cox model estimates 56%, suggesting that the Cox model provides a better fit for predicting survival outcomes. By integrating gene expression data into survival modeling, this study offers a more accurate and interpretable way to understand patient outcomes. The findings highlight the Cox mixture cure model as a valuable tool for guiding clinical decisions.

**Keywords:** Cox model, Mixture cure fraction model, Principal components, Proportional hazard model, Weibull model

## 1. Introduction

Survival analysis is crucial in understanding the prognostic factors affecting cancer patients' outcomes. Gene expression data offers a rich source of information but poses challenges due to its high dimensionality. Colorectal cancer is the third most common cancer worldwide with an estimated death of 930,000 in 2020 (Morgan et al., 2023), yet one of the cancer cases involving high dimensional microarray data in molecular and genetic biology, where the sample size is usually in the hundreds, with tens of thousands of genes (Al-Thanoon et al., 2018);(Algamal et al., 2018). Multicollinearity and overfitting are significant challenges when applying statistical classification and feature selection

methods to high-dimensional datasets, consequently, making it trivial in model selection and classification. An unsupervised machine learning algorithm approach called principal component analysis (PCA) is one of the techniques that can be employed to measure the dimensionality of covariates. Though many studies used PCA in high dimensional microarray data (Lenz et al., 2016);(Razzaque & Badholia, 2024). Its application while putting into cognizance the three major approaches (Scree plot, Eigenvalue-one-criterion and proportion of variance accounted for) is essential. The main idea of using PCA is to reduce the dimensionality of data whose variables are interrelated (Gewers et al., 2022) into a set of uncorrelated variables that successively maximize variability (i.e. statistical information) without losing information from the datasets.

Early cancer detection using gene expression data is crucial for quality patient care. Accurate data analysis is essential to avoid misdiagnosis and its associated risks. The high dimensionality of gene expression datasets, with numerous features per gene, requires substantial computational resources and can introduce multicollinearity issues (Das et al., 2024). Due to the importance of these issues, efficient and effective techniques are required to improve the classification accuracy and the selection of a small subset of genes appropriately. However, the PCA effectively addresses multicollinearity by creating orthogonal variables that capture most data variance. The Eigenvalue-one–criterion method of the PCA significantly helps assess multicollinearity, with values greater than 1 suggesting the retention of principal components (Kyriazos & Poga, 2023). Moreover, reducing the gene expression to a significant dimension will lead to an easy discovery and diagnosis of patients affected with the cancer disease (Hossain et al., 2019); (Ding et al., 2021).

The event of interest in this study is clearly defined as death from colon cancer. This means the study focuses on observing and analyzing the time from diagnosis to the occurrence of death specifically caused by colon cancer. Patients who do not experience this event within the 5-year follow-up period are considered censored, meaning their exact survival time is unknown, but it is known to exceed the duration of their follow-up. By modeling this event, the study aims to predict survival probabilities and survival times using gene expression data as covariates. Both cured patients (those who survive beyond 5 years without death from colon cancer) and uncured patients (those who experience the event) are accounted for in the analysis using mixture cure models.

Furthermore, after properly selecting the relevant covariates for the estimate, the next is to model the survival probability by an appropriate model that could better fit the dataset, accurately representing a relationship between the covariates and the survival time of the colon cancer patients. In this research, a maximum likelihood estimation MLE technique was later employed to maximize the parameters of the statistical models. Recently, (Badisy et al., 2023) evaluated Morocco's colorectal patients' overall survival rates at 3 years and, using a novel method that combined survival random forest with the Cox model, revealed strong predictive indicators. Moreover, (Atinafu et al., 2020) assessed the survival status and predictors of mortality among colorectal cancer patients using Kaplan–Meier analysis with a log-rank test and bivariate and multivariable analysis through the Cox proportional hazard model. Moreover, (Xie et al., 2024)used univariate, Lasso, and multivariate Cox regression models to create an immune-related lncRNA signature, followed by constructing a nomogram in R to predict survival in colorectal cancer patients. Furthermore, (Bai et al., 2020) applied univariate Cox regression analysis to examine the association of immune-related genes with the prognosis in patients with colorectal cancer.

However, despite extensive research on additive risk and Cox models, linear regression survival models for high-dimensional microarray data, particularly for right-censored data, are still rare. This study proposes using Cox (semi-parametric) and Weibull (parametric) models as baselines to model the survival time and probability of colon cancer patients, where the event of interest is death. In this research, the event of interest is death, specifically among individuals who succumbed to colon cancer after being diagnosed with the disease. In contrast, the patients who do not experience this event within the 5-year follow-up period are considered censored. These models will be integrated with a mixture

cure fraction model to account for cured patients. The study aims to identify the best methodology for modeling survival time by comparing these approaches, focusing on dimensionality reduction, and applying advanced survival models using gene expression data as covariates.

## 2. Materials and methods

### 2.1 Data collected and Filtering

The dataset comprises gene expression profiles of 2,000 genes from a heterogeneous sample of 62 colon cancer patients, originally collected by (Alon et al., 1999). Among these patients, 40 had tumor tissues, while 22 did not, whose samples were from non-tumorous regions of the colon in the cancer patients, making it a complete time-to-event data. The data was sourced from the Microarray Databases site compiled by the Princeton University Gene Expression Project (2002), from patients with a comprehensive medical report and diagnosed with colon cancer. The data is normalized afterward by scaling and filtering using a variance threshold of 0.5, removing any column with zero variance, and performing Principal component analysis to reduce the dimension of the data set using the 'princomp ()' functions in R studio software. The diagnosis and the general survival time were not explicitly mentioned in the data, therefore the research adopted a simulation-based technique, like that of (Bender et al., 2005); (Infante et al., 2023); (Rutter et al., 2023)to simulate a sufficient survival time of the patients for a follow-up period of 5 years, which is from 0 times to 5 years, which is approximately 60 months.

In contrast to (Alon et al., 1999), The survival time in this study is defined as the time from diagnosis to death from colon cancer. The study specifically models the survival probabilities of colon cancer patients using gene expression data as covariates, with the event of interest being death due to colon cancer. Moreover, the research specifically indicates right-censored observations, which account for the subjects who did not experience the event of interest, that is death from colon cancer. In this research, patients who do not experience the event of interest (death from colon cancer) or are lost to follow-up within 5 years are considered censored, indicating that their exact survival times are unknown but exceed the observed follow-up duration.

In this study, the colon tissue samples from 62 patients (Alon et al., 1999) were analyzed, guided by the principles of the Central Limit Theorem (CLT) (Abraham De Moivre, 1733). The CLT states that, regardless of the original data distribution, the sampling distribution of the sample mean (or other statistics) will approximate a normal distribution when the sample size is moderately large, typically between 30 and 50, even if the data is slightly skewed. This property enables reliable statistical inference for population-level insights. Although the sample size is sufficient under this framework, dimensionality reduction is crucial to ensure model stability, prevent overfitting, and derive meaningful statistical conclusions.

### 2.2 Dimension Reduction Using PCA

In this research, the principal component analysis technique (Pearson 1901) is used to reduce the dimensionality of the colon gene expression data while preserving as much variability as possible. The original correlated variables are transformed into a set of uncorrelated variables that are a linear combination of the original gene expression, called principal components (PCs), serving as covariates. Sixty (60) PCs were retained to account for most of the variation from the 2000 gene expressions of 62 patients.

## 2.3 Performing PCA analysis

**Data presentation**

The gene expression data in Matrix ***X*** form (patient's *X* gene expression):

$$X = \begin{bmatrix} Gene1 & Gene2 & \dots & Geneg \\ x_{11} & x_{12} & \dots & x_{1g} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{62,1} & x_{62,2} & \dots & x_{62,2000} \end{bmatrix} \tag{2.1}$$

The survival data is;

$$\begin{bmatrix} patient & Time & Event \\ 1 & t_1 & \delta_1 \\ 2 & t_2 & \delta_2 \\ . & . & . \\ . & . & . \\ N & t_N & \delta_N \end{bmatrix} \tag{2.2}$$

**Data Standardization**

Standardization to make each feature (gene) have zero mean and unit variance,

a. Mean Vector Calculation**:**

$$\mu_i = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \tag{2.3}$$

where $n$ is the number of samples which is 62 in this case, and $X_{ij}$ is the expression level of gene $j$ and sample $i$.

b. We subtract the mean from each element in the corresponding column:

$$X_{centered,ij} = X_{ij} - \mu_j \tag{2.4}$$

$$X_{Centered} = X_{ij} - \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 \\ . & . & . \\ . & . & . \\ . & . & . \\ \mu_{2000} & \mu_{2000} & \mu_{2000} \end{bmatrix} \tag{2.5}$$

c. Standardize the data:

$$X_{standardized,ij} = \frac{X_{centered,ij}}{\sigma_j} \tag{2.6}$$

d. we compute the standard deviation for each gene expression as follows:

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_{ij} - \mu_j\right)^2} \tag{2.7}$$

**Variance-covariance**

Let $X$ be the 62 × 2000 matrix of the standardized gene expression dataset and compute the variance matric $C$ as follows:

$$C = \frac{1}{n-1} X_{standardized}^{T} X_{standardized} \tag{2.8}$$

where $n$ is the 62 samples of the gene expression of 2000 dimension.

- Transpose of the centered Matrix $X_{centered}^{T}$

$$X_{centered}^{T} = \begin{bmatrix} x_{11} - \mu_1 & \cdots & x_{62,1} - \mu_1 \\ \vdots & \vdots & \vdots \\ x_{1,2000} - \mu_{2000} & \cdots & x_{62,2000} - \mu_{2000} \end{bmatrix} \tag{2.9}$$

- Matrix multiplication

$$X_{centered}^{T} X_{centered} = \begin{bmatrix} \sum_{i=1}^{62}(x_{i1} - \mu_1)^2 & \cdots & \sum_{i=1}^{62}(x_{i1} - \mu_1)(x_{i,2000} - \mu_{2000}) \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{62}(x_{i,2000} - \mu_{2000})(x_{i1} - \mu_1) & \cdots & \sum_{i=1}^{62}(x_{i,2000} - \mu_{2000})^2 \end{bmatrix} \tag{2.10}$$

- Normalization by $n - 1$:

$$C = \frac{1}{61}\begin{bmatrix} \sum_{i=1}^{62}(x_{i1} - \mu_1)^2 & \cdots & \sum_{i=1}^{62}(x_{i1} - \mu_1)(x_{i,2000} - \mu_{2000}) \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{62}(x_{i,2000} - \mu_{2000})(x_{i1} - \mu_1) & \cdots & \sum_{i=1}^{62}(x_{i,2000} - \mu_{2000})^2 \end{bmatrix} \tag{2.11}$$

**Eigenvalues and Eigenvectors for the covariance Matric *C*:**

$$CV = \lambda V \tag{2.12}$$

which expands to

$$\begin{bmatrix} C_{11} & \dots & C_{1,2000} \\ . & . & . \\ . & . & . \\ . & . & . \\ C_{2000,1} & \dots & C_{2000,2000} \end{bmatrix} \begin{bmatrix} V_1 \\ . \\ . \\ . \\ V_{2000} \end{bmatrix} = \lambda \begin{bmatrix} V_1 \\ . \\ . \\ . \\ V_{2000} \end{bmatrix} \tag{2.13}$$

**Principal Components scores:**

$$Z = X_{standardized} V \tag{2.14}$$

which expands to

$$Z = \begin{bmatrix} x'_{11} & x'_{12} & \dots & x'_{1,2000} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ x'_{62,1} & x'_{62,2} & \dots & x'_{62,2000} \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1,2000} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ v_{2000,1} & v_{2000,2} & \dots & v_{2000,2000} \end{bmatrix} \tag{2.15}$$

where *V* here is the eigenvectors for the 2000 dimension. The first element $Z_{11}$ is computed by the row and column variables as follows

$$Z_{11} = x'_{11}.v_{11} + x'_{21}.v_{21} + \dots + x'_{1,2000}.v_{2000,1} \tag{2.16}$$

Similarly, the computation continues for the other elements for each row and column. Hence, the element $z_{ij}$ of matrix Z is computed as:

$$Z_{ij} = \sum_{k=1}^{2000} X'_{ik}.V_{kj} \tag{2.17}$$

For the entire principal components, each entry in Z is a dot product between the $i-th$ row of the $X_{standardized}$ and the $j-th$ column of *V*. However, it is generally not feasible to do PCA manually, in this research the 'princomp' function in R studio software was used to compute the principal components.

## 2.4 Cox Proportional Hazard Mixture Cure Fraction Model

The Cox proportional hazards mixture cure fraction model is used to analyze the survival data of the colon patient with the genes retained as the covariates (PCs) when a fraction of the population is assumed to be cured, meaning they will not experience the event of interest (death). This model incorporates the survival function for uncured patients and the probability of being cured.

## 2.5 Model Formulation for the Cox Proportional Hazard Mixture Cure Fraction Model

- **Derivation with the PCs as the covariates**

Let $X = (PC_1, PC_2, \ldots\ldots, PC_{60})^T$ be the vector of 60 PC derived from the gene expression data. We incorporate the principal components (PCs) into the model as covariates. The Cox proportional hazard function for the uncured patients can be written as:

$$h(t / pc_1, pc_2, \ldots, pc_{60}) = h_o(t) \exp\left(\sum_{i=1}^{60} \beta_i pc_i\right) \tag{2.18}$$

where $h_0(t)$ is the baseline hazard function of the event over time, independent of the patient-specific covariates, *X is* the vector of the principal components ($PC_i$), that is the covariates, and $\beta_i$ is the vector of the regression coefficients.

The Cure fraction with logistic regression is given by the following expression:

$$\pi(X) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{60} \gamma_i pc_i\right)} \tag{2.19}$$

$\pi(X)$ is the probability of being cured, $\gamma_i$ is the vector of the coefficient for the cure fraction model.

Survival function for uncured patients is given by:

$$S(t / X) = \exp(-H_0(t) \exp\left(\sum_{i=1}^{60} \beta_i PC_i\right) \tag{2.20}$$

$S\left(\frac{t}{X}\right)$ represents the probability that the colon cancer patient will survive beyond time *t*, $H_0(t)$ the cumulative hazard when all *X* covariates are equal to zero, and $\beta_i$ is the vector of the regression coefficients corresponding to each covariate, while the overall survival function can be written as:

$$S(t / X) = \frac{1}{1 + \exp\left(-\sum_{i=1}^{60} \gamma_i pc_i\right)} + \left(1 + \frac{1}{1 + \exp\left(-\sum_{i=1}^{60} \gamma_i pc_i\right)}\right) \exp\left(-H_0(t) \exp\left(\sum_{i=1}^{60} \beta_i PC_i\right)\right) \tag{2.21}$$

where $h\left(\frac{t}{X}\right)$ describes how the hazard of colon cancer changes over time for the patients who are not cured, considering the effects of the principal components (covariates), $\pi(X)$ represents the probability of the patients being cured influenced by the covariates i.e. the gene expressions retained as the (PCs). The log-likelihood function combines contributions from uncensored (event) and censored observations for right-censored data. We let $\delta_i$ be the event indicator (1 if the event occurred, 0 if censored) for the $i - th$ individuals.

The log-likelihood contribution for the $i^{th}$ individuals is given by

$$LogL_i = \delta_i \log(ph(t_i / X_i) S_u(t_i / X_i)) + (1 - \delta_i) \log(PS_u(t_i / X_i) + (1 - p) \tag{2.22}$$

substitute the expression for $h\left(\frac{t}{X}\right)$ and $S\left(\frac{t}{X}\right)$:

$$LogL_i = \delta_i (\log p + \log h_0(t_i) + \sum_{i=1}^{60} \beta_j PC_{ij} - H_0(t) \exp\left(\sum_{j=1}^{60} \beta_j PC_{ij}\right) + (1 - \delta_i) \log\left(P \exp(-H_0(t) \exp\left(\sum_{j=1}^{60} \beta_j PC_{ij}\right) + (1 - P)\right) \tag{2.21}$$

The overall log-likelihood which is the sum of individual log-likelihood is shown below:

$$LogL = \sum_{i=1}^{n} \log L_i \tag{2.23}$$

becomes

$$LogL = \sum_{j=1}^{n}\left\{\delta_i\left(\log P + \log h_0(t_i)\right) + \sum_{i=1}^{60}\beta_j PC_{ij} - H_0(t)\exp\left(\sum_{j=1}^{60}\beta_j PC_{ij}\right) + (1-\delta_i)\log P\exp\left(-H_0(t)\right)\exp\left(\sum_{i=1}^{60}\beta_j PC_{ij}\right) + (1-p)\right\} \tag{2.24}$$

## 2.6 Model Formulation for the Weibull Proportional Hazard Mixture Cure Fraction Model

- **Derivation with the PCs as the covariates**

*Let* $X = (PC_1, PC_2, \ldots\ldots, PC_{60})^T$ be the vector of 60 PC derived from the gene expression data. We incorporate the principal components (PCs) into the model as covariates.
Weibull proportional hazard function for uncured patients is given by;

$$h(t/X) = h_0(t)\exp\left(\sum_{i=1}^{60}\beta_i PC_i\right) \tag{2.25}$$

where $h_0(t_i)$ is the baseline Weibull hazard function of the event over time, $X$ is the vector of the principal components ($PC_i$), that is the covariates, and $\beta_i$ is the vector of the regression coefficients.

$$h_0(t_i) = K\lambda t^{k-1} \tag{2.27}$$

where a shape parameter $K > 0$ and a scale parameter $\lambda > 0$ parameterize the Weibull distribution. Thus,

$$h(t/X) = K\lambda t^{k-1}\exp\left(\sum_{i=1}^{60}\beta_i PC_i\right) \tag{2.28}$$

is the cure fraction, $\pi(X)$, modeled using the logistic regression function as shown below:

$$\pi(X) = \frac{1}{1+\exp\left(-\sum_{i=1}^{60}\gamma_i PC_i\right)} \tag{2.29}$$

The survival function for uncured patients is:

$$S_u(t/X) = \exp(-H_0(t)\exp\sum_{i=1}^{60}\beta_i PC_i) \tag{2.30}$$

where

$$H_0(t) = \int_0^t (u)du = \lambda t^k \tag{2.31}$$

and

$$S(t/X) = \exp\left(-\lambda t^k \exp\left(\sum_{i=1}^{60}\beta_i PC_i\right)\right) \tag{2.32}$$

Therefore, the overall survival function, $S\left(\frac{t}{X}\right)$, becomes:

$$S(t/X) = \pi(X) + (1-\pi(X))S_u(t/X) \tag{2.33}$$

where $\pi(X)$ is the probability of being cured, $(1-\pi(X))S_u(t/X)$, is the survival function for the uncured individual colon patients. We substitute the expression for $\pi(X)$ and $S_u\left(\frac{t}{X}\right)$ becomes:

$$S(t/X) = \frac{1}{1+\exp\left(-\sum_{i=1}^{60}\gamma_i PC_i\right)} + \left(1 - \frac{1}{1+\exp\left(-\sum_{i=1}^{60}\gamma_i PC_i\right)}\exp(-\lambda t^k \exp\left(-\sum_{i=1}^{60}\gamma_i PC_i\right)\right) \tag{2.34}$$

The log-likelihood function combines contributions from uncensored (event) and censored observations for right-censored data. We let $\delta_i$ be the event indicator (1 if the event occurred, 0 if censored) for the $i-th$ individuals, and $t_i$ is the observed survival time of the patients. The log-likelihood contribution for the $i-th$ individual is:

$$LogL_i = \delta_i \log(ph(t_i/X_i)S_u(t_i/X_i)) + (1-\delta_i)\log(PS_u(t_i/X_i) + (1-p) \tag{2.35}$$

Substituting the expression for $h\left(\frac{t}{X}\right)$ and $S\left(\frac{t}{X}\right)$:

$$LogL_i = \delta_i \log(PK\lambda t^{k-1}\exp\left(\sum_{i=1}^{60}\beta_i PC_i\right)\exp(-\lambda t^k \exp\left(\sum_{i=1}^{60}\beta_j PC_{ij}\right)) + (1-\delta_i)\log(P\exp(-\lambda t^k \exp\left(\sum_{i=1}^{60}\beta_j PC_{ij}\right) + (1-p) \tag{2.36}$$

The overall log-likelihood is given by

$$LogL = \sum_{i=1}^{n}\log L_i,$$

and can also be written as

$$LogL = \sum_{i=1}^{n}\left\{\delta_i\left(\log P + \log K + \log\lambda + (K-1)\log t_i + \sum_{i=1}^{60}\beta_j PC_{ij} - \lambda t^k \exp\left(\sum_{i=1}^{60}\beta_j PC_{ij}\right)\right) + (1-\delta_i)\log\left(P\exp\left(\sum_{i=1}^{60}\beta_j PC_{ij}\right)\right) + (1-p)\right\} \tag{2.37}$$

## 3. Results

The following result, table 1 provides the principal components derived from the PCA, including their Eigenvalues, the percentage of variance they individually explained, and the cumulative percentage of the variance. The output suggests selecting the first 60 principal components considering the eigenvalue-one criterion approach. The approach indicates that retaining components with eigenvalues > 1 can replace the 2000 gene expression covariates with a reduced number of components while sacrificing only a negligible amount of information about the total variation in the system. The higher eigenvalues indicate that the component explains a larger proportion of the variance in the data. However, by the proportion of variance accounted for approach, we are to retain the PCs to the cumulative % of variances from 70 to 90%.

**Table 1. Proportion of Total Variance Explained by Each Principal Component**

| Component | Initial Eigenvalues (total) | % of variance | Cumulative % of the variance |
|---|---|---|---|
| 1 | 899.113 | 44.956 | 44.956 |
| 2 | 196.925 | 9.846 | 54.802 |
| 3 | 135.301 | 6.765 | 61.567 |
| 4 | 113.104 | 5.655 | 67.222 |
| 5 | 65.669 | 3.283 | 70.506 |
| 6 | 62.582 | 3.129 | 73.635 |
| 7 | 46.638 | 2.332 | 75.967 |
| 8 | 44.354 | 2.218 | 78.184 |
| 9 | 33.342 | 1.667 | 79.851 |
| 10 | 31.05 | 1.553 | 81.404 |
| 11 | 27.666 | 1.383 | 82.787 |
| 12 | 24.135 | 1.207 | 83.994 |
| 13 | 20.824 | 1.041 | 85.035 |
| 14 | 19.465 | 0.973 | 86.008 |
| 15 | 16.915 | 0.846 | 86.854 |
| 16 | 15.726 | 0.786 | 87.64 |
| 17 | 15.485 | 0.774 | 88.415 |
| 18 | 13.327 | 0.666 | 89.081 |
| 19 | 12.627 | 0.631 | 89.712 |
| 20 | 11.762 | 0.588 | 90.3 |
| 21 | 11.042 | 0.552 | 90.853 |
| 22 | 10.495 | 0.525 | 91.377 |
| 23 | 9.952 | 0.498 | 91.875 |
| 24 | 9.623 | 0.481 | 92.356 |
| 25 | 9.124 | 0.456 | 92.812 |
| 26 | 8.279 | 0.414 | 93.226 |
| 27 | 8.149 | 0.407 | 93.634 |
| 28 | 7.669 | 0.383 | 94.017 |
| 29 | 6.999 | 0.35 | 94.367 |
| 30 | 6.651 | 0.333 | 94.7 |
| 31 | 6.356 | 0.318 | 95.017 |
| 32 | 6.04 | 0.302 | 95.319 |
| 33 | 5.848 | 0.292 | 95.612 |
| 34 | 5.78 | 0.289 | 95.901 |
| 35 | 5.405 | 0.27 | 96.171 |
| 36 | 5.062 | 0.253 | 96.424 |
| 37 | 4.904 | 0.245 | 96.669 |
| 38 | 4.674 | 0.234 | 96.903 |

| | | | |
|---|---|---|---|
| 39 | 4.342 | 0.217 | 97.12 |
| 40 | 4.234 | 0.212 | 97.332 |
| 41 | 3.972 | 0.199 | 97.53 |
| 42 | 3.868 | 0.193 | 97.724 |
| 43 | 3.679 | 0.184 | 97.908 |
| 44 | 3.577 | 0.179 | 98.087 |
| 45 | 3.49 | 0.175 | 98.261 |
| 46 | 3.335 | 0.167 | 98.428 |
| 47 | 3.186 | 0.159 | 98.587 |
| 48 | 3.077 | 0.154 | 98.741 |
| 49 | 2.775 | 0.139 | 98.88 |
| 50 | 2.692 | 0.135 | 99.014 |
| 51 | 2.57 | 0.128 | 99.143 |
| 52 | 2.457 | 0.123 | 99.266 |
| 53 | 2.209 | 0.11 | 99.376 |
| 54 | 2.13 | 0.106 | 99.483 |
| 55 | 1.953 | 0.098 | 99.58 |
| 56 | 1.84 | 0.092 | 99.672 |
| 57 | 1.648 | 0.082 | 99.755 |
| 58 | 1.597 | 0.08 | 99.835 |
| 59 | 1.289 | 0.064 | 99.899 |
| 60 | 1.17 | 0.058 | 99.958 |

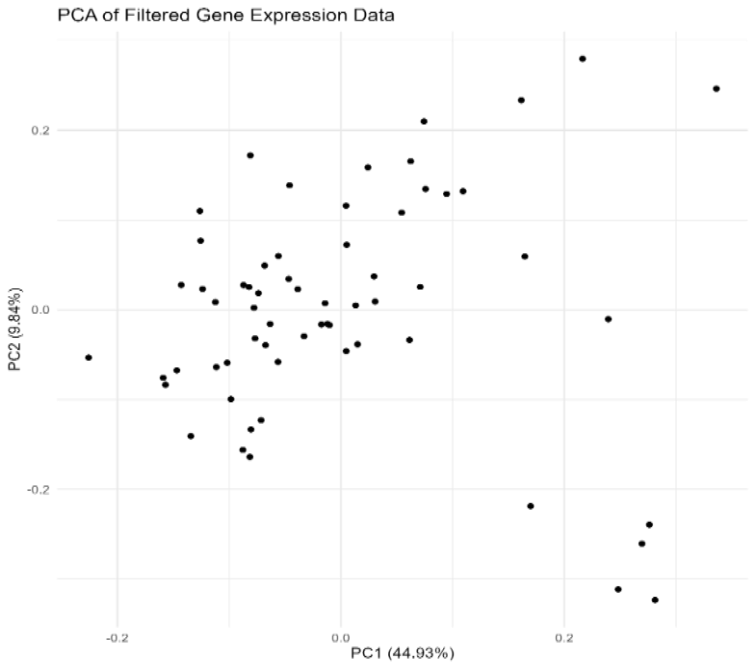
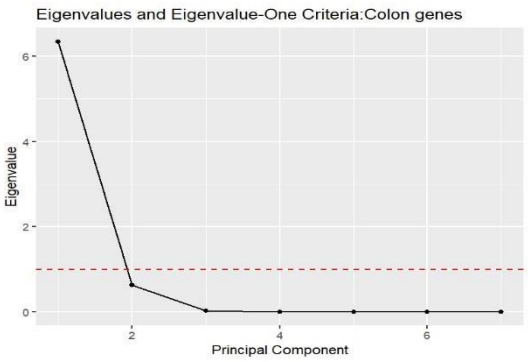
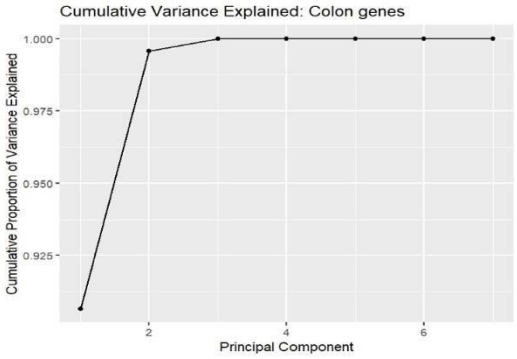



**Figure 1: Plot view of the Two-dimensional PCs for the gene expression data of the colon cancer patients**

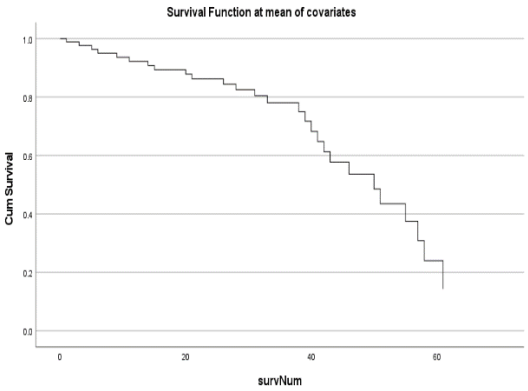
Figure 1 represents the visual image of the principal components plot with the x-axis representing the first principal component (PC1), which accounts for 44.93% of the total variance in the data. This suggests that PC1 captures a substantial portion of the variability among the gene expression samples. The second (PC2) follows accounting for 9.84% of the total variability capturing less variation than the (PC1). However, since the 2D plot suggests retention of too few PCs, we employ alternatives by exploration of the Scree-plot, Proportion of variance accounted for, and the Eigenvalue one criterion approaches to determine the number of PCs to be retained as our covariates for subsequent analysis.



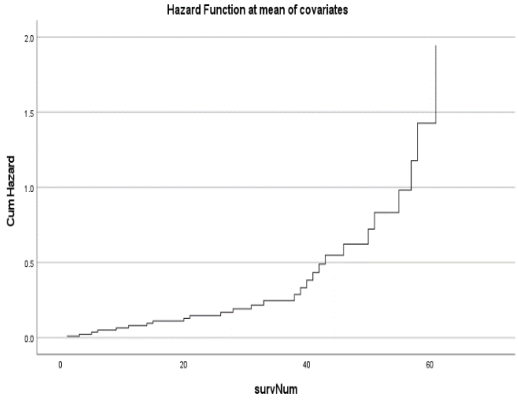

(i)



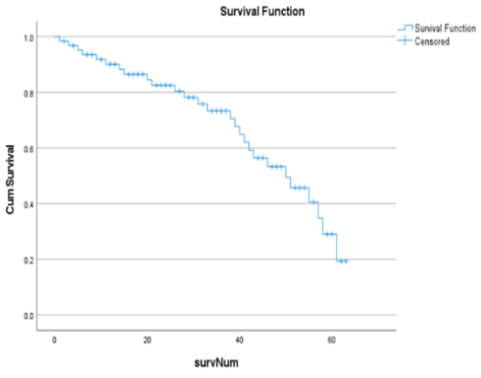

(ii)



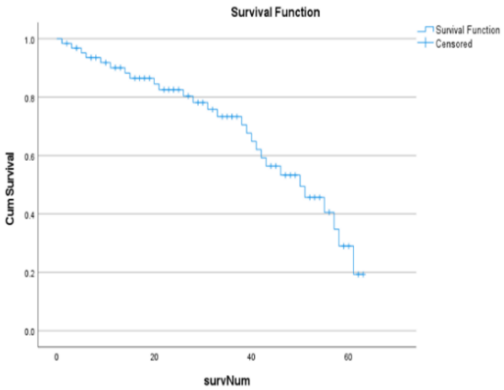

(iii)




(iv)




(v)




(vi)

**Figure 2. (i) The scree plot for the suggested principal components is to be retained from the Eigenvalues by the percentage of the variances. (ii) The scree plot for the suggested principal**

**components with the cumulative variance of 75.967 % is maintained according to the proportion of variance accounted for approach. (iii) The visual hazard plot of the colon cancer patients based on the mean values of the extracted principal components from gene expression data. (iv) The stepwise hazard function curve is based on the mean values from the principal components as covariates. (v) Survival function curve with censoring and covariates. (vi) Hazard function curve with covariates and censoring**

Figure 2 (i) represents the scree plot of the eigenvalue one criterion suggested by the research for a threshold of 75.967 cumulative percentage of the variance for the PCs to be retained by the proportion of variance accounted for, with the red-shaded line representing the eigenvalue of 1, indicating retention of all eigenvalues > 1 which is a common threshold used to determine the number of PCs to retain. This is meant to check for the elbow points on the plot and to visualize if the PCs capture a significant proportion of the variance at the 75.967% threshold. The Figure 2 (ii) scree plot suggests retaining a single PC as it breaks at a single PC for it retains a variability of 44.93% of the total variation and the proportion of variance accounted for suggests retaining 7 number of the PCs as covariates for modeling the colon cancer survival periods based on the results in Table 1, but are arbitrary, as a result, the approach has sometimes been criticized for its subjectivity [16]. Consequently, the research adopted the entire 60 PCs as the covariates retained by the Eigen-value-one criterion approach to avoid loss of vital information from the gene expression data that are significant to the colon cancer incidence and survival probability. Figure 2 (iii) and (iv) give the general survival rates and the hazard rates curves by the influence of the principal components (covariates) accounting for most of the variation from the original data source of 2000 dimensions. The Figure 2 (iv) curve illustrates the evolution of risk over time, showing the accumulated risk of the event occurring up to a specific time point. Moreover, Figure 2 (v) and (vi) curves represent the cumulative survival probability estimate for the patients beyond a specific time and the cumulative hazard probability estimate at different times throughout the follow-up period respectively with censoring throughout the period.

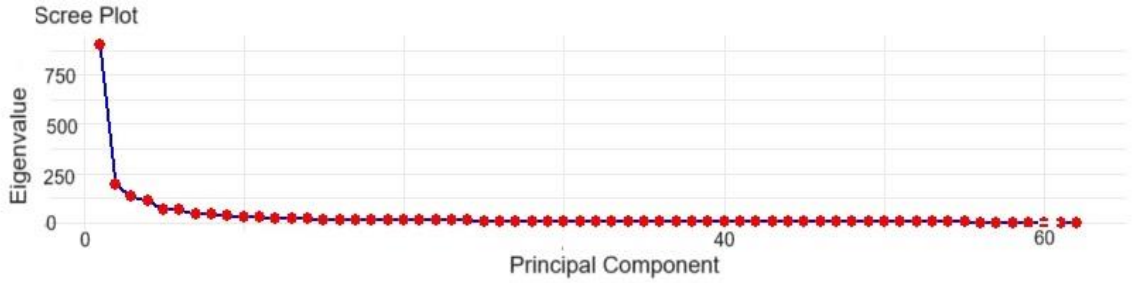



**Figure 3. Scree Plot of Principal Components Retained Using the Eigenvalue > 1 Criterion**

The scree plot in Figure 3 gives the graphical presentation of the eigenvalues > 1 retained from PCs to represent the entire gene expression of 2000 of the 62 colon cancer patients. The *x*-axis represents the total of 60 principal components accounting for the total variability and the y-axis represents the eigenvalues corresponding to each principal component.

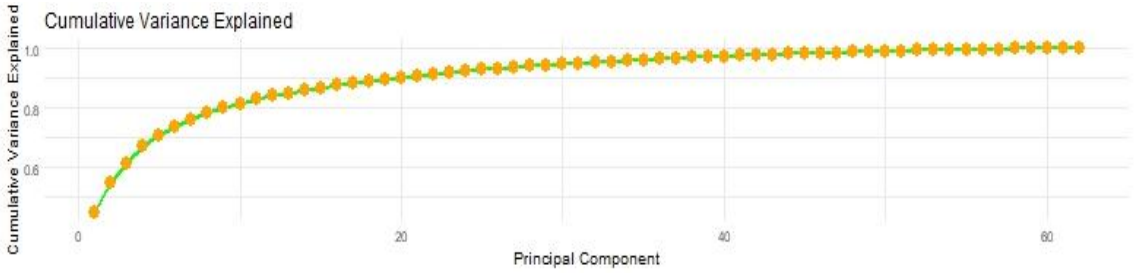



**Figure 4. Scree Plot Showing Cumulative Variance Explained by Principal Components**

The Figure 4 plot represents the cumulative variance explained by the principal components with the *x*-axis representing the principal components from 1 to 60 and the y-axis representing the cumulative proportion of total variance explained by the principal components, ranging from 0 to 1 (or 0% to 100%) from the gene expression colon cancer data.

**Table 2. Estimated PCs Associated with High Hazard Rates Using the Cox Proportional Hazard Model via MLE**

| Principal Component ($PC_i$) | Coefficient ($\beta_i$) | Hazard Ratio (exp(($\beta_i$)) | $p$-value ($P_i$) |
|---|---|---|---|
| PC1 | 15.16 | 3.84 X10$^6$ | <2e-16 |
| PC2 | 12.54 | 3.09X10$^5$ | <2e-16 |
| PC3 | 5.35 | 210.0 | <2e-16 |
| PC9 | 2.93 | 18.89 | 1.49e-11 |
| PC12 | 5.85 | 347.2 | <2e-16 |
| PC13 | 3.85 | 47.16 | <2e-16 |
| PC15 | 5.89 | 362.6 | <2e-16 |
| PC19 | 4.39 | 80.87 | <2e-16 |
| PC20 | 3.58 | 35.97 | <2e-16 |
| PC26 | 3.49 | 32.90 | <2e-16 |
| PC27 | 12.67 | 3.19X10$^5$ | <2e-16 |
| PC28 | 10.88 | 5.32X10$^4$ | <2e-16 |
| PC30 | 7.98 | 2933 | <2e-16 |
| PC31 | 17.37 | 3.48X10$^7$ | <2e-16 |
| PC34 | 28.00 | 1.45X0$^{12}$ | <2e-16 |
| PC36 | 8.87 | 7147 | <2e-16 |
| PC40 | 25.99 | 1.94X10$^{11}$ | <2e-16 |
| PC42 | 4.66 | 106.2 | <2e-16 |
| PC44 | 5.78 | 325.1 | <2e-16 |
| PC50 | 11.75 | 1.26X10$^5$ | <2e-16 |
| PC52 | 49.02 | 1.95X10$^{21}$ | <2e-16 |
| PC54 | 57.75 | 1.20X10$^{25}$ | <2e-16 |
| PC56 | 6.85 | 946.6 | <2e-16 |
| PC57 | 10.16 | 25950 | <2e-16 |

Table 2 shows the 24 PCs with exp ($\beta_i$) $\gg 1$, showing a stronger association between the PCs and the risk of the event from the estimated 60 PCs, indicating a significantly increased hazard.

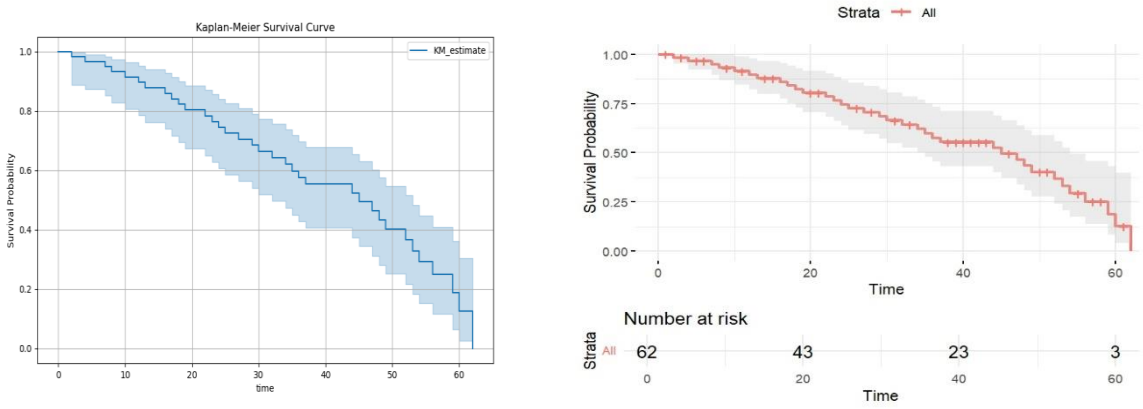
Moreover, explaining that a small increase in the specified PCs could lead to a large increase in the risk of the event (death) of the colon cancer patient. The smaller P-values of the PCs also indicate the highly significant association between them and the Hazard. Generally, the hazard ratios are substantial ($\gg 1$) indicating that for every unit increase in the $PC_i$ , the risk of the event (death) increases exactly times the number of the Hazard Ratio (exp(($\beta_i$)) values.

**Table 3. Estimated PCs Using the Weibull Proportional Hazard Model via MLE**

| Principal Components | Coefficient $\beta_i$ | Std. Error | *z*-value | *p*-value |
|---|---|---|---|---|
| **Intercept ($\beta_0$)** | 3.70750 | 0.14301 | 25.93 | <2e-16 |
| PC_1 | -0.09026 | 0.22943 | -0.39 | 0.69 |
| PC_2 | -0.22414 | 0.15785 | -1.42 | 0.16 |
| PC_3 | 0.07874 | 0.29745 | 0.26 | 0.79 |
| PC_4 | 0.16433 | 0.29499 | 0.56 | 0.58 |
| PC_5 | -0.01424 | 0.31675 | -0.04 | 0.96 |
| PC_6 | 0.22909 | 0.21067 | 1.09 | 0.28 |
| PC_7 | -0.24532 | 0.22299 | -1.10 | 0.27 |
| PC_8 | 0.07265 | 0.32247 | 0.23 | 0.82 |
| PC_9 | -0.16821 | 0.36802 | -0.46 | 0.65 |
| PC_10 | 0.01931 | 0.35086 | 0.06 | 0.96 |
| PC_11 | -0.02939 | 0.28513 | -0.10 | 0.92 |
| PC_12 | -0.09068 | 0.58775 | -0.15 | 0.88 |
| PC_13 | -0.13917 | 0.40794 | -0.34 | 0.73 |
| PC_14 | -0.03323 | 0.43157 | -0.08 | 0.94 |
| PC_15 | 0.02454 | 0.13908 | 0.18 | 0.86 |
| PC_16 | -0.00111 | 0.43003 | 0.00 | 1.00 |
| PC_17 | 0.24757 | 0.20933 | 1.18 | 0.24 |
| PC_18 | -0.11486 | 0.28182 | -0.41 | 0.68 |
| PC_19 | -0.20698 | 0.45221 | -0.46 | 0.65 |
| PC_20 | 0.24512 | 0.52662 | 0.47 | 0.64 |
| PC_21 | 0.02006 | 0.80824 | 0.02 | 0.98 |
| PC_22 | 0.03705 | 0.42479 | 0.09 | 0.93 |
| PC_23 | -0.05492 | 0.21595 | -0.25 | 0.80 |
| PC_24 | -0.03659 | 0.39387 | -0.09 | 0.93 |
| PC_25 | 0.12967 | 0.36788 | 0.35 | 0.72 |
| PC_26 | 0.17978 | 0.45077 | 0.40 | 0.69 |
| PC_27 | 0.22016 | 0.36952 | 0.60 | 0.55 |
| PC_28 | -0.17726 | 0.33725 | -0.53 | 0.60 |
| PC_29 | 0.07516 | 0.28726 | 0.26 | 0.79 |
| PC_30 | -0.26306 | 0.64525 | -0.41 | 0.68 |
| PC_32 | 0.01172 | 0.28746 | 0.04 | 0.97 |
| PC_33 | 0.05052 | 0.31882 | 0.16 | 0.87 |
| PC_34 | 0.04446 | 0.50963 | 0.09 | 0.93 |

| | | | | |
|---|---|---|---|---|
| PC_35 | -0.13883 | 0.12017 | -1.16 | 0.25 |
| PC_37 | -0.15226 | 0.31282 | -0.49 | 0.63 |
| PC_38 | 0.16117 | 0.32670 | 0.49 | 0.62 |
| PC_39 | 0.03993 | 0.31855 | 0.13 | 0.90 |
| PC_40 | 0.03362 | 0.61409 | 0.05 | 0.96 |
| PC_41 | 0.10981 | 0.23249 | 0.47 | 0.64 |
| PC_42 | -0.05350 | 0.25811 | -0.21 | 0.84 |
| PC_43 | -0.15935 | 0.32249 | -0.49 | 0.62 |
| PC_44 | -0.07370 | 0.26484 | -0.28 | 0.78 |
| PC_45 | 0.13871 | 0.66629 | 0.21 | 0.84 |
| PC_46 | 0.20069 | 0.34541 | 0.58 | 0.56 |
| PC_47 | -0.04556 | 0.27114 | -0.17 | 0.87 |
| PC_48 | 0.01446 | 0.26437 | 0.05 | 0.96 |
| PC_49 | -0.02140 | 0.25101 | -0.09 | 0.93 |
| PC_50 | 0.28416 | 0.55376 | 0.51 | 0.61 |
| PC_51 | 0.19285 | 0.25288 | 0.76 | 0.45 |
| PC_52 | -0.00791 | 0.46210 | -0.02 | 0.99 |
| PC_53 | 0.02943 | 0.16400 | 0.18 | 0.86 |
| PC_54 | 0.10676 | 0.53881 | 0.20 | 0.84 |
| PC_55 | 0.09673 | 0.12380 | 0.78 | 0.43 |
| PC_56 | 0.16650 | 0.22484 | 0.74 | 0.46 |
| PC_57 | -0.13708 | 0.11624 | 0.24 | 0,24 |
| PC_58 | 0.09441 | 0.25730 | 0.37 | 0.71 |
| PC_59 | -0.00539 | 0.31708 | -0.02 | 0.99 |
| PC_60 | -0.10365 | 0.10880 | -0.95 | 0.34 |

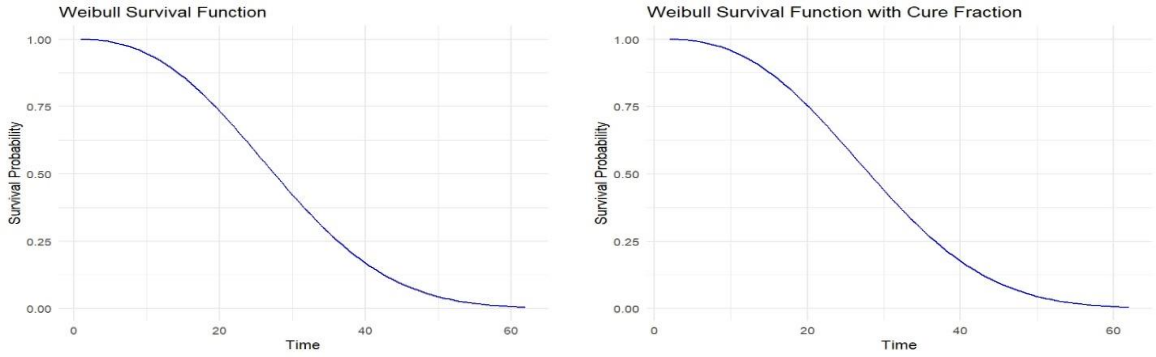
In Table 3, the highly significant intercept ($\beta_0$) with a coefficient of 3.70750 and a very small *p*-value of < 2e-16 represents the baseline hazard (or log-hazard), which is the logarithm of the hazard rates when all covariates (PC_1, PC_2, ..., PC_60) are zero with a positive high z-value of 25.93 indicating that baseline hazard is significantly different from zero. The PC's coefficient ($\beta_i$) represents the log-relative hazard (the natural log of the hazard ratio) associated with the principal components. The *p*-values indicate the statistical significance of each coefficient and the z-value shows the deviation of the coefficients away from zero. However, in this model, all the PC's coefficients have a *p*-value greater than 0.05, and all the std. errors are larger relative to the coefficients leading to smaller *z*-values and higher *p*-values in all the PCs indicating that they are not statistically significant, suggesting that the principal components do not strongly influence the survival time of the colon cancer patients.




(i) (ii)

**Figure 5. (i) Kaplan Meier curve estimating the survival rate of all the groups of patients over time (ii) Kaplan-Meier curve with the censoring indicator for all groups of patients over time**

In Figure 5 (i), the Kaplan- Meier curve shows the survival probabilities of the groups without censoring for all the colon cancer patients with the PC covariates for 60 months (Approximately 5 years). The median survival time by the plot is at the 30$^{th}$ month with an estimated survival probability of approximately 70% indicating that 30% of the patients are likely to survive the event of interest (death) beyond that time. The shaded area around the survival curve indicates the confidence limit for the survival probability estimates, wherever it widens indicates less precision of the survival time estimates and shows better at the narrower levels. In Figure 5 (ii), Kaplan Meier's plot incorporates censoring and the number at risk at different periods of 60 months, with all 62 patients at the start of the study (time 0) under observation, 43 remain at risk at time 20, meaning 19 patients have experienced the event (death) or were censored by this time, 23 are at risk at time 40, meaning additional patients have either experienced the event or been censored, and 3 remain at risk at time 60, indicating a very small number of patients are still being observed at the end of the study period. The survival curve shows a more gradual decline compared to the previous plot. This suggests that while events occur steadily, the rate is slower and more uniform. The tail end of the curve shows a slight flattening, suggesting the possibility of long-term survivors or a cure fraction. The confidence intervals are wider towards the end, indicating more uncertainty in the survival estimates as the number of subjects at risk decreases.




(i) (ii)

**Fig. 6. (i) Weibull survival distribution curve for the complete data with the PCs as covariates (ii) The Weibull survival distribution curve for the cure fraction**

Figure 6 (i) shows the survival rate of colon cancer patients, using gene expression principal components (PCs) as covariates, with the data fitted to the Weibull proportional hazard mixture cure fraction model while Figure 6 (ii) represents the survival probability for the patients with the proportion of those that are cured. The Weibull curve shows a decreasing survival probability over time, which is expected in most real-world scenarios in both curves and Figure 6 (ii) the distribution of the survival probability plateaus at a non-zero value, indicating the presence of a cure fraction. However, the Weibull shows the insignificance of all the PCs (covariates) retained by the PCA indicating it is a poor fit for the scenario and suggests the preference of the Cox proportional hazard model and the mixture cure fraction with covariates and right-censoring as the best for modeling the survival time of the colon cancer patients with gene expressions as covariates.

**Table 4: Parameter Estimates for Weibull and Cox PHMCF Models using MLE with a sample of size 62**

| Model | Log-Likelihood (Model) | Log-Likelihood (Intercept Only) | Chi-squared (Chisq) | Degrees of Freedom (df) | *p*-value | Iterations | Shape ($\alpha$) | Scale ($\lambda$) | Cure Fraction |
|---|---|---|---|---|---|---|---|---|---|
| Weibull PHMCF | -132.1 | -169 | 73.69 | 60 | 0.11 | 2000 | 0.5749 | 3.9089 | 1.00 |
| Cox PHMCF | -133.6 | …….. | 70.87 | 58 | 0.02 | 1000 | …… | …… | 0.56 |

Table 4 shows the results of the MLE explored for Weibull and Cox to maximize their parameters. The estimates of the Weibull parameter show that the log-likelihood for the model with covariates is higher than that of the model with only the intercept, indicating the model with covariates fits the data better. The chi-square value of (73.69) indicates an improvement in fit, but not statistically significant, since the p-value > 0.05. It assesses the overall fit of the model. The degree of freedom (60) represents the number of parameters the model estimates (including covariates and intercepts). The p-value of (0.11) across all the parameters estimated shows that despite the model fitting better, modeling with the addition of covariates is not statistically significant. The number of iterations for the optimization algorithm used to fit the model was 2000. The shape parameter ($\alpha = 0.5749$) for $\alpha < 1$, indicates that the hazard decreases over time, suggesting that the risk of the event (death) as a result of the colon with such genes decreases as time progresses. The relatively large scale ($\lambda = 3.9089$) parameter indicates that the event (death) occurs later in the time scale. The estimated survival probabilities for 60 months (approximately 5 years) show a consistent decrease over time, which is expected in survival analysis with the Weibull mixture cure model. However, the large fluctuations in CI's especially at later times indicate greater uncertainty which could imply less robustness in the estimates. The model estimated that 100% of the patients are cured, which may be an overestimation. By this information the dataset is suitable for the Weibull model as it allows for estimating hazard shape and scale, providing detailed insights into how risk changes over time. However, the overestimated cure fraction (100%) raises concerns about its applicability to cure rate analysis.

The Cox model uses partial likelihood estimation, focusing on how covariates affect the hazard ratio relative to a baseline. Therefore, the focus is on the proportional effects of covariates rather than the baseline hazard. As a result, the Cox model doesn't directly estimate the baseline hazard or provide a straightforward log-likelihood for an intercept-only model, making full likelihood calculations less applicable in this research. Unlike the Weibull where the log-likelihood for the intercept only serves as a baseline for comparison of a model fit. Similarly, the chi-square (70.87) value for the Cox assesses

how well the model fits the data compared to the intercept-only model, with a p-value < 0.05, indicating a better fit with the covariates, over the intercept-only with 58 degrees of freedom. Moreover, the number of iterations for the optimization algorithm used to fit the model was 1000. The Cox model estimates that 0.56 percent of the patients are cured and 0.44 uncured. However, the Cox PHMCF Model does not estimate shape or scale parameters but offers a more moderate and potentially realistic cure fraction of 56%, which better indicates the proportion of patients cured, indicating that the dataset is well-suited for the Cox model, particularly when focusing on the proportional hazards of covariates without the need for a specific baseline hazard form. The realistic cure fraction (56%) makes the Cox model more practical and reliable for interpreting survival and cure rates.

## 4. Discussion

To evaluate the effectiveness of Weibull and Cox baseline hazard functions within proportional hazard models incorporating mixture cure fractions, particularly when modeling data that include high dimensional covariates and right-censoring, the research took advantage of the colon cancer DNA microarray data set [15]. The data contains 40 tumors and 22 normal colon human gene tissues of (62) observations of 2000 gene expressions acquired using an Affymetrix oligonucleotide array, making it a $Mxn$ data set. Consequent to the high dimensionality of the data, the principal component analysis technique was employed to reduce the dimension of the redundant dataset while retaining all vital information needed for the evaluation and inference. Three of the common data reduction approaches in the PCA were explored: Eigenvalue-one-criteria, the proportion of variance accounted, and scree plot to better assess and logically retain the $PC_i$ as the covariates of the data for subsequent analysis using the proposed models. A few 60 $PC_i$ were retained as covariates according to the eigenvalue one criterion approach, 7 pcs according to the proportion of variance accounted for, and only 1 according to the scree plot approach. Eigenvalue one criterion suggestion of retaining 60 $PC_i$ and used as covariates was used as it gives more reliable $PC_i$ , [16].

Furthermore, the retained 60 $PC_i$ (covariates) were estimated by CPHM and found that all of them are associated with a high hazard rate with 24 of them exhibiting a greater hazard ratio of exponentiated coefficients exp $(\beta_i) \gg 1$ indicating that the influence of these $PC_i$ is so strong that even small changes in their values could lead to large increases in hazard that is the relative risk of the death of the colon cancer patients associated with each $PC_i$ (covariates), moreover explaining that for every unit increase in the $PC_i$ , the risk of the event (death) increases exactly times the number of the Hazard Ratio. The entire coefficients $\beta_i$ are large and positive, indicating an increased effect of $PC_i$ on the hazard rate. The corresponding p-values are all extremely low and less than 0.05 typically suggesting that the $PC_i$ are having a significant impact on the hazard rate or are all statistically significant contributors to the hazard rate of the colon patient's event(death). In contrast, the estimated $PC_i$ by the WPHM shows that only the intercept coefficient is significant, suggesting a baseline hazard that is constant regardless of the $PC_i$ values. In this model, all the $PC_i$ coefficients have p-values greater than 0.05, and the standard errors are relatively large compared to the coefficients. This results in low z-values and high p-values, indicating that none of the $PC_i$ are statistically significant. The estimation by the WPHM indicates that none of the principal components $PC_i$ in this model have a statistically significant effect on the hazard (risk of the event, death) at the conventional 0.05 level of significance to the colon cancer patients.

Lastly, the general survival time was estimated by the Cox proportional hazard mixture cure fraction model with the retained covariates $(PC_i)$ and right-censoring. The results showed that the survival probabilities start very high, close to 1.0 at early time points, with narrow confidence intervals, indicating high precision and a strong likelihood that almost all individuals survive during this period

and gradually decrease over time. The survival probability continues to decrease significantly, reflecting that fewer individuals are expected to survive as time progresses. The confidence interval widens at this point indicating greater uncertainty in the survival probability estimates later. The cure proportion was estimated to be 56% and non-cure 44.4% showcasing the heterogeneity of the population. The outcome indicates that the cure proportion is not susceptible, while the cure is susceptible to the event(death), and remains at risk at the follow-up period of 5 years. Similarly, the estimate with the Weibull proportional hazard mixture cure fraction model showed that the Weibull model also provides precise estimates, with confidence intervals that are relatively narrow at early time points. However, as time progresses, the confidence intervals widen more significantly than in the Cox model, indicating increased uncertainty in long-term survival estimates at the same follow-up period of 5 years. The mixture cure fraction with the model as a baseline shows that 100% of the population will be cured at the follow-up period of 5 years which could be attributable to the insignificant of the covariates to the hazard risk estimated at the preliminary stage of the analysis by the Weibull proportional hazard model.

## 5. Conclusion

Modeling the survival time of cancer using high-dimensional DNA microarray data is an important research area. However, the challenges faced by high-dimensional data, especially in gene reduction and selection, often lead to the failure of many penalized likelihood methods in identifying a small, significant subset of the genes. To address this problem, the present study proposed and applied the concept of an unsupervised machine learning algorithm approach called principal component analysis (PCA) to perform gene reduction and estimation of its coefficients simultaneously. Given the basic rule in survival analysis regarding the event-to-covariate ratio, the sample size of 62 relatives to the PCs remains crucial even though PCA has reduced the dimensionality from 2000 to 60 as the covariates. However, the model selection process naturally addresses this ratio issue by employing AIC as a goodness-of-fit evaluation metric, guaranteeing that the best-fitting model is selected while considering the model's complexity.

Afterward, the study went further to propose two models; Cox (semi-parametric) and Weibull (parametric) models as a baseline hazard function to estimate the general survival probability of colon cancer patients using the retained 60 $PC_i$ by the eigenvalue 1 criteria approach of the PCA with proportional hazard incorporating mixture cure fraction models. From the findings which were computed using the principal components (covariates) and survival time with right-censoring from the colon cancer microarray data set, it was confirmed that the Cox proportional hazard mixture cure fraction model appears to be the better model based on its flexibility, precision, smaller AIC (129.1606) and better fit to the data. It provides more reliable estimates and aligns closely with the observed survival patterns. The Weibull got a larger AIC (388.2784) compared to that of the Cox. While the Weibull model is useful in specific contexts, the Cox model's adaptability and stronger performance in fitting the data make it the preferable choice in this analysis. Generally, the results established the detail that the CPHMCF model is a very feasible technique that can analyze DNA microarray cancer data accurately. In addition, the proposed CPHMCF model results can be applied practically to other related high-dimensional data for cancer classification and prognosis. It could be applied in instances where the patients exhibit some higher covariates, such as aggressive cancer subtypes, late-stage diagnoses, or high-risk genetic profiles. By providing estimates and accounting for cure fractions, the model can help in many cancer treatment strategies. These results demonstrate the model's potential value in more general medical research by improving decision-making and predicting accuracy in a range of cancer prognosis studies. As a result, we may implement the suggested CPHMCF model in the medical research field with effectiveness.

## 6. Acknowledgment

The Authors thank the Tertiary Education Trust Fund (TETFUND), Nigeria for providing financial support to the achievement of the paper.

## 7. References

Algamal, Z. Y., Alhamzawi, R., & Mohammad Ali, H. T. (2018). Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression. *Computers in Biology and Medicine*, *97*, 145–152. https://doi.org/10.1016/j.compbiomed.2018.04.018

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

Al-Thanoon, N. A., Qasim, O. S., & Algamal, Z. Y. (2018). Tuning parameter estimation in SCAD-support vector machine using firefly algorithm with application in gene selection and cancer classification. *Computers in Biology and Medicine*, *103*, 262–268. https://doi.org/10.1016/j.compbiomed.2018.10.034

Atinafu, B. T., Bulti, F. A., & Demelew, T. M. (2020). Survival Status and Predictors of Mortality Among Colorectal Cancer Patients in Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia: A Retrospective Follow-up Study. *Journal of Cancer Prevention*, *25*(1), 38–47. https://doi.org/10.15430/JCP.2020.25.1.38

Badisy, I. E., BenBrahim, Z., Khalis, M., Elansari, S., ElHitmi, Y., Abbas, F., Mellas, N., & Rhazi, K. E. (2023). *Risk factors affecting patients survival with colorectal cancer in Morocco: Survival Analysis using an Interpretable Machine Learning Approach*. https://doi.org/10.21203/rs.3.rs-2435106/v1

Bai, J., Zhang, X., Xiang, Z.-X., Zhong, P.-Y., & Xiong, B. (2020). Identification of prognostic immune-related signature predicting the overall survival for colorectal cancer. *European Review for Medical and Pharmacological Sciences*, *24*(3), 1134–1141. https://doi.org/10.26355/eurrev_202002_20164

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713–1723. https://doi.org/10.1002/sim.2059

Das, A., Neelima, N., Deepa, K., & Özer, T. (2024). Gene Selection Based Cancer Classification With Adaptive Optimization Using Deep Learning Architecture. *IEEE Access*, *12*, 62234–62255. https://doi.org/10.1109/ACCESS.2024.3392633

Ding, D., Lang, T., Zou, D., Tan, J., Chen, J., Zhou, L., Wang, D., Li, R., Li, Y., Liu, J., Ma, C., & Zhou, Q. (2021). Machine learning-based prediction of survival prognosis in cervical cancer. *BMC Bioinformatics*, *22*(1), 331. https://doi.org/10.1186/s12859-021-04261-x

Gewers, F. L., Ferreira, G. R., Arruda, H. F. D., Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. D. F. (2022). Principal Component Analysis: A Natural Approach to Data Exploration. *ACM Computing Surveys*, *54*(4), 1–34. https://doi.org/10.1145/3447755

Hossain, Md. A., Saiful Islam, S. M., Quinn, J. M. W., Huq, F., & Moni, M. A. (2019). Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *Journal of Biomedical Informatics*, *100*, 103313. https://doi.org/10.1016/j.jbi.2019.103313

Infante, G., Miceli, R., & Ambrogi, F. (2023). Sample size and predictive performance of machine learning methods with survival data: A simulation study. *Statistics in Medicine*, *42*(30), 5657–5675. https://doi.org/10.1002/sim.9931

Kyriazos, T., & Poga, M. (2023). Dealing with Multicollinearity in Factor Analysis: The Problem, Detections, and Solutions. *Open Journal of Statistics*, *13*(03), 404–424. https://doi.org/10.4236/ojs.2023.133020

Lenz, M., Müller, F.-J., Zenke, M., & Schuppert, A. (2016). Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Scientific Reports*, *6*(1), 25696. https://doi.org/10.1038/srep25696

Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C. J., Laversanne, M., Vignat, J., Ferlay, J., Murphy, N., & Bray, F. (2023). Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from GLOBOCAN. *Gut*, *72*(2), 338–344. https://doi.org/10.1136/gutjnl-2022-327736

Razzaque, A., & Badholia, D. A. (2024). PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification. *Measurement: Sensors*, *31*, 100945. https://doi.org/10.1016/j.measen.2023.100945

Rutter, C. M., Nascimento De Lima, P., Maerzluft, C. E., May, F. P., & Murphy, C. C. (2023). Black-White disparities in colorectal cancer outcomes: A simulation study of screening benefit. *JNCI Monographs*, *2023*(62), 196–203. https://doi.org/10.1093/jncimonographs/lgad019

Xie, T., Fu, D.-J., Li, K.-J., Guo, J.-D., Xiao, Z.-M., Li, Z., & Zhao, S.-C. (2024). Identification of a basement membrane gene signature for predicting prognosis and estimating the tumor immune microenvironment in prostate cancer. *Aging*, *16*(2), 1581–1604. https://doi.org/10.18632/aging.205445