Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

A Multi-Model Survival Analysis of Lung Cancer Using Parametric Techniques

Pitta Shankaraiah, Mokesh Rayalu. G*

School of Advanced sciences, Vellore Institute of Technology, Vellore, India
E-mail: mokesh.g@vit.ac.in

Received: 9 October 2024; **Revised:** 8 November 2024; **Accepted:** 14 November 2024

Abstract: Lung cancer remains one of the leading causes of cancer-related mortality worldwide, underscoring the critical need for effective prognostic tools. This study utilizes survival analysis to explore the factors that influence the survival outcomes of North Central Cancer Treatment data related to lung cancer. The main goal of this study is to compare and contrast various statistical models, including the Weibull, Exponential, Log-gaussian, Gumbel, and Rayleigh models. We have computed important functions, such as the survival function, the hazard function, and the cumulative hazard function, for all the considered distributions. The Anderson-Darling and Cramer Von-Mises tests, which are Goodness of fit tests, effectively compare and assess various parametric regression survival models. The Weibull survival model is interpreted to be the most effective and efficient way to study the lung cancer dataset, which is concluded upon evaluating the results of Anderson-Darling statistic 0.28745, Cramer Von-Mises statistic 0.0450, Mean Survival Probability 0.9697, Mean Cumulative Survival Probability 0.0303, Akaike Information Criterion 1,650.753 and Bayesian Information Criterion 1623.329 of the Weibull, Exponential, Log-gaussian, Gumbel, and Rayleigh parametric regression survival models.

Keywords: Akaike Information Criterion, Bayesian Information Criterion, Goodness of Fit, Parametric Models, Survival Functions, Survival Probability

MSC: 60E05, 62B15, 62F30, 62G05, 62P10

Abbreviation

AIC, BIC	Akaike Information Criterion and Bayesian Information Criterion
AD, CvM	Andeson-Darling and Cramer Von-Mises tests
KM, WD	Kaplan-Meier and Weibull Distribution
MSP, MCSP	Mean Survival Probabilities and Mean Cumulative Survival Probabilities
Pdf, CDF	Probability density function and Cumulative Distribution Function
PRSM, GoF	Parametric Regression Survival Models and Goodness of Fit
SF, HF, CHF	Survival Function, Hazard Function, and Cumulative Hazard Function

DOI: https://doi.org/10.37256/cm.5420245879

1. Introduction

Due to its high death rate and substantial impact on public health, lung cancer represents a major global health concern [1]. Lung cancer is the leading cause of death associated to cancer than combined effects of colorectal, prostate, and breast cancers, as reported by the World Health Organization [2]. Its high incidence and advanced diagnosis stage, which restricts available treatments and adversely affects survival rates, are what make it so insidious [3]. The survival rate of lung cancer patients is significantly lower than that of individuals with other cancer types. The challenging objective is to identify the nodule regions located within the soft lung tissues during the initial stage of lung cancer. Computed Tomography and Chest Radiography are utilized for the identification of pulmonary nodules indicative of lung cancer [4]. Based on projections of 1.79 million deaths from the disease by 2020, lung cancer is foremost curable cancer globally malignancies of the stomach and liver rank second in terms of mortality, following malignancies of the colon [5]. Many of deaths globally are caused by heart disease and stroke, with lung cancer ranking only sixth in terms of frequency of death [6]. The most common cancer killer on a global scale is lung cancer. It was anticipated that 236,740 persons get a lung cancer diagnosis by 2022 and 238,340 by 2023. Compared to the previous year, there has been an increase in the number of new cases to 1,600 and a rise in deaths to 460. One out of every sixteen men and one out of every seventeen women will get a lung cancer diagnosis at some point in their lives [7]. Recent years have shown a consistent decline in the overall incidence of lung cancer cases. The annual decline in the incidence rate for men was 2.6% and for women it was 1.1% from 2006 to the present. As a result of improvements in diagnosis and treatment, mortality rates are falling at an even quicker rate [8].

The primary goals of the present research study are as follows:

To identify the pairwise relationship by considering all the variables using Pearson correlation.

To ascertain the Survival Function (SF), Hazard Function (HF), and Cumulative Hazard Function (CHF) through Exponential, Log-gaussian, Gumbel, and Rayleigh models using the Weibull Distribution (WD).

To apply parametric models to the data of lung cancer patients in order to determine the mean survival time, Mean Survival Probability (MSP), and Mean Cumulative Survival Probability (MCSP).

Lastly, use the Anderson-Darling (AD), Cramer Von-Mises (CvM), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values to identify the best as well as most effective Parametric Regression Survival Model (PRSM).

The Goodness of Fit (GoF) tests and Model selection criteria are essential for evaluating parametric regression models. The AD test assesses the fit of the regression model to the data, focusing on the tails of the distribution, while the CvM test evaluates the overall fit, emphasizing the central region. This GoF tests enables model validation by verifying the adequacy of the parametric regression model. The AIC estimates the relative loss of information when approximating the true model with a simpler one, while BIC penalizes models with more parameters, balancing complexity and fit. Collectively, AD, CvM, AIC, and BIC serve three primary purposes: model validation, model selection, and overfitting/underfitting detection. The remaining sections of this article are arranged as follows: In Part 2, the publications of various authors that are pertinent to the study of lung cancer are summarized. Part 3 covers the details of the data processing, look at the elements that affect lung cancer, and apply the mathematical framework. Part 4 displays the results and discussion that were obtained. Part 5 provides an explanation of the conclusion sections.

2. Literature survey

An investigation by Geremew et al. [9] was aimed to evaluate the incidence of tuberculosis in pregnant women in northwest Ethiopia and to identify risk factors. Of the 289 affected women, 29 (10.03%) developed active TB. Throughout the course of a thousand persons per month under surveillance, 17.4 instances total were noted. The Kaplan-Meier (KM) survival curve and the Weibull regression model were used, respectively, to predict the probability of survival and highlight the risk factors for tuberculosis. To choose the best model amongst the parametric and Semi-parametric models, the AIC and BIC have been used. The study conducted by Wong et al. [10] looks on the mortality risk that changes over time following a lung cancer lobectomy. Making use of parametric survival models and a hazard function

shown over time in accordance with the best fit statistical distribution, it examined 2,284 patients between 2015 and 2022. At 30, 90, and 180 days, the cumulative death rate was 1.3%, 2.9%, and 4.9%, according to the results. At the time of the postoperative phase, a moment's hazard rate has been increased. Within the initial 30 days, it quickly dropped and stabilized at 180 days. According to a study discussed by Everest et al. [11], the observed survival time based on current data with the expected survival time based on parametric expansions. Following an analysis of thirty-two randomized controlled trials, it examined US Food and Drug Administration permissions for oncology from 2006 to 2015 and computed the restricted mean survival time, mean deviation, mean absolute error, and mean absolute percentage error by applying KM curves. The predicted mean survival time significantly outperformed the updated predictions, according to the results. As time was stretched and the proportion of censored patients rose, the mean absolute error between the projected and observed survival rate increased. Cislo et al. [12] compared finite mixture models to typical survival models for heterogenous data fit and analyzed the cost-effectiveness using mean survival times. Digitizing public overall survival rate as well as progression free survival curves and fitting regression models with diverse distributions, the AIC exists while comparing the accuracy of models. The 3-Weibull mixture and 2-Weibull mixture models exceeded others by over 40 points for PFS and overall survival. The 3-Weibull and 2-Weibull mixture models nearest to 17.58 months KM mean estimate and all models estimated survival time within 10% of the KM mean. A study made by Wulandari et al. [13] has three methods for survival analysis viz., semiparametric, nonparametric, and parametric. Because of its simplicity and versatility, Weibull regression is a widely used kind of parametric modeling that requires the survival time distribution. When the assumptions of Cox proportional hazard model are broken, stratified Cox regression, which is a semi-parametric, can be employed. While evaluating the duration of exclusive breastfeeding in Indonesian infants aged 0 to 6 months, a study comparing Weibull regression and stratified Cox regression discovered that Weibull regression was more superior to stratified Cox regression. Sato et al. [14] described data from 112 patients with lessened side effects from treatment for advanced pancreatic ductal adenocarcinomas retrospectively analyzed patients who received chemotherapy at a single institution. The Eastern Cooperative Oncology Group performance status, poor mass in the skeletal muscle's ascites, and carcinoembryonic antigen levels were found to be associated with the highest risk of death. The study focused on the Cox, Weibull, and standardized Exponential models using hazard ratios and AIC. The Weibull-Exponential distribution model, which implies a connection among mass in skeletal muscles and pancreatic ductal adenocarcinomas results, was used to assess a single year survival probability.

According to a study proposed by Almongy et al. [15], statistical technique for examining COVID-19 death rates in Italy, Mexico, and the Netherlands is presented in the research. It presents the great features of the extended odd Weibull Rayleigh distribution. Applying maximum likelihood, maximum product spacing, and Bayesian estimation techniques, real data analysis and Monte Carlo simulation findings are taken into account. A study of Ahmed [16], uses parametric techniques such as Weibull, Gumbel, Exponential, and Log-logistic regression models to derive the SF of the patients with lung cancer. Upon comparing the techniques with corrected, Akaike, and Bayesian information, the Gumbel distribution model was concluded to be the most effective. There is an inverse link between failure times and SF, as the latter increases and the former lowers. Wegbom et al. [17], has found using survival analysis that the factors influencing under-five mortality were increasing in Nigeria. Because of it's adaptability and lack of baseline hazard function specification, Cox model is frequently utilized. This study compares various models using data from 2013 Nigeria Demographic and Health Survey. The best model was determined using the Cox-Snell residual and the AIC. According to the study, parametric models performed better than the Cox model. In Malaysian lung cancer patients, in a study measured by Jamil et al. [18], the fixed covariate of right censored data used a parametric survival model. Using R-statistical software for several sample sizes (50, 100, 150, and 200), the model was evaluated with the appropriately censored data. The outcomes demonstrated that the simulation method for correctly censored data was enhanced by varying the shape and scale parameter values and sample sizes, and the Weibull regression survival model was appropriate for the analysis. A study conducted by Wang et al. [19] was on the genetic relationship between analysis of age at onset and the Hepatocyte Nuclear Factor-1-beta gene compares Cox Proportional Hazard model with parametric survival models while looking at 23 single nucleotide polymorphisms inside the Hepatocyte Nuclear Factor-1-beta gene in the Marshfield population, which includes 716 cancer cases and 2,848 non-cancer controls. The Weibull distribution is the most accurate model for all 23 SNPs, according to

the results, with the Gamma distribution coming in second. This offers the first proof of several genetic variations in the Hepatocyte Nuclear Factor-1-beta gene linked to age at onset in cancer.

A study conducted by Saho et al. [14] was on a comprehensive study on survival models, incorporating both parametric models and semi-parametric models. Their analysis revealed that the WD yielded superior results based on the AIC. However, the study did not include standard error and a GoF tests for the parametric models. Furthermore, Almongy et al. [15] extending the analysis from the WD to the Rayleigh distribution using a COVID-19 dataset. This study focused solely on the GoF to determine efficacy, without utilizing the AIC or BIC metrics. Moreover, Ahmed [16] utilizing various PRSM with lung cancer data. The effectiveness of these models was determined solely using AIC and BIC metrics. However, to better assess the efficacy of the models, incorporating GoF tests and analyzing the significance among association of covariates would provide more comprehensive results. Notably, previous research emphasizes the amalgamation of parametric distributions concerning parameter count and distribution shape, yet it overlooks the association between features and their impacts, representing a potential avenue for further investigation.

3. Dataset modeling and simulation

The survival data for patients with advanced lung cancer in the North Central Cancer Treatment Group from Terry Therneau is utilized in this study. The patients performance ratings show that they are capable of performing daily, regular activities. All these variables are featured in the 168 rows and 9 columns of the data sample. Every unit in this data has been utilized to establish the time and status which are independent variables and all the other variables Age (x_1), Sex (x_2), Ph. E (x_3), Ph. K (x_4), Pa. K (x_5), MC (x_6), and WT (x_7) are dependent. Every piece of information has been shown in Table 1.

Table 1. The list of variables representing the lung cancer dataset

S NO.	Variables	Description
1	Time	Time represented in days
2	Status	Censoring status (Censored 1 and Dead 2)
3	Age (x_1)	Survival time in days
4	Sex (x_2)	Male 1 and Female 2
5	Ph. E (x_3)	Ecog performance rating (Good 0 to Dead 5)
6	Ph. K (x_4)	Physician-rated karnofsky performance rating (Bad 0 to Good 100)
7	Pa. K (x_5)	Patient-rated karnofsky performance rating (Bad 0 to Good 100)
8	MC (x_6)	Calories taken during meals
9	WT (x_7)	Loss of weight throughout the past six months

Considering the constructed dataset within the context of patients suffering from advanced lung cancer, Figure 1 illustrates the entire functionality of the workflow which is used to handle the data. Figure 1 shows a conceptual framework for comparing and evaluating several PRSMs that are specific to data on lung cancer. Several PRSMs, such as the Weibull, Exponential, Log-gaussian, Gumbel, and Rayleigh models, are applied to lung cancer repository data at the start of the process. These models were chosen because they are relevant to survival analysis and have unique capacities to depict different hazard rate patterns over time. Several statistical metrics, including AD, CvM, AIC, and BIC, are computed to evaluate the performance of the model while taking into account the various predictor variables x_1, x_2, $----$, x_7. To understand how these factors affect patient survival, survival analysis is carried out using both non-parametric and PRSM approaches. This investigation finds the model that best predicts survival for patients with lung cancer and helps to clarify the factors that affect survival outcomes. Significant demographic variables that may affect survival rates, are x_1 and x_2. These factors frequently reveal disparities in survival rates, with younger patients and females typically exhibiting better prognoses for specific cancer kinds. While evaluating a patient's ability to carry out daily tasks, the performance scores

x_3, x_4, x_5 are essential. Greater functional status, which is correlated with better survival outcomes, is indicated by higher scores [20]. Two nutritional status indicators that may have an impact on general health and possibly the prognosis of cancer patients is x_6 and x_7.

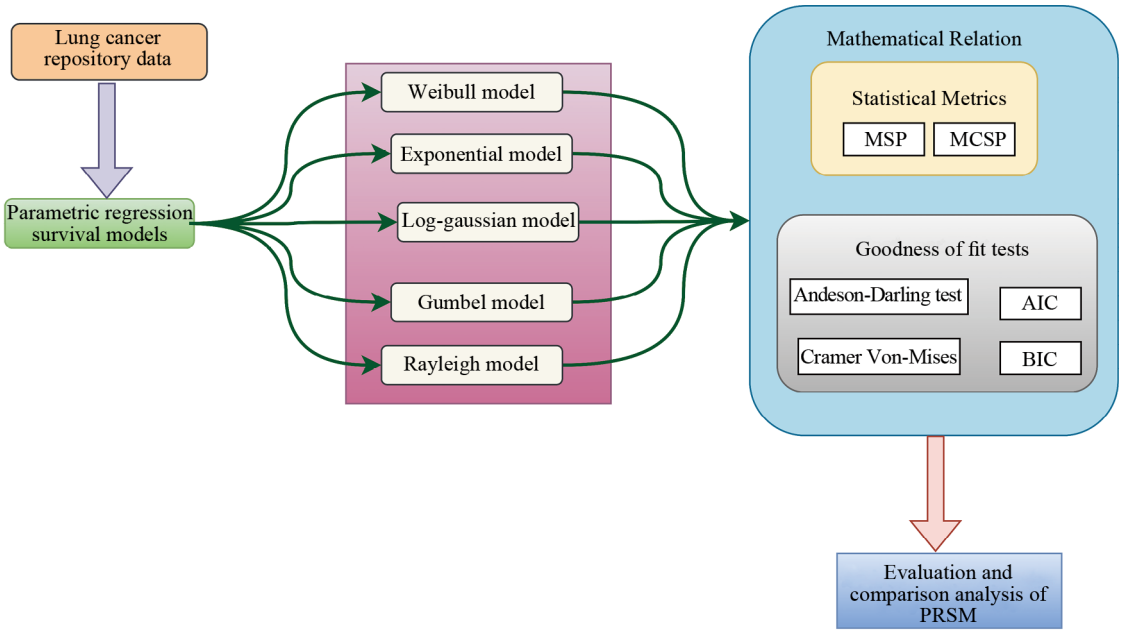

Figure 1. A flowchart for estimating the optimal PRSM

In this study, any possible interactions between each of the variables have been examined. Table 2 describes the relationship between time and a feature lung cancer by using the Karl Pearson correlation strategy [21].

The Pearson correlation coefficient for each pair of variables x_i and x_j is given by

$$r_{(x_i,\, x_j)} = \frac{\operatorname{cov}(x_i,\, x_j)}{\sigma_{x_i} \cdot \sigma_{x_j}}; \quad i,\, j = 1,\, 2,\, \ldots,\, n; \quad -1 \le r \le 1$$

Where

$$\operatorname{cov}(x_i,\, x_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x_i})(x_{jk} - \bar{x_j})$$

$$\sigma_{x_i} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x_i})^2}$$

$$\sigma_{x_j} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x_j})^2}$$

A low negative correlation has been found between time and status which implies that: (1) the probability of the event, i.e., the failure rate or death rate, decreasing slightly with time; (2) the performance status of x_3 tends to be better but the relationship is not very strong; and (3) the performance score of x_4 assessed by patients tends to be slightly lower. x_5, x_6, x_7, and individuals' gender show a negligible association with time, as indicated by a low positive correlation. This could mean that doctors see improvement in the karnofsky performance score over time, that there has been a little increase in calories consumed during meals, and that there has been weight reduction throughout the last six months. Data from the present study determines how the analysis is applied [20].

Table 2. Pairwise analysis of correlation features in lung cancer

Pairwise correlation	Time	Status	(x_1)	(x_2)	(x_3)	(x_4)	(x_5)	(x_6)	(x_7)
Time	1	-0.16	-0.08	0.11	-0.19	0.09	0.18	0.07	0.03
Status		1	0.16	-0.22	0.24	-0.16	-0.19	0.02	0.05
(x_1)			1	-0.13	0.31	-0.33	-0.24	-0.24	0.05
(x_2)				1	-0.01	-0.02	0.07	-0.17	-0.17
(x_3)					1	-0.82	-0.54	-0.11	0.18
(x_4)						1	0.53	0.06	-0.13
(x_5)							1	0.17	-0.18
(x_6)								1	-0.11
(x_7)									1

3.1 *Mathematical derivation of Weibull distribution*

Two characteristics define the WD: shape and scale. The shape measure, which is commonly represented by the sign k, establishes the distribution's shape [22]. The spread of the distribution is determined by the scale measure, which is commonly represented by the symbol λ. For particular values of the shape parameter $k = 1, 2$. The WD is a particular variant of the Exponential and Rayleigh distributions. The Weibull and Gumbel distribution are related by extreme value theory, which explains the distribution of the highest and lowest value of the sample size of Weibull-distributed variables.

The Probability density function (Pdf) of the WD is given by:

$$f(x) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k};\ k > 0,\ \lambda > 0,\ t \geq 0.$$

The CDF is:

$$F(t) = 1 - e^{-(t/\lambda)^k}$$

The SF is the enhance of the CDF:

$$S(t) = 1 - F(t) = e^{-(t/\lambda)^k} \tag{1}$$

The HF represent the ratio of the Pdf to the SF:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}}{e^{-(t/\lambda)^k}} = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \tag{2}$$

The CHF is the integral of the HF from 0 to t:

$$H(t) = \int_0^t h(t)\, dt$$

Substitute HF, then we get

$$H(t) = \int_0^t \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} dt$$

Let $u = \dfrac{t}{\lambda}$, then $du = \dfrac{1}{\lambda} dt$, and the integral becomes:

$$H(t) = \int_0^{t/\lambda} k u^{k-1}\, du = \left[u^k\right]_0^{t/\lambda} = \left(\frac{t}{\lambda}\right)^k$$

Thus, the CHF is:

$$H(t) = \left(\frac{t}{\lambda}\right)^k \tag{3}$$

3.2 *Exponential-Weibull model: Application in survival functions*

The Exponential distribution is a continuous probability distribution that delineates the duration between events in a Poisson process, where occurrences transpire continuously, independently, and at a constant mean rate [23]. The WD with a shape parameter $k = 1$ is a specific instance of the Exponential distribution, that consists of the scale parameter λ. The Pdf of the Exponential distribution is expressed as:

$$f(t) = \lambda e^{-\lambda t};\ \lambda > 0,\ t \geq 0.$$

The CDF is the integral of the Pdf from 0 to t:

$$F(t) = \int_0^t \lambda e^{-\lambda t}\, dt$$

After simplifying, the CDF is:

$$F(t) = 1 - e^{-\lambda t}$$

The SF is the enhance of the CDF:

$$S(t) = 1 - F(t) = e^{-\lambda t} \tag{4}$$

The HF represent the ratio of the Pdf to the SF:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \tag{5}$$

Exponential distribution, the HF is constant. The CHF is the integral of the HF from 0 to t:

$$H(t) = \int_0^t h(t)\,dt = \int_0^t \lambda\,dt = \lambda t \tag{6}$$

3.3 *Log-gaussian distribution Via Weibull: A unified approach*

A random variable t follows WD with two parameters $k,\ \lambda$, then the Pdf of the WD is:

$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{k}\right);\ k > 0,\ \lambda > 0,\ t \geq 0.$$

To map a WD variable T to a Log-gaussian distribution, you can apply a logarithmic transformation. Consider a new variable Y defined as:

$$Y = \ln(T)$$

Now, if T is Weibull distributed, we need to check if Y follows a Gaussian distribution.
Assume $Y = \ln(T)$. The transformation will be:

$$f(y) = f(e^y) \cdot \frac{d}{dy}(e^y)$$

Since $\frac{d}{dy}(e^y) = e^y$, we have:

$$f_Y(y) = \frac{k}{\lambda}\left(\frac{e^y}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{e^y}{\lambda}\right)^{k}\right) \cdot e^y$$

Simplifying:

$$f_Y(y) = \frac{k}{\lambda} \exp\left(\frac{k-1}{y} - ky - \left(\frac{e^y}{\lambda}\right)^k\right)$$

The Log-gaussian distribution is obtained by taking the logarithm of a Gaussian distribution. If Y follows a Gaussian distribution, i.e., $Y \sim N(\mu,\ \sigma^2)$, then $T = e^Y$ follows a Log-gaussian distribution.

The Pdf of a Log-gaussian distribution is:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right)$$

The CDF is:

$$F(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right)$$

where Φ is the CDF of the standard Gaussian distribution.

Using the relationship of SF and CDF

$$S(t) = 1 - F(t)$$

$$S(t) = 1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{7}$$

By using relationship of Pdf and SF the computed HF is

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

Substitute the Pdf and SF of the Log-gaussian distribution then

$$h(t) = \frac{1}{S(t)\sigma t}\Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{8}$$

Using the relationship between the CDF and the SF, we know that:

$$H(t) = -\ln(S(t))$$

Substitute the SF of the Log-gaussian distribution:

$$H(t) = \Phi\left(\frac{\log t - \mu}{\sigma}\right) \tag{9}$$

3.4 *Investigating the Gumbel distribution Via Weibull distribution*

A Gumbel distribution is used to model the distribution of the optimum of multiple samples with various changes. Use of it is common in extreme value theory. Let's work through the transformation process to convert a WD into a Gumbel distribution and then derive the Pdf, CDF, SF, HF, and CHF for the Gumbel distribution.

The WD has a Pdf expressed as

$$f(t) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{k}\right);\ k > 0,\ \mu > 0,\ t \geq 0.$$

We want to transform the Weibull-distributed variable T into a new variable Y such that Y follows a Gumbel distribution.

Let:

$$Y = \frac{T^{1/k} - \mu}{\lambda}$$

Solve for T in terms of Y:

$$T = (\lambda Y + \mu)^{k}$$

We know that the Pdf of Y can be derived from the Pdf of T by:

$$f(y) = f(t)\left|\frac{dt}{dy}\right|$$

First, calculate $\frac{dt}{dy}$:

$$\frac{dt}{dy} = \frac{d}{dy}(\lambda y + \mu)^{k} = k(\lambda y + \mu)^{k-1} \cdot \lambda$$

So, the Pdf of Y is:

$$f(y) = f(t) \cdot \left|\frac{dt}{dy}\right|$$

Substitute the expressions for $f(t)$ and $\frac{dt}{dy}$:

$$f(y) = \frac{k}{\lambda}\left(\frac{(\lambda y+\mu)^k}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{(\lambda y+\mu)^k}{\lambda}\right)\right) \cdot k\lambda\,(\lambda y+\mu)^{k-1}$$

Simplifying, we get Pdf of the Gumbel distribution

$$f(y) = \frac{1}{\lambda}\exp\left(-\frac{(y-\mu)}{\lambda}\right)\exp\left(-\exp\left(-\frac{(y-\mu)}{\lambda}\right)\right)$$

This is the Pdf of the Gumbel distribution.
The CDF is defined as:

$$F(t) = \exp\left(-\exp\left(-\frac{t-\mu}{\lambda}\right)\right)$$

The SF is related to the CDF by:

$$S(t) = 1 - F(t)$$

Substituting the CDF into this equation, and then we get

$$S(t) = \exp\left(-\exp\left(-\frac{t-\mu}{\lambda}\right)\right) \tag{10}$$

The HF is defined as the ratio of the Pdf to the SF:

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

Now, substitute the Pdf and SF, Simplifying the equation:

$$h(t) = \frac{1}{\lambda}\exp\left(-\frac{t-\mu}{\lambda}\right) \tag{11}$$

The CHF is related to the SF by:

$$H(t) = -\ln(S(t))$$

Substituting the SF into this equation:

$$H(t) = -\ln\left(\exp\left(-\exp\left(-\frac{t-\mu}{\lambda}\right)\right)\right)$$

Simplify using properties of logarithms then CHF for the Gumbel distribution is:

$$H(t) = \exp\left(-\frac{t-\mu}{\lambda}\right) \tag{12}$$

3.5 *Exploring the Rayleigh distribution Via Weibull distribution*

The Rayleigh distribution is often used to model the magnitude of a vector that has two independent, normally distributed components [24]. Below are the derivations for the Pdf, CDF, SF, HF, and CHF of the Rayleigh distribution.

The Pdf of the Rayleigh distribution is:

$$f(t) = \frac{t}{\lambda^2} e^{-\frac{t^2}{2\lambda^2}};\ \lambda > 0,\ t \geq 0.$$

The CDF is:

$$F(t) = 1 - e^{-\frac{t^2}{2\lambda^2}}$$

The SF is the enhance of the CDF:

$$S(t) = 1 - F(t) = e^{-\frac{t^2}{2\lambda^2}} \tag{13}$$

The HF is the ratio of the Pdf to the SF:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{t}{\lambda^2} e^{-\frac{t^2}{2\lambda^2}}}{e^{-\frac{t^2}{2\lambda^2}}} = \frac{t}{\lambda^2} \tag{14}$$

The CHF is the integral of the HF from 0 to t:

$$H(t) = \int_0^t h(t)\,dt = \int_0^t \frac{t}{\lambda^2}\,dt = \frac{1}{\lambda^2}\int_0^t t\,dt = \frac{1}{\lambda^2}\left[\frac{t^2}{2}\right]_0^t = \frac{t^2}{2\lambda^2} \tag{15}$$

In this PRSM, the variable time t is considered as an explanatory variable, while all others are treated as measured variables. The Table 3 below provides an explanation of the formulae for the HF and CHF, which are used to determine the SF.

Table 3. Functions and their corresponding PRSM, SF, HF, and CHF

Functions	SF $S(t)$	HF $h(t)$	CHF $H(t)$
Weibull	$e^{-\frac{t^k}{\lambda^k}}$	$\frac{kt^{k-1}}{\lambda^k}$	$\frac{t^k}{\lambda^k}$
Exponential	$e^{-\lambda t}$	λ	λt
Log-gaussian	$1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{1}{S(t)\sigma t}\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\Phi\left(\frac{\log t-\mu}{\sigma}\right)$
Gumbel	$e^{-e^{-\frac{t-\mu}{\lambda}}}$	$\frac{1}{\lambda}e^{-\frac{t-\mu}{\lambda}}$	$e^{-\frac{t-\mu}{\lambda}}$
Rayleigh	$e^{-\frac{t^2}{2\lambda^2}}$	$\frac{t}{\lambda^2}$	$\frac{t^2}{2\lambda^2}$

The WD is a flexible model commonly utilized in reliability and survival research due to its flexibility to depict diverse hazard rate patterns. It encompasses numerous well-established distributions as specific or analogous instances. For instance, when the Weibull shape parameter $k = 1$, the distribution reduces to the exponential distribution, which signifies a constant hazard rate throughout time. When $k = 2$, the WD converges to the Rayleigh distribution, which is frequently utilized across multiple domains, characterized by a hazard rate that increases linearly with time. While not a simple conversion, both the Weibull and log-gaussian distributions are proficient in modeling skewed survival data, each representing distinct forms of skewed distributions relative to the Weibull. The WD is theoretically connected to the Gumbel distribution in extreme value theory, as it models minimum extreme values, whereas the Gumbel distribution is generally linked to maximum values. These links underscore the WD versatility and its efficacy in modeling many forms of time-to-event data.

3.6 *Evaluation metrics*

Important metrics for describing the probability of survival over time in survival analysis are MSP and MCSP [25]. While fitting distributions to time-to-event data, both of them are especially helpful. The MSP is the average probability across the whole duration that a person will survive past a given time t. The MSP is determined by the SF $S(t)$, which is the probability of surviving past a given time t.

$$\text{MSP} = \frac{1}{T}\int_0^T S(t)\,dt \tag{16}$$

The probability that a person will live past a specific time is called the MCSP, and it is usually calculated by integrating the survival probabilities across time. It offers a cumulative survival rate for a specific time frame.

$$\text{MCSP} = \frac{1}{T}\int_0^T \left(\int_0^t S(t)\,dt\right) dt \tag{17}$$

where $S(t)$ is the SF.

Both are crucial instruments for survival analysis; they support model comparisons, long-term survival estimations, and evaluations of how well various probability distributions fit the data.

A statistical technique identified as the AD test is used to ascertain whether a given data sample fits into a particular distribution [26]. The Kolmogorov-Smirnov test has been improved, with a greater focus on the distribution's tails. The AD test statistic is defined as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)\left(\ln F(X_i) + \ln(1 - F(X_{n+1-i}))\right)\right] \tag{18}$$

The CvM test is another GoF test that evaluates how well a distribution fits a collection of data. Additionally, it makes a comparison between the sample data's empirical distribution and the cumulative distribution function of a theoretical distribution, such as the AD test. The CvM test statistic is given by:

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[\frac{2i-1}{2n} - F(X_i)\right]^2 \tag{19}$$

One can choose the best distribution for a dataset by comparing these test statistics and p-values, since higher p-values and lower test statistics indicate a better fit. These tests are very useful for evaluating various models used in survival analysis. The overall AIC and BIC calculations are particularly applicable to PRSM [27]. AIC Enter the appropriate log-likelihood calculation into the AIC equation using the number of parameters unique to each distribution. Similar to this, BIC is determined by using a certain log-likelihood formula together with the number of parameters and data. Models can be compared using these requirements in order; to overall, the model with the lower AIC or BIC is chosen.

The AIC equation was:

$$\text{AIC} = -2\log(L) + 2k \tag{20}$$

The BIC equation was:

$$\text{BIC} = -2\log(L) + \log(n)\cdot k \tag{21}$$

where: n is the number of observations; L is the likelihood value; k is the number of estimated parameters.

Particularly in the field of statistical modeling and regression analysis, AIC and BIC are two often used criteria for model selection and complexity penalties. Both seek to assess the GoF of a statistical model while penalizing model complexity in order to prevent overfitting. Certain flexibility and skewness traits are shared by the WD and the Log-gaussian distribution. Though the HF of each distribution are unique, the WD is especially adaptable, enabling it to simulate growing, decreasing, or constant hazard rates that, in some situations, overlap with the Exponential, Rayleigh, and Gumbel distributions.

4. Results and discussion

A thorough evaluation of the $S(t)$, $h(t)$, and $H(t)$ in PRSM is done through R software using the formulae mentioned in Table 3. The Figures 2, 3, and 4 visually represent the functions of survival models, which also highlight the dynamic linkages that the model's equations mandate among time, survival probability, and hazard rates. Plotting these curves helps researchers to analyze and communicate the results of the model by providing a clearer grasp of the time evolution of survival probabilities.

The below Figures 2, 3, and 4 are prepared in order to convey the insights from Table 3 into focus. As shown in Figure 2 the five distinct curves have been obtained by using equations (1, 4, 7, 10), and (13). Each curve begins with a survival probability of 1.0 at time zero and decreases over time, suggesting that the likelihood of survival rate decreases

with increasing period. This graph can be used to compare the estimates of failure or mortality probabilities provided by various statistical techniques. abilities off with increasing time, suggesting a swift initial drop in the likelihood of surviving that gradually slows down.

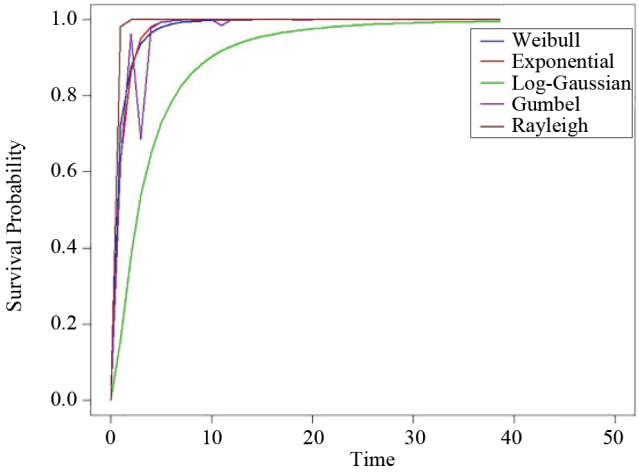

Figure 2. Survival probability of PRSM

Similarly, the results in the Figure 3 have been generated by utilizing the equations (2, 5, 8, 11), and (14). The result in Figure 3 represents the evolution of the HF for each distribution, with initial points on the y-axis that vary throughout time. The WD shows a decreasing hazard rate as time goes on. An Exponential distribution shows a constant rate of risk or an on-going probability of failure over time. A sharply rising hazard rate is indicated by the Gumbel and Rayleigh Distribution. The probability of failure rises with time, according to the Log-gaussian distribution, which shows an increasing hazard rate. With the help of equations (3, 6, 9, 12), and (15), the CHF has been estimated and its results has been shown in Figure 4. Figures 2 and 4 are excellent resources which provide details of survival probability measurements. The MSP values and MCSP across several parametric regression models can be identified and compared using equations (16) and (17). Given that survival probability is as a crucial parameter with a range of 0 to 1, the comparative evaluation made possible by these numbers clarifies which model produces the highest survival probabilities over time as well as overall cumulative survival probabilities.

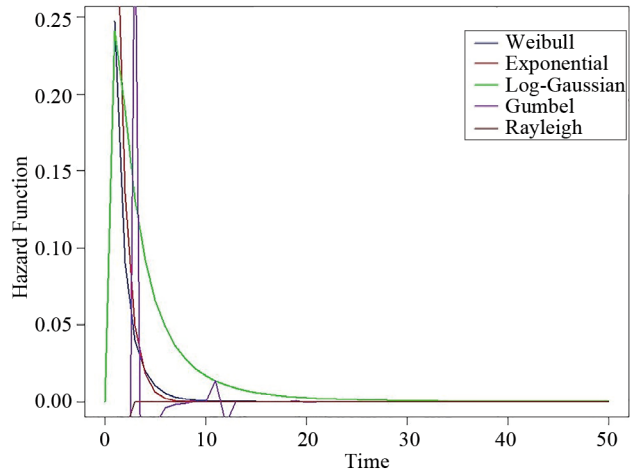

Figure 3. Hazard functions of PRSM

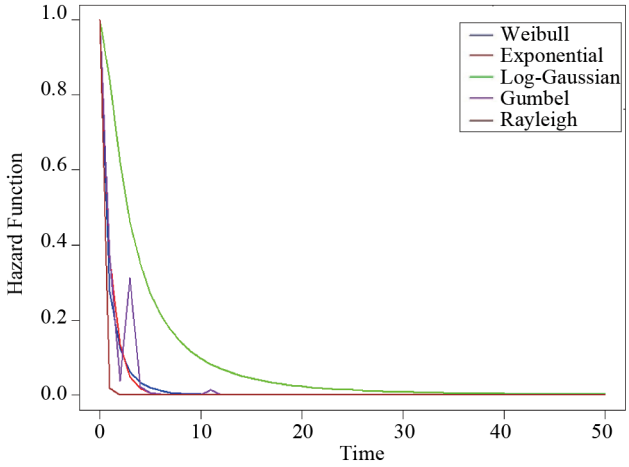

Figure 4. Cumulative hazard functions of PRSM

The Weibull model demonstrates the most significant quality of fit among the models according to the GoF test findings. It displays a high p-value (0.6164) and the lowest AD statistic (0.28745) computed using equation (18), suggesting a good fit to the observations. An additional indication of the Weibull model's adequacy comes from the CvM test, which also shows a low statistic (0.0450) obtained from equation (19) with a high p-value (0.5877). As an illustration of a poor match, other models, such the Gumbel and Log-gaussian, have far lower p-values (0.00513 for the Gumbel and 0.03691 for the Log-Gaussian). With a CvM p-value just below 0.05 (0.04468), the Exponential model exhibits a borderline fit but performs decently. Despite fitting the data reasonably, the Rayleigh model falls short of the Weibull model in terms of performance. By comparison with the other models in this research, the Weibull model offers the best description of the data. In particular, the Weibull model stands out within the parameters of the study with relevant differences. With a MSP value of an astounding 0.9697, it has been represented the value among the population under study, indicating its ability at estimate survival outcomes. Moreover, the Weibull model similarly shows a positive cumulative survival trend across the evaluation period, with a MCSP value of 0.0303.

Table 4. Goodness of fit tests for various PRSM

Models	Goodness of fit tests			
	AD	$p =$ values	CvM	$p =$ values
Weibull	0.28745	0.6164	0.0450	0.5877
Exponential	0.69355	0.06893	0.12912	0.04468
Log-gaussian	0.80298	0.03691	0.15069	0.02303
Gumbel	1.1493	0.00513	0.18989	0.007119
Rayleigh	0.59402	0.1192	0.10102	0.1079

Table 5. Statistical metrics for different survival distributions

Statistical Metrics	Weibull	Exponential	Log-gaussian	Gumbel	Rayleigh
Significant factors	4	3	3	4	6
Mean survival probability	0.9697	0.969	0.9579	0.9653	0.9693
Mean cumulative survival probability	0.0303	0.0310	0.0420	0.0347	0.0199
AIC	1,650.753	1,671.981	1,676.139	1,734.956	1,675.602
BIC	1,623.329	1,644.846	1,648.704	1,707.522	1,648.168

Table 5 elucidates important insights in support of these conclusions. This table includes the MSP values, MCSP values, and the computed AIC and BIC values for each of the five PRSM that are being examined using equations (20) and (21). Researchers can assess and choose the best model based on statistical rigor and predicted accuracy by using the AIC and BIC values, which are quantitative measurements of model performance and complexity. To conclude, a comprehensive understanding of the PRSM under study is provided by the integrated analysis shown in Tables 4, and 5 along with Figures 2, 3, and 5. Finaly, these results lead to improved knowledge and making decisions in the field of survival analysis and prognostication by illuminating the important variables influencing survival results as well as demonstrating the predictive ability and comparative efficacy of various models.

Especially in the frame of advanced-survival lung cancer, the combination of Non-parametric and parametric regression (Weibull, Exponential, Log-gaussian, Gumbel, and Rayleigh) survival models received significant traction in a variety of domains, including biomedical research, clinical trials, epidemiological studies, and health care outcome analysis. In order to determine a correlation association between time and other relevant factors related to lung cancer, the Pearson correlation is applied in this investigation. To ensure resilience and reliability while demonstrating survival processes, parametric models typically rely on the application of the uniform property for convergence. The Weibull model expertly captures skewed survival time distributions and is well-suited for circumstances where the hazard rate increases over time. In contrast, the symmetric form of the Gaussian model provides well-balanced insights into survival dynamics. However, the generalized extreme value theory-based extreme survival model, which focuses on the extreme values' tail distribution, does a better job at simulating uncommon and extreme events. Because each model is chosen according to the particular features of the survival dataset being studied, it offers unique insights and analytical powers.

Ultimately, the Weibull survival model was supported by significant factors like MSP, MCSP, AD, CvM, AIC, and BIC. The proposed work has shown to be the best method for analyzing the lung cancer dataset. The study helps us to emphasize on how well the model works in order to define survival outcomes in advanced-stage lung cancer. The WD adaptability in simulating various HF type makes it especially well-suited for this investigation. Using the fitted Weibull model, create prediction models to assess survival rates for new patients. Clinical judgment and individualized treatment regimens may be guided by it. New studies should capitalize on the advantages of the WD by improving prediction accuracy and comprehending the impact of many factors on survival. However, this study has certain limitations. It solely focused on parametric-based analysis, neglecting to analyze the lung cancer data in relation to nonparametric-based analysis. Furthermore, the study utilized data from a specific NCCT health center only.We can extend the study in the future by collecting data from multiple health centers, potentially leading to effective conclusions on lung cancer.

5. Conclusion

The application of mathematical regression models is crucial for assessing survival data that includes various variables and factors. Five different PRSM that are frequently used to model survival processes are reviewed in this research study. After a comprehensive analysis of the results obtained from the Weibull, Exponential, Log-gaussian, Gumbel, and Rayleigh regression survival models, it is evident that the Weibull model performs better than the other

models based on mathematical relation, statistical metrics and GoF tests. In particular, the MSP values of 0.9697 for the Weibull model are noticeably higher, suggesting a more hopeful outlook for survival outcomes. Subsequently, the Weibull model exhibits an altered MCSP of 0.0303, indicating improved precision in predicting survival probabilities throughout different time spans. Consequently, when compared to the other models, the Weibull model has the lowest AD statistic 0.28745, CvM statistic 0.0450, AIC value 1,650.753, and BIC value 1,623.329 respectively. One of the greatest techniques for advanced lung cancer analysis is the WD, which offers an adaptable structure for comprehending, interpreting, extending, and forecasting events in a variety of fields.This suggests a better fit between model complexity and accuracy, highlighting the Weibull model's effectiveness in gathered the subtleties of survival Conditions in the lung cancer dataset. Because of its outstanding results on a variety of parameters, it is the go to option for practitioners and scholars who want solid insights into the mechanisms involved in survival in this domain.

Acknowledgement

We like to convey our sincere appreciation to the management of Vellore Institute of Technology and the School of Advanced Sciences for their extensive support and resources.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Chan HF, Hsu WH, Chen JP, Lee JH. Factors associated with survival of patients with advanced lung cancer and long travel distances. *Journal of the Formosan Medical Association*. 2024; 123(2): 273-282.

[2] Mustafa M, Azizi AJ, IIIzam E, Nazirah A, Sharifa S, Abbas S. Lung cancer: Risk factors, management, and prognosis. *IOSR Journal of Dental and Medical Sciences*. 2016; 15(10): 94-101.

[3] Leiter A, Veluswamy RR, Wisnivesky JP. The global burden of lung cancer: Current status and future trends. *Nature Reviews Clinical Oncology*. 2023; 20(9): 624-639.

[4] Pradhan KS, Chawla P, Tiwari R. HRDEL: High ranking deep ensemble learning-based lung cancer diagnosis model. *Expert Systems with Applications*. 2023; 213(8): 118956.

[5] Altuhaifa FA, Win KT, Su G. Predicting lung cancer survival based on clinical data using machine learning: A review. *Computers in Biology and Medicine*. 2023; 165(1): 107338.

[6] Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B, et al. Lung cancer immunotherapy: Progress, pitfalls, and promises. *Molecular Cancer*. 2023; 22(1): 40.

[7] Amicizia D, Piazza MF, Marchini F, Astengo M, Grammatico F, Battaglini A, et al. Systematic review of lung cancer screening: Advancements and strategies for implementation. *Healthcare*. 2023; 11(14): 2085.

[8] Torre LA, Siegel RL, Jemal A. Lung cancer statistics. In: *Lung Cancer and Personalized Medicine: Current Knowledge and Therapies*. Heidelberg: Springer; 2016.

[9] Geremew H, Dessie AM, Anley DT, Feleke SF, Geremew D. Tuberculosis and its associated risk factors among HIV-positive pregnant women in northwest Ethiopia: A retrospective follow-up study. *Heliyon*. 2023; 9(11): e21382.

[10] Wong MSH, Pons A, De Sousa P, Proli C, Jordan S, Begum S, et al. Determining the optimal time to report mortality after lobectomy for lung cancer: An analysis of the time-varying risk of death. *JTCVS Open*. 2023; 16: 931-937. Available from: https://doi.org/10.1016/j.xjon.2023.08.009.

[11] Everest L, Blommaert S, Chu RW, Chan KK, Parmar A. Parametric survival extrapolation of early survival data in economic analyses: A comparison of projected versus observed updated survival. *Value in Health*. 2022; 25(4): 622-629.

[12] Cislo PR, Emir B, Cabrera J, Li B, Alemayehu D. Finite mixture models, a flexible alternative to standard modeling techniques for extrapolated mean survival times needed for cost-effectiveness analyses. *Value in Health*. 2021; 24(11): 1643-1650.

[13] Wulandari I, Kurnia A, Sadik K. Weibull regression and stratified cox regression in modelling exclusive breastfeeding duration. *Journal of Physics: Conference Series*. 2021; 1940(1): 012001.

[14] Sato H, Goto T, Hayashi A, Kawabata H, Okada T, Takauji S, et al. Prognostic significance of skeletal muscle decrease in unresectable pancreatic cancer: Survival analysis using the weibull exponential distribution model. *Pancreatology*. 2021; 21(5): 892-902.

[15] Almongy HM, Almetwally EM, Aljohani HM, Alghamdi AS, Hafez EH. A new extended rayleigh distribution with applications of COVID-19 data. *Results in Physics*. 2021; 23(60): 104012.

[16] Ahmed LA. Parametric models in survival analysis for lung cancer patients. *Ibn AL-Haitham Journal for Pure and Applied Sciences*. 2021; 34(2): 108-118.

[17] Wegbom AI, Kiri VA, Essi ID. Comparison between semi-parametric cox and parametric survival models in estimating the determinants of under-five mortality in Nigeria: Application in Nigerian demographic and health survey. *African Journal of Mathematics and Statistics Studies*. 2019; 2(2): 1-12.

[18] Jamil SA, Abdullah MA, Kek SL, Olaniran OR, Amran SE. Simulation of parametric model towards the fixed covariate of right censored lung cancer data. *Journal of Physics Conference Series* 2017; 890(1): 012172.

[19] Wang K, Liu X, Pan Y, Owusu D, Xu C. Comparison of cox regression and parametric models for survival analysis of genetic variants in Hnf1b gene related to age at onset of cancer. *Journal of Data Science*. 2017; 15(3): 423-442.

[20] Sultana J, Jilani AK. Predicting breast cancer using logistic regression and multi-class classifiers. *International Journal of Engineering and Technology*. 2018; 7(4.20): 22-26.

[21] Zhu H, You X, Liu S. Multiple ant colony optimization based on pearson correlation coefficient. *IEEE Access*. 2019; 7: 61628-61638. Available from: https://doi.org/10.1109/ACCESS.2019.2915673.

[22] Csalódi R, Bagyura Z, Abonyi J. Mixture of survival analysis models-cluster-weighted Weibull distributions. *IEEE Access*. 2021; 9: 152288-152299. Available from: https://doi.org/10.1109/ACCESS.2021.3127576.

[23] Nadler DL, Zurbenko IG. Developing a weibull model extension to estimate cancer latency. *International Scholarly Research Notices Epidemiology*. Egypt: Hindawi Publishing Corporation; 2013.

[24] Palm BG, Bayer FM, Cintra RJ, Pettersson MI, Machado R. Rayleigh regression model for ground type detection in SAR imagery. *IEEE Geoscience and Remote Sensing Letters*. 2019; 16(10): 1660-1664.

[25] Mozumder SI, Rutherford MJ, Lambert PC. Estimating restricted mean survival time and expected life-years lost in the presence of competing risks within flexible parametric survival models. *BMC Medical Research Methodology*. 2021; 21: 1-20. Available from: https://doi.org/10.21203/rs.2.23839/v1.

[26] Berlinger M, Kolling S, Schneider J. A generalized Anderson−Darling test for the goodness-of-fit evaluation of the fracture strain distribution of acrylic glass. *Glass Structures and Engineering*. 2021; 6(3-4): 195-208.

[27] Amran SE, Abdullah MA, Kek SL, Jamil SA. Analysis of survival in breast cancer patients by using different parametric models. *Journal of Physics: Conference Series*. 2017; 890(1): 012169.