MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

A Novel Feature Selection Method for Ultra High Dimensional Survival Data

Nahid Salma[a,b], Ali Hussain Mohammed Al-Rammahi[a], Majid Khan Majahar Ali[a*]

[a]School of Mathematical Sciences, Universiti Sains Malaysia, Penang, Malaysia; [b]Department of Statistics and Data Science, Jahangirnagar University, Savar, Dhaka, Bangladesh-1342

Abstract Finding relevant features in ultra-high dimensional survival data is one of the most important and fundamental objectives in biology discovery and statistical acquisition. Conventional survival regression algorithms are challenged by the exponential increase in raw data. In real-world scenarios, data processing with ultra-high dimensionality has an impact, particularly on two-component structures like the kidneys, lungs, and eyes. Future system stability and the frequency of illness are both affected by gene interactions between two components. The traditional statistical procedures employed for the survival system are restricted to single component. To date, for ultra-high-dimensional survival data with two compartments, no feature selection method is available. Thus, with the goal to determine the optimal methods in this situation, this study suggested and contrasted the performance of ten variable selection approaches for ultra-high dimensional Renal Cell Carcinoma (RCC) survival data containing two compartments. The study attempted to combine Freund's baseline hazard function as the baseline hazard of Cox model (Lasso Freund, Robust Lasso Freund, Elastic Net Freund) and integrated with sure independence screening (SIS) and iterative sure independence screening (ISIS) (i.e., LF-SIS, RLF-SIS, ENF-SIS, LF-ISIS, RLF-ISIS, ENF-ISIS) in an attempt to tackle this issue. Additionally, two basic approaches, LASSO and EN, were taken into consideration and EN is combined with SIS and ISIS (EN-SIS, EN-ISIS). Result shows that based on the validating model measures, including MSE (340.000), SSE (25300.0) and RMSE (16.490) suggest, the Robust Lasso Freund-Iterative Sure Independence Screening (RLF-ISIS) and Robust Lasso Freund-Sure Independence Screening (RLF-SIS) strategy performs superior to the other suggested approaches in terms of greater precision in picking variables. Though both methods showed lower R^2 (0.71) which advocates the presence of the outliers in the dataset. Additionally, the box-plot of some selected predictive genes confirms the presence of outliers. Furthermore, two methods, RLF-ISIS and RLF-SIS, have been used to identify 49 and 68 genes that have both direct and indirect effects on patients with RCC. Finally, it can be concluded that although RLF-SIS and RLF-ISIS outperform other proposed approaches, they may, however, be regarded as a variable selection strategy but they might not be the optimal choice for ultra-high dimensional survival data with outliers. Nevertheless, the study can be expanded in the future by applying competitive risk theory to a sequential and parallel structure, which serves as the basis for most complex mechanical systems found in manufacturing facilities. Notably, no feature selection method is available for ultra-high-dimensional survival data with outliers and two-compartments. Therefore, to address this particular issue, further research should focus on developing an advanced hybrid feature selection approach, with a particular emphasis on deep learning strategies.

Keywords: Ultra-high dimension, renal cell carcinoma, cox model, freund model, feature selection.

***For correspondence:**
majidkhanmajaharali@usm.my

Received: 27 June 2024
Accepted: 09 Sept. 2024

Introduction

As a result of high-throughput technology, more and more High-Dimensional or Ultra-High-Dimensional (UHD) data is being manufactured related to genomics [46, 25]. The ultimate purpose of such massive genomics data is to improve understanding of the biological, environmental, and behavioral factors that contribute to a disease. Individual differences in lifestyle, environment, and genes will be taken into consideration in treatment as well as prevention of diseases [11]. The precision medicine projects revolve around this central idea [28, 3]. High-throughput variables or predictors are frequently gathered for

10.11113/mjfas.v20n5.3665

survival outcomes in the modern age of precision medicine. However, the key trends and extensive biological facts hidden within the data sometimes remain unresolved [32]. Hence, efficient computational techniques are necessary for mining such unusually huge data.

Survival analysis in statistics uses survival time as a random variable to represent a qualitative shift from one condition to another, i.e., alive to dead [38]. Healthcare survival data often face three distinct problems: high dimensionality (the sample size (*n*) is generally much smaller than the number of predictors (*p*), for instance gene characteristics (i.e., *p*>>*n*) or ultra-high-dimensionality (*p* is incredibly huge, for example *p* > 10^5), data censoring (i.e., inadequately recorded time-to-event), and breaking the assumption of proportional hazard [25, 9]. However, as the amount of data increases, data contain a large amount of unnecessary, distorted, insignificant, and insufficient information [46]. Hence, as the number of model parameters increases, the lack of observation in relation to the number of variables leads to incorrect estimates [52]. Also, Excessive processing of precise projections results in low accuracy (due to multicollinearity). As noted by Hampel *et al*. [17], a genuine dataset might consist of one to ten percent outliers. Outliers are unexpected observations in a dataset that differ from the majority, and the model performs poorly in the presence of an outlier [22]. Thereafter, these ultra-high-dimensional data present a challenge to conventional survival regression techniques, which are either impossible to adapt or are likely to have poor conformity as a result of overfitting [57]. Although, survival data have been analyzed using conventional statistical techniques including Kaplan-Meier and then Cox proportional hazard (CPH) frameworks [40, 7, 21] but Cox model has been developed for small data sets, therefore, cannot handle ultra or high-dimensional data [7, 45]. Thus, the recent focus has been on creating original methods for choosing risk factors and survival prediction.

In such situations, Machine Learning (ML) has provided scientist different way to investigate the complicated relationships of variables and hazard prediction [20]. A key idea in machine learning is feature (or risk factors) selection, which is crucial for reducing overfitting and enhancing model effectiveness [32]. It is not new to employ machine learning algorithms to select important risk factors and forecast the likelihood of cancer patients dying using clinical data [34, 35, 23, 30, 36, 56]. A number of regularization techniques are offered for high-dimensional variable picking, incorporating the Dantzig selector [6, 13], the LASSO [43], the Smoothly Clipped Absolute deviation (SCAD) [17], the adaptive LASSO [48], as well as the minimax concave penalty [52] and so on. Owing to major constraints such as when p > > n, it has trouble selecting more predictive variables than instances and it often chooses just one predictor from a bunch of associated predictors [74], LASSO is no longer practical for use in UHD analysis. Yet, as ultra-high dimensionality encompasses simultaneous issues of computational simplicity, statistical precision, and algorithmic rigidity, the penalized approaches such as SCAD, Adaptive LASSO could not work satisfactorily because of extremely significant number of predictors [16]. An enhanced variant of LASSO for managing large correlations is the Elastic Net (EN) approach, which was implemented out by Zou and Hastie [49]. However, EN requires model tuning, which can make it operationally costly to use on huge data sets. Additionally, since there is no set method for selecting tuning parameters (λ_1 and λ_2), it can be difficult to make a decision [8]. This led to restrictions on EN use in UHD.

Apart from the techniques outlined earlier, scholarly works suggest other feature screening methodologies, focusing on reducing dimensionality by incorporating predictors strongly related to outcomes. The sure independence screening (SIS) approach was suggested by [18] which is based on an original sure screening notion. It ranks the significance of every candidate factor based on its marginal Pearson associations with the outcome. While features are ranked according to their marginal utility, SIS may have some problems related to independence learning. First, compared to other significant covariates, a few unimportant covariates that share a strong correlation with the essential ones may have greater marginal utility. Secondly, following the screening stage, certain significant factors which are jointly connected but slightly uncorrelated with the outcome may be overlooked [18]. With the objective to tackle these problems, [18] also presented the iterative SIS, an expansion of the SIS methodology. The basic concept is to use SIS to continually revise the calculated set of key variables conditioned on the projected set of factors given the preceding phase [8].

Beside the feature selection problem in UHD, in real-world scenarios, data processing with ultra-high dimensionality has an impact, particularly on two-component structures like the kidneys, lungs, and eyes. Future system stability and the frequency of illness are both affected by gene interactions between two components. In more detail, when a patient loses one kidney due to cancer, disease, or large stone, the remaining kidney shows a higher failure rate, affecting the remaining normal function. This phenomenon also occurs in the lungs and ears. The current statistical procedures employed for the survival systems are restricted to single component systems. Furthermore, high-dimensional problems presented significant obstacles for the older statistical approaches for variable selection, for instance stepwise regression, all subsets regression, and ridge regression, necessitated the use of very sophisticated statistical approach [1, 50]. Thus, in the recent past, computational scientists and statisticians have made phenomenal efforts to design and construct completely novel techniques for modeling and variable selection with regard to the emerging issues. Nevertheless, no generally recognized model has been created for the efficient handling of ultra-high dimensional survival data that includes two-component systems. Thus, with the goal to determine the optimal methods in this situation, this study suggested and contrasted the performance of ten variable selection approaches for ultra-high dimensional renal cell carcinoma (RCC) survival data containing two-compartments. Freund's (1961) [21] model is a crucial and efficient model for life testing in biological systems with dual chemicals. Where and the well-known Cox’s proportional hazards model [7] is the frequently utilized framework for analyzing right-censored survival data. An exponential distribution with one parameter defines the baseline hazard function of Cox proportional hazard model. Unluckily, structures with two or more components cannot be handled by it. Therefore, Freund's baseline hazard function in parallel two-component or multi-component systems can be utilized to solve this problem. Thus, the study attempted to combine Freund’s baseline hazard function as the baseline hazard of Cox model (Lasso Freund, Robust Lasso Freund, Elastic Net Freund) and integrated with sure independence screening (SIS) and iterative sure independence screening (ISIS) (i.e., LF-SIS, RLF-SIS, ENF-SIS, LF-ISIS, RLF-ISIS, ENF-ISIS) in an attempt to tackle this issue. Additionally, two basic approaches, LASSO and EN, were taken into consideration and EN is combined with SIS and ISIS (EN-SIS, EN-ISIS). Therefore, a total of ten techniques were evaluated and tested for ultra-high dimensional renal cell carcinoma (RCC) survival data. The renal cell carcinoma (RCC), a critical health condition in the kidneys is made up of several different kinds of kidney tumors. Nearly 90% of all kidney malignancies are RCC [4]. To date, three forms of kidney cancers are available worldwide [42]. Kidney renal clear cell carcinoma otherwise known as clear cell RCC (KIRC or ccRCC), which ccounts for approximately 70–75% of all renal cancers.

Related Works

Numerous techniques described in the scientific literature can be used to pick features. An overview of the pertinent studies on these techniques is given in this section (Table 1). The first two applications of least squares with penalty were LASSO [43] and ridge regression [70]. First introduced by [49], the EN regression is an expanded variant of penalized regression that performs better in variable selection. However, there are certain drawbacks of LASSO, EN and Ridge regression, such as algorithms stability and computational simplicity [18, 69]. Consequently, using an embedded approach in conjunction with a filter as well wrapper is beneficial, particularly in cases whenever the number of characteristics greatly outweighs the extent of the data collection [69].

The research [61] Id a generic procedure called concordance index screening (CI-SIS) to handle categorical outcome in data with extremely high dimensions. Test findings indicate that the suggested approach may successfully discover genes linked to certain illness. After analyzing four genuine high-dimensional datasets, the authors of the paper [62] offered Deep Feature Screening (DeepFS), which performed better than ISIS, Plasso, CAE, FsNet, LassoNet, TSFS, and other alternatives. Author [63] advocated Sparse support vector machines with L_0 approximation, taking into account datasets on ovarian cancer. After removing the superfluous qualities, a unique feature selection technique (DRPT) was proposed by [64]. This technique discovers correlations among the remaining features. Author [2] examined data on kidney cancer and suggested LASSO-Freund in a two-component concurrent configuration to extend the Cox PH-based iterative sure independence screening. Random bits forest recursive clustering elimination (RBF-RCE), SVM-RCE, and RF were compared by the author [66]. However, the study utilized High Dimensional data, not Ultra High Dimensional data**.** Authors [40, 67, 68] have also suggested feature selection techniques for HD data. In order to differentiate among early-phage and late-stage ccRCC, author [71] employed RNAseq expression data out of The Cancer Genome Atlas (TCGA) project in Kidney Renal Clear Cell Carcinoma (KIRC) patients. With the help of an apprised Korean Renal Cell Carcinoma (KORCC) record that included information on 10,068 individuals who had undergone an operation for customized renal cell carcinoma (RCC), the researchers [72] created an original forecasting model for survival and recurrence in patients with RCC after surgery. The feature selection process was executed using an elastic net after data pre-processing. Though author [71, 72] investigated renal cell carcinoma data but they did not consider ultra-high dimensional data.

To the extent of our expertise, there are no hybrid feature selection methods for ultra-high dimensional RCC survival data that includes parallel two-component systems. Thus, more study in this field is required.

Table 1. Summary of the previous study

SL No	Author's	Data Source	Number of compartments	Methods	Findings
1	[43]	Prostate cancer data	1	Compared among Ridge regression, garotte and LASSO	Suggest a novel strategy named LASSO to estimating linear models
2	[49]	Prostate cancer data	1	Compared among OLS, Ridge Regression, LASSO, Naïve Elastic Net and EN	Provide a novel approach to selection of variables as well as regularization called the elastic net.
3	[61]	Lung Dataset	1	LDA, RF, CI-SIS	Create a generic concordance index screening (CI-SIS) process to manage categorical response in ultra-high dimensional data. The suggested approach can effectively identify genes linked to certain diseases, according to experimental findings.
4.	[62]	Four large-scale datasets with low sample sizes and high dimensions (Colon, Leukemia, B-cell chronic lymphocytic leukemia and Prostate Cancer	1	Deep Feature Screening (DeepFS)	DeepFS outperformed ISIS, Plasso, CAE, FsNet, LassoNet, TSFS, and other options.
5.	[63]	Obese, Ovarian cancer	2	Advised Sparse support vector machines with L_0 approximation	Advised method outperformed.
6.	[64]	Gene Expression Data	1	Proposed a novel feature selection technique (DRPT)	The author presents a novel feature selection technique (DRPT) that finds correlations amongst all remaining characteristics after eliminating the unnecessary ones.
7.	[65]	4 real HD datasets (*Riboflavin, Eyedata, Boston Housing and Longley)*	1	This work presents a novel blended feature selection method that combines Elastic Net regularized regression (K-EN) with feature filtering using Kendall's tau.	Proposed method outperformed
8.	[39]	Prostate cancer and Lung Cancer	1	Offer the LASSO along with SCAD R packages SIS and ISIS, which integrate in the CoxPH model.	Offered method works well.
9	[2]	Kidney cancer	2	LASSOFreund is used within a two-component concurrent setup to expand the iterative sure independence screening utilizing Cox PH.	Recommended method performed well.
10	[66]	High-dimensional data (colon, prostate cancer, DLBCL, GLI)	1	Compared random bits forest recursive clustering elimination (RBF-RCE), SVM-RCE, RF.	RBF-RCE is best.

SL No	Author's	Data Source	Number of compartments	Methods	Findings
11.	[67]	Prostate cancer	1		In light of the MCAR and MAR missing processes, the study suggested a screening process as per the ML of the linear correlation coefficient.
12.	[68]	Results of an investigation on breast cancer concerning survival	2	Proposed (Matched Case-Control Logistic Regression), PL-Cox	Proposed method outperformed.
13.	[40]	HD data of dementia		Both filter and wrapper Methods	Random Forest Minimal Depth algorithm outperforms.
14.	[71]	Clear cell renal cell carcinoma	2	SVM, LR, NB, MLP, RF	SVM outperformed.
15.	[72]	The Korean Renal Cell Carcinoma (KORCC)	2	EN	Developed a novel predicting technique to forecast the survival and relapse of RCC patients following surgery.

Materials and Methods

To accomplish our aim, we carefully developed a study plan that we strictly followed, displayed in Figure 1. Shortly, with the aim to develop a hybrid model for feature selection of UHD data, the study started with combining Freund's baseline hazard function as the baseline hazard of Cox model. The UHD RCC data was then pre-processed and partitioned into 70:30 ratio where 70% was used as training and rest 30% was considered to validate all techniques. A total of 10 ML feature selection approaches then run on the UHD data including two traditional feature selection methods: LASSO and EN and eight proposed feature selection approach. The proposed approaches were divided into two groups where in one, SIS was integrated with LF, EN, ENF and RLF: LF-SIS, EN-SIS, ENF-SIS, RLF-SIS and in another group, ISIS was combined with LF, EN, ENF and RLF produces LF-ISIS, EN-ISIS, ENF-ISIS, RLF-ISIS. The performance of all 10 feature selection methods was evaluated based on SSE, MSE, RMSE and R^2 . To further investigate the cause of the low R^2 , a boxplot analysis was used to examine any outliers in the dataset. Finally, the optimal feature selection technique was selected and employed to extract the significant features/genes associated with RCC.

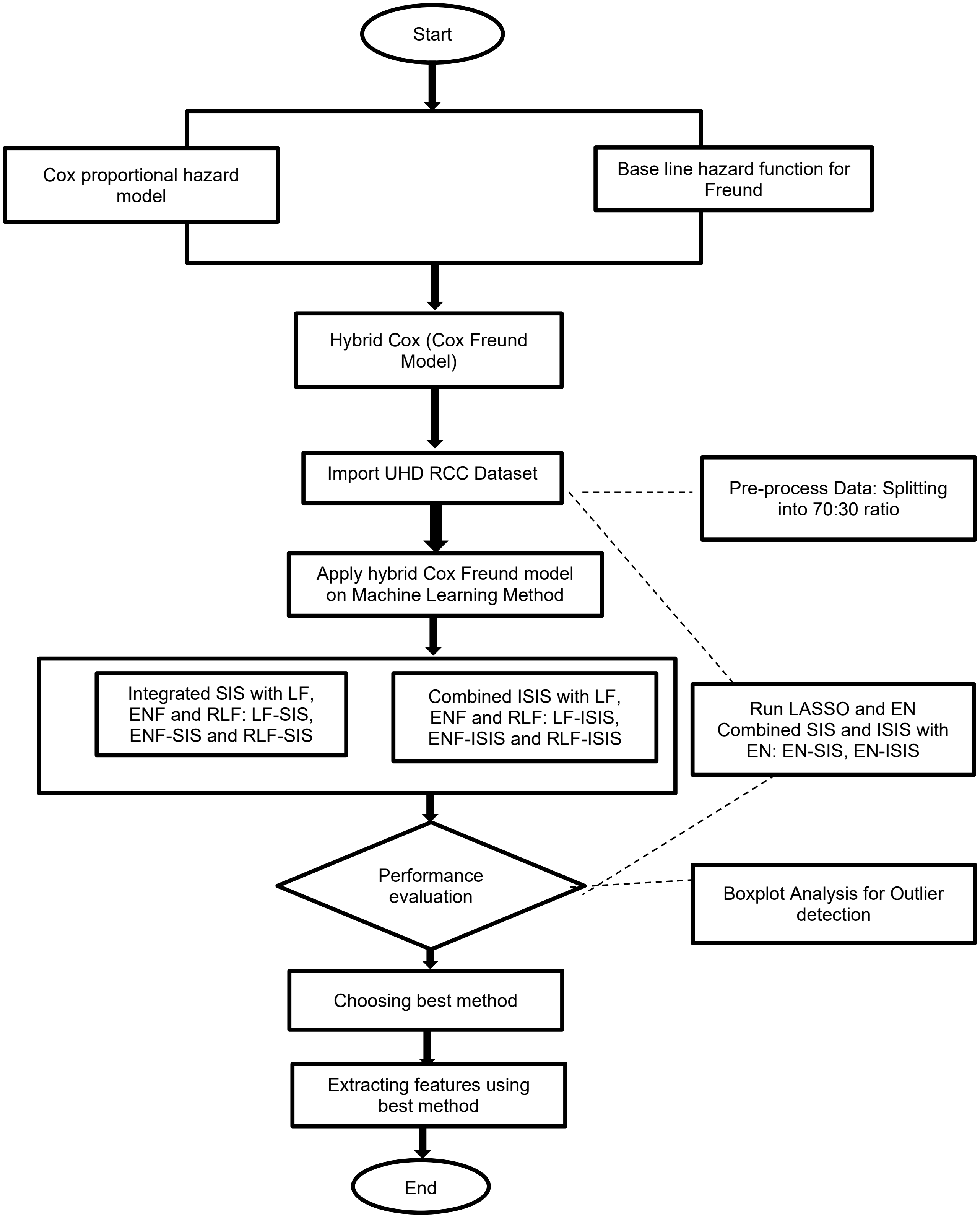
Start
Cox proportional hazard model
Base line hazard function for Freund
Hybrid Cox (Cox Freund Model)
Import UHD RCC Dataset
Pre-process Data: Splitting into 70:30 ratio
Apply hybrid Cox Freund model on Machine Learning Method
Integrated SIS with LF, ENF and RLF: LF-SIS, ENF-SIS and RLF-SIS
Combined ISIS with LF, ENF and RLF: LF-ISIS, ENF-ISIS and RLF-ISIS
Run LASSO and EN Combined SIS and ISIS with EN: EN-SIS, EN-ISIS
Performance evaluation
Boxplot Analysis for Outlier detection
Choosing best method
Extracting features using best method
End

Figure 1. Flowchart of the study's overall methodology

Data Description

From R package 'kidpack', the gene expression data was collected. It is a data frame with 4224 rows of Differentially Expressed Genes (DEG) for 74 people whom it believes may have kidney cancer. Since the dataset contains huge number of predictors (DEG), the dataset can be considered as the ultra-high dimensional data. A total of 74 kidney tumor samples with varying histological types, differentiation grades, stages, and chromosomal abnormalities and follow-up data were included in the study. Patients' survival status, survival rates are also in the dataset. Samples were hybridized using a shared reference that was created by combining several kidney tumor samples. The overall information about the dataset can be found and the data can also be downloaded from this link: E-DKFZ-1 < ArrayExpress < BioStudies < EMBL-EBI. The outcome variable of this study is survival hazard, whether a patient death = 1 or alive = 0 during the study period and the predictors are 4224 genes.

Data Processing

The dataset was partitioned into two parts, 70% was used as training and rest 30% was considered to validate all techniques.

Cox Proportional Hazard Model

Cox's proportional hazards model [7] is the frequently utilized framework for analyzing right-censored survival data. It uses an independent lifetime distribution with a predetermined hazard function. The hazard function, which represents the likelihood that an occurrence will happen at time *t*, is how the Cox model is stated can be written as:

$$h(t) = h_0(t) \times \exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p) \text{ or}$$
$$h(t) = h_0(t) \times \exp(x^T \beta) \qquad (1)$$

Here, *t* serve as the time of survival, $h(t)$ represents the function of hazard and $\{x_1, x_2, ..., x_p\}$ considered as *p* covariates values whereas coefficients are denoted as $\{\beta_1, \beta_2, ..., \beta_p\}$ which calculate the impact of explanatory factors on the time to survival and $h_0(t)$ is the unknown baseline hazard function [26, 40]. The unknown parameters are calculated from partial likelihood by maximizing it.

Freund Model

One of the most important and useful models for applications and life testing in biological systems with dual chemicals is Freund's (1961) model [21]. The Freund thought that two identical components, for instance the eyes, ears, kidneys, lungs, etc., would be distributed together in a single system. Till a single of the two components—referred to as the parallel system—remains, the system will keep running. Freund's main assumption is that the rates of danger for the two components are equal. However, if one component fails, the risk rating of the others will go up and stay at that level unless the whole thing is turned off altogether. Additionally, consider two distinct arbitrary variables *(X, Y)* that are distributed exponentially and represent the lifetimes of the two components *(A, B). X** denotes the component (A)'s survival time in the event that component (B) is substituted with an element of the same type each time it fails, and *Y** denotes the component (B)'s duration of stay in the event that component (A) is replaced with an element of the same type each time it fails (Freund, 1961). Since Freund's model is a double exponential model, the following is an expression for his pdf:

$$f(x,y) = \begin{cases} A_1 B_2 \exp\{-B_2 y - (A_1 + B_1 - B_2)x\}, & 0 < x < y \\ B_2 A_2 \exp\{-A_2 x - (A_1 + B_1 - A_2)y\}, & 0 < y < x \end{cases} \qquad (2)$$

For all values of *x, y, A_1, A_2, B_1,* and B_2 greater than *0,* For the survival time of parts 1 and 2, equation 1 uses x and *y* as arbitrary variables. An exponential distribution with parameters A_1 and B_1, respectively, is supposed to be followed by the random variables *x* and *y.*

The baseline hazard function for Freund's with both components is formatted as thereafter:

$$Fh_0 = \begin{cases} Fh_{01} \\ Fh_{02} \end{cases} \qquad (3)$$

Where Fh_0 is represented Freund's baseline hazard function.

$$Fh_{01} = \left\{ \frac{\begin{bmatrix} a_2\{\frac{1}{2}(3a_1 + a_2 - a_3)\}e^{-\frac{1}{2}(3a_1+a_2-a_3)t} \\ +\frac{1}{2}(a_1 - a_2)(a_1 - a_3)e^{-\frac{1}{2}(a_1-a_2)t} \end{bmatrix} a_1 a_3 a_4}{\begin{matrix} -\frac{1}{2}(a_1 + a_2)a_1 a_3 e^{-(a_1-a_4)t} + \frac{1}{2}(a_1 - a_2)a_1 a_4 e^{-(a_1-a_2)t} \\ -\frac{1}{2}(a_1 + a_2)(a_1 - a_4)a_3 e^{-a_2 t} - \frac{1}{2}(a_1 - a_2)(a_1 - a_3)a_4 e^{-a_2 t} \end{matrix}} \right.$$

where Freund's baseline hazard function for kidney one is represented by the symbol Fh_{01} [1,2]

$$Fh_{02} = \frac{(a_4 - a_3)(\frac{3}{2}a_1 + \frac{1}{2}a_2 - a_3)e^{-(\frac{3}{2}a_1+\frac{1}{2}a_2-a_3)t+(a_1+a_4)[\frac{1}{2}(a_1+a_2)]e^{-(a_1-a_4)t}a_1 a_3 a_4}}{\frac{1}{2}(a_1 + a_2) - (a_3 - a_4)\begin{bmatrix} \frac{1}{2}(a_1 + a_2)]a_1 a_3 e^{-(a_1-a_4)t} + \frac{1}{2}(a_3 - a_4)a_1 a_4 e^{-\frac{1}{2}(a_1+a_2)t} \\ -\frac{1}{2}(a_1 + a_2)(a_1 - a_4)a_3 - (a_3 - a_4)(a_1 - a_3)a_4 e^{-a_1 t} \end{bmatrix}}$$

where, Fh_{02} refers to Freund's baseline hazard function for kidney tow [1, 2].

Cox Proportional Hazard Models by Freund Model (Hybrid Cox Regression)

In equation 1, after combining Freund's baseline hazard function as the baseline hazard of Cox model, we get the following

$$h(t) = Fh_0(t) \times \exp(x^T \beta)$$

$Fh_0(t)$ is the baseline hazard function of Freund in two-component systems.

Feature Extraction Methods

In real-world scenarios, data processing with ultra-high dimensionality has an impact, particularly on multi-component structures like the kidneys, lungs, and eyes. Nevertheless, no generally recognized model has been created for the efficient handling of ultra-high dimensional survival data that includes concurrent multi-component systems. With the goal to determine the optimal methods in this situation, this study suggested and contrasted the performance of ten variable picking techniques for ultra-high dimensional survival data with many compartments. The approaches can be divided into 2 distinct categories: conventional methods and proposed methods. Among conventional, LASSO and Elastic Net (EN) was used. The proposed methods are: Lasso Freund-Sure Independence Screening (LF-SIS), Robust Lasso Freund-Sure Independence Screening (RLF-SIS), Elastic Net Freund- Sure Independence Screening (ENF-SIS), Elastic Net- Sure Independence Screening (EN-SIS), LASSO Freund-Iterative Sure Independence Screening (LF-ISIS), Robust Lasso Freund- Iterative Sure Independence Screening (RLF-ISIS), Elastic Net-Iterative (EN-ISIS), and Elastic Net Freund-Iterative (ENF-ISIS).

LASSO

The data were examined by the Least Absolute Shrinkage and Selection Operator (LASSO) with the goal to execute cross-validation analysis using the L1 norm penalty function in order to find key variables. [43, 33]. In light of their favorable qualities for picking variables and regularization, shrinkage approaches are gaining appeal in biosystems during the massive data age. By boosting log-partial likelihood and adjusting the tuning value to manage punishment parameter, the LASSO approach calculates coefficients. Lasso's overall expression is easily expressed in the following way:

Assume that $(1x_i, y_i)$, $i = 1,\ldots,N$ represents a sample of N randomly distributed vectors that are Independent and Identically Distributed (IID). $1x_i \in R^p$ where $y_i \in R$ indicates the matching response vector and $x_i = (x_{i1,} x_{i2}, \ldots, x_{ip})$ denotes the row vector of observations regarding the p-explanatory variables of the i^{th} unit of sample. Then the general form of LASSO estimator is as:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \in R^p} \left[\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right]$$

where λ is represents the punishment parameter as well as the Lagrange multiplier. The matrix from of the above equation can be written like this:

$$\hat{\beta}_{LASSO} = \arg\min_{b \in R^p} \left[\frac{1}{N} \sum_{i=1}^{N} \left\| Y - X\beta \right\|_2^2 + \lambda \sum_{j=1}^{p} \left| \beta \right|_1 \right]$$

Where Y points to the column vectors $(n \times 1)$ of the output findings, X represents the matrix $(n \times p)$ containing the relevant variables found of interest, and $\|.\|_1, \|.\|_2$ denotes the L1, L2 vector norms respectively.

Elastic Net

Even while LASSO works well for a variety of variable selection issues, it becomes ineffective when there are significantly more covariates (p) than there are samples [8]. Where the number of predictor variables selected cannot exceed the sample size limit. Furthermore, there exist strong relationships among various sets of variables. An upgraded version of LASSO was originally suggested by [49] for handling strong correlations: the Elastic Net (EN) approach. To increase accuracy in forecasting, the EN employs compensation periods L1-LASSO and L2-right, automatically recognizes the variables, and carries out continual shrinking. This method functions similarly to a stretchy fishing net, preserving all the larger fish (important covariates) while eliminating unimportant factors. Here is how the Elastic Net Estimator is defined.

$$J(\beta, \lambda_1, \lambda_2) = \sum_{j=1}^{p} \left[\lambda_1 \left| \beta_j \right| + \lambda_2 \beta_j^2 \right]$$

The equation involves the Lasso penalty (the first part of the above equation) for sparse variables, and the Ridge penalty (the second part of the above equation) for highly correlated features, promoting the average computation. Specifically, any linear model, particularly the regression or classification approaches, can be utilized with the elastic net penalty [40].

$$\hat{\beta}^{elastic\,net} = (1 + \lambda_2) \arg\min_{\beta \in R^p} \left[\frac{1}{N} \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} \left| \beta_j^2 \right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right]$$

The elastic net's parameters, λ1 and λ2, require an ideal ratio rather than a single parameter, often resulting in a parameter sum of both. The elastic-net penalty is used to minimize the regression loss function in order to get the elastic-net coefficient estimate.

$$\sum_{j=1}^{p} \left[\alpha \left| \beta_j \right| + (1 - \alpha) \beta_j^2 \right] \le \kappa$$

where κ denotes the extra α parameter, which is the total of λ_1 and λ_2 .

Sure Independence Screening (SIS)

For data with high dimension, SIS is a two-stage method for choosing significant covariates [39, 18, 37, 55]. There are two stages to this technique, which are as follows: Phase one: The initial screening phase where the primary effects are analyzed in a rough way using marginal utilities. The second step is the selection stage, where a penalized regression with LASSO penalty is used for both estimating parameters and pick variables. The steps of SIS are as follows:

1. Let's begin by say the data collection $\{(x_i, y_i), i = 1, \ldots, n\}$ has a sample size of $n \text{ and } x \in R$.

   The covariate-specific marginal utility $X_i, i = 1, 2, \ldots, p$ can be computed as follows.

   $$L_m = \min_{\beta_0, \beta_j} \frac{1}{n} \sum L(y_i \beta_0 + x_m \beta_m)$$

   where a generic loss function is represented by L (.,.). More obviously, compute the p marginal utilities by fitting p bivariate models, such as the generalized linear model (GLM).

2. The partial likelihood of every parameter is maximized as follows to determine the utility.

   $$U_m = \max_{\beta_m} (\sum_{i=I} \delta x_{im} \beta_m - \sum_{i=I} \delta_i \log\{ \sum_{j \in R(y_i)} \exp(x_{im} \beta_m)\})$$

   where $R(y_i)$ is the risk set prior to the event y_i . x_{im} is the m^{th} factor amongst the p component, and δ_i denotes censoring indicator. Sort the covariates in ascending order based on this marginal utility. Accordingly, based on the characteristic or variable's marginal importance, the lowest U is the crucial covariate.

3. Sort the predictors in chronological order based on these marginal utility. Accordingly, the most influential predictors are the least L_j , which depends on the feature or variable's marginal utility.

4. Introduce the initial d characteristics. $d = \lfloor n/\log n \rfloor$, is a common formula, where $\lfloor . \rfloor$ denotes the floor function. Thus, $\overset{\%}{A}$ referred to as a subset of pre-approved factors.

5. Estimating the model parameters of the regression with penalties represents the last stage in the SIS process, as the subsequent illustrates.

   $$(\hat{\beta}_0, \hat{\beta}_m) = \underset{(\beta_0, \beta_m) \in R^{d+1}}{\arg\min} = \frac{1}{n} \sum_{i=1}^{n} L(i, \beta_0 + x_{i,M^*} \beta_{M^*}) + \sum_{j \in M^*} \lambda(|\beta_j|$$

   where the sub-vector yielded $x_i \in R_p$ through $d << p$ pre-approved variables M^* is denoted by the notation $x_{i,M^*} \in R_d$. The LASSO penalty is denoted by $\lambda(\|\beta_j\|)$ [18] explain the rationale behind the method's name (SIS). The mentioned algorithm's initial sorting step has a good chance of choosing all significant predictors when d is large enough. In the second step, choosing variables is achieved by the penalizing regression of LASSO, which also assesses the primary influences of the rest covariates.

ISIS

A significant flaw in the SIS approach is that factors will not be identified in the later round if they are ignored in the initial one. To put it another way, if a predictor is simultaneously uncorrelated yet has a larger peripheral association to the outcome over certain significant factors in the portion, or if a marker is marginally unconnected but jointly connected with the outcome [16, 39]. Iterative Sure Independence Screening (ISIS), which was introduced by [16], is an ongoing SIS technique designed to strengthen SIS and address the aforementioned issues. According to [18, 15, 3, 55-59], the processes of the ISIS technique have been summarized as below:

1. All of the statistically significant factors are recovered having a likelihood of one utilizing the SIS approach. Yet, the iterative sure independence screening (ISIS) strategy is employed when multiple important factors are only weakly uncorrelated by the answer [5, 3].
2. Regression parameter estimates β_{i1} are obtained by the iterative SIS using a penalty-based picking of features phase after an index list $\hat{I}_1$ is chosen using the Sure independence screening approach. The estimate $\hat{M}_1$ in $\hat{I}_1$ is changed depending on the positive components of $\hat{\beta}_{i1}$. The covariate m conditional usefulness, provided the fact that M does not contain the covariate, is outlined below:

$$U_{m|\grave{M}_1} = \max_{\beta_m \beta_{\grave{M}}} \left(\sum_{I=1}^{n} \begin{array}{c} \delta_i (x_i \beta_m + x^T{}_{\grave{M}_1,i} \beta_{\grave{M}_1} \\ - \sum\limits_{j \in R(y_i)}^{n} \delta_i \log(\sum\limits_{j \in R(y_i)} \exp(x_{jm}^T \beta_m + x^T{}_{\grave{M}_1,i} \beta_{\grave{M}_1})) \end{array} \right)$$

3. We employ the 2nd SIS step during this ISIS phase by using the subsequent formula to determine every factor's marginal usefulness.

4. To find the model's coefficients, we minimize the aforementioned equation using penalized regression. The following is the result of applying a penalized regression technique: an average model that is extremely similar to the original model

$$-\sum_{I=1}^{n} \delta_i (x^T_{\grave{M}_1 \cup \grave{I}_2,i} \beta_{\grave{M}_1 \cup \grave{I}_2})$$

$$+\sum_{i=1}^{n} \delta_i \log\{ \sum_{j \in R(y_i)} \exp(x^T_{\grave{M}_1 \cup \grave{I}_2,j} \beta_{\grave{M}_1 \cup \grave{I}_2,j)} \} + p\lambda(\beta_j)$$

Newer subset $\grave{M}_2$ of the chosen factors are produced by the values of $\beta_{\grave{M}_1 \cup \grave{I}_2}$ which are greater than zero.

5. Lastly, we carry out phases three and four unless we get to the set or d's stated set i.e.,

$$(\grave{M}_j = \grave{M}_{j-1})$$

Robust Lasso Freund Model with SIS (RLF-SIS)

According to [25], the robust Cox regression using Lasso regression is expressed as follows.

$$\sum_{t=1}^{n} \rho(y(t) - \sum_{i=1}^{p} \sum_{i=1}^{p} h_0(t) \exp(x_i^T \beta_j)) + p\lambda(|\beta_j|) \tag{4}$$

where $p_\lambda(|\beta_{j,l}|)$ is the penalty function, *k* is the tuning constant, and *ρ (.)* denotes the Huber loss function [27] as defined by [19]. The breakdown point (BP) and efficiency features are used as indicators to assess how effective penalized robust approaches are. While there exists a high ratio of contamination among the data, the BP serves as an index of an estimator's robustness. Since the least squares estimator's BP is as low as 1/n, an OLS estimator may prove to be worthless based on even one outlier observation.

Equation 4 is used to combine the Cox model, the SIS approach's operation, and the baseline hazard function for Freund's to create equation 5.

$$\sum_{t=1}^{n} \rho(y(t) - \sum_{i=1}^{p} \sum_{i=1}^{p} Fh_0(t) \exp(x_i^T \beta_{LASSO-SIS})) + p\lambda(|\beta|) \tag{5}$$

where $p_\lambda(|\beta_{j,l}|)$ is thepenalty function and Fh_0 baseline hazard function for Freund's.

Robust Lasso Freund Model with ISIS (RLF-ISIS)

Likewise, equation 6 is generated in the following manner whenever the Cox regression model, the Freund's baseline hazard function, and the ISIS approach are integrated.

$$\sum_{t=1}^{n}\rho(y(t)-\sum_{i=1}^{p}\sum_{i=1}^{p}Fh_0(t)\exp(x_i^T\beta_{LASSO-ISIS}))+p\lambda(|\beta|) \tag{6}$$

Efficiency features and the breakdown point (BP) are employed as metrics to assess the efficacy of robust penalized techniques. A measure of an estimator's robustness in cases where the contamination ratio of the data is high is called the BP [54].

LASSO-SIS with Cox-Freund Model (LF-SIS)

For the analysis of right-censored survivor data, the most popular paradigm is Cox's proportional hazards model [7]. It uses an independent distribution of the lifespan with a predetermined hazard function.

$$y(t)=h_0(t)\exp(x^T\beta) \tag{7}$$

In Equation (7), the patient's risk at each given time is represented by y(t), a fixed-length p vector, with the communal baseline hazard by ho(t). The PL function of Cox model is

$$L(\beta)=\prod_{i=1}^{N}\frac{\exp(x_{j(i)}^T\beta)}{\sum_{j\in R_i}\exp(x_j^T\beta)}$$

When PL is maximized to yield the standard Cox's estimator β, where Ri is the failure indicator at that moment. Let $p\lambda(\beta)$ be a penalty function that is non-differentiable at zero [2]. Take into account the discounted PL estimator.

$$\hat{\beta}_{LASSO-cox}=\arg\min_{\beta}\frac{1}{n}\left[\sum_{i=1}^{N}-x_{i,j}^T\beta+\log(\sum_{i\in R_i}\exp(x_i^T\beta_{LASSO-sis}))\right]+p\lambda(\beta) \tag{8}$$

$p\lambda(\beta)=\lambda\sum_{i=j}^{p}|\beta|$ represents the lasso penalty that Tibshirani (1997) used to fit the Cox regression model. Consequently, formula 8 takes on the following form when the Cox PH function and Freund baseline hazard function are integrated with the SIS approach activated.

$$y(t)=f(Fh_0)\arg\min_{\beta}\frac{1}{n}\left[\sum_{i=1}^{N}-x_{i,j}^T\beta+\log(\sum_{i\in R_i}\exp(x_i^T\beta_{LASSO-sis}))\right]+p\lambda(\beta) \tag{9}$$

LASSO-ISIS with Cox-Freund Model (LF-ISIS)

When the ISIS approach is activated along with the Freund baseline hazard function and Cox PH function, the formula (9) becomes:

$$y(t)=f(Fh_0)\arg\min_{\beta}\frac{1}{n}\left[\sum_{i=1}^{N}-x_{i,j}^T\beta+\log(\sum_{i\in R_i}\exp(x_i^T\beta_{LASSO-isis}))\right]+p\lambda(\beta)$$

Where, $p\lambda(\beta)=\lambda\sum_{i=j}^{p}|\beta|$ to fit a lasso penalized Cox regression model

The Elastic Net-SIS with Cox-Freund Model (ENF-SIS)

Here is how the EN estimator is defined,

$$\hat{\beta}^{(EN)} = \beta \arg\min_{\beta} \left\| y - \sum_{j=1}^{p} x_{j} \beta_{j} \right\|^{2} + \lambda_{2} \left\|\beta\right\|^{2} + \lambda_{1} \left\|\beta\right\|_{1}$$

Where $\|\beta\|_1$ and $\|\beta\|^2$ are defined as

$$\|\beta\|_{1} = \sum_{p=0}^{p} \left|\beta_{p}\right|$$

$$\|\beta\|^{2} = \sum_{p=0}^{p} \left|\beta_{p}{}^{2}\right|$$

Again, the cox model's partial likelihood function is [1,2]

$$L(\beta) = \prod_{i=1}^{N} \frac{\exp(x_{j(i)}^{T}\beta)}{\sum_{j\in R_i} \exp(x_{j}^{T}\beta)}$$

$$subject\,to\,restriction : \alpha \left|\beta + (1-\alpha)\sum \beta_{i}^{2}\right| \le c.$$

A scaled log partial likelihood is maximized when the PL is maximized.

$$\frac{2}{n} l(\beta) = \frac{2}{n}\left[\sum_{i=1}^{N} x_{j(i)}^{T}\beta - \log(\sum_{j\in R_i} \exp(x_{j}^{T}\beta)\right]$$

For simplicity, by scale with a factor of 2/n. Now, using Lagrange's equation, the issue is as follows:

$$\hat{\beta}^{(EN-cox)} = \arg\min_{\beta}\left[\frac{2}{n}(\sum_{i=1}^{N} x_{j(i)}^{T}\beta - \log(\sum_{j\in R_i}\exp(x_{j}^{T}\beta_{EN-SIS})) - \lambda p_{\alpha}(\beta)\right] \tag{10}$$

where

$$\lambda p_{\alpha}(\beta) = \lambda(\alpha\sum_{i=1}^{p}\left|\beta_{i}\right| + \frac{1}{2}(1-\alpha)\sum_{i=1}^{p}\beta_{i}^{2}$$

Thus, the following is what would happen to equation (10) if Freund's model with Cox-EN was taken into account:

$$y(t) = f(Fh_{0}) \arg\min_{\beta}\left[\frac{2}{n}(\sum_{i=1}^{N} x_{j(i)}^{T}\beta - \log(\sum_{j\in R_i}\exp(x_{j}^{T}\beta_{EN-SIS})) - \lambda p_{\alpha}(\beta)\right]$$

The Elastic Net-ISIS with Cox-Freund Model (ENF-ISIS)

The formula in the instance of the ISIS methodology

$$y(t) = f(Fh_{0}) \arg\min_{\beta}\left[\frac{2}{n}\sum_{i=1}^{N} x_{j(i)}^{T}\beta - \log(\sum_{j\in R_i}\exp(x_{j}^{T}\beta_{EN-ISIS})) - \lambda p_{\alpha}(\beta)\right]$$

Performance Evaluation Criteria

The anticipating model's accuracy has been evaluated using the coefficient of determination (R^2), mean squared error (MSE), sum of squares error (SSE), and root mean squared error (RMSE). Regression tasks, which often use evaluation measures like MSE, SSE, RMSE, and R-squared, are therefore the focus of our research [65,47].

Mean Squared Error (MSE)

The statistical instrument employed in forecasting to assess the regression line's correctness called the mean squared error, or MSE. It measures the distances from points to the regression line, excluding negative signals, and gives more weight to larger differences. The calculative formula of MSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where, Y is the i^{th} true value, $\hat{Y}$ denotes the i^{th} estimated value, and n represents the overall count of dataset. An MSE score which is nearer to zero indicates that the model is more in line with the data [47, 10].

Sum of Squares Error (SSE)

The Sum of Squares for Error (SSE) measures variation residuals in regression models, with a lower SSE indicating better data description and more stable predictive capacity [10].

$$SSE = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

In this case, X_i denotes i^{th} observation's value, and n is the overall count of dataset.

Root Mean Squared Error (RMSE)

An indicator called Root Mean Squared Error (RMSE) is utilized in both machine learning and statistics to determine how accurate an algorithm for prediction is. It examines the disparity between anticipated and true values; greater forecasting accuracy is demonstrated by lower values. Considerably significant divergence between the residual and the actual truth is apparent by a higher RMSE. Squaring the MSE results yields the computed value of RMSE. It deals with the aberrations from the real value and evaluates the overall extent of the errors. A zero value of the RMSE implies the model fits the data perfectly. The calculative formula can be written as follows [65]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} = \sqrt{MSE}$$

Coefficient of Determination (R^2)

A goodness-of-fit metric for models based on the percentage of explained variation is the coefficient of determination. R^2 is commonly understood to indicate the percentage of the dependent variable's variation that can be accounted for by variations in the determinants.

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\overline{y})^2}$$

R^2 has two possible values: 0 is the least and 1 is the largest. To put it in simple terms, a model's R^2 will approach 1, the more successful it is at predicting outcomes [47, 14].

Software

All the analysis was performed in R language version 4.3.2.

Results and Discussion

Performance Evaluation

Since LASSO and EN are the two fundamental methods for feature selection, the study initially used them for RCC data and simultaneously assessed the effectiveness of each method. Table 2 shows the performance of LASSO and EN. When compared with LASSO, EN's MSE, SSE, and RMSE values are significantly smaller (LASSO: MSE=726.33, SSE=53748.6, RMSE=26.950; EN: MSE=587.67, SSE=43487.5, RMSE = 24.241). This is evident from the fact that when the data dimension increases, LASSO performs poorly [49]. Furthermore, as genes are typically interrelated, the collection includes 4224 differentially expressed genes. As a result, the dataset may exhibit multicollinearity. The previous study [25] showed that, the efficacy of LASSO is adversely affected by highly linked relationships among the real set of factors along with unimportant ones. EN is the solution of such situation [49]. To increase accuracy in forecasting, the EN applies penalized periods L1-Lasso and L2-right, recognizes the variables, and carries out continual shrinking. Consequently, EN outperformed LASSO in this instance. Interestingly, the EN and LASSO coefficients of determination are almost identical, at 0.53 and 0.54, respectively. In other words, the data set accounts for around 54% of the variation in the dependent variable. For a further 44% of the data, there are still unanswered questions [31, 14]. However, this coefficient of determination values suggests that none of these methods are appropriate for this particular dataset.

Table 2. Performance of LASSO and EN as feature selection technique

No.	Methods	NGS	MSE	SSE	RMSE	R^2	PVS
1	**LASSO**	23	726.33	53748.6	26.950	0.54	0.54%
2	**EN**	92	587.67	43487.5	24.241	0.53	2.17%

The study took into consideration SIS in LASSO Freund, Robust LASSO Freund, Elastic Net, and Elastic Net Freund since it believes that the dataset contains multicollinearity. Table 3 additionally evaluates and displays the effectiveness of the strategies that have been suggested and taken into consideration. Combining LASSO with Freund and SIS (LF-SIS) raised the number of chosen genes from 23 to 41. In addition, LF-SIS has better values for all other indexes than LASSO, including MSE, SSE, RMSE, and R^2 (LASSO: MSE=726.33, SSE=53748.6, RMSE=26.950, and R^2 = 0.54; LF-SIS: MSE= 712.630, SSE= 52734.6, RMSE= 26.695, and R^2 = 0.673). Even yet, the coefficient of determination value, R^2 = 0.673., showed a significant improvement. However, 33% of the dataset's volatility cannot be explained by the predictors [22,65]. Consequently, for this dataset, LF-SIS might not be the best option when it comes to variable selection.

Table 3. Performance of LASSO, EN and Proposed methods (LF-SIS, RLF-SIS, EN-SIS and ENF-SIS) as a feature selection technique

No.	Methods	NGS	MSE	SSE	RMSE	R^2	PVS
1	**LASSO**	23	726.33	53748.6	26.950	0.54	0.54%
2	**EN**	92	587.67	43487.5	24.241	0.53	2.17%
3	**LF-SIS**	41	712.630	52734.6	26.695	0.673	1%
4	**RLF-SIS**	**68**	**342.009**	**25308.7**	**18.493**	**0.710**	1.6%
5	**EN-SIS**	40	600.15	44410.88	24.497	0.67	0.946%
6	**ENF-SIS**	85	429.98	318 18.69	20.736	0.68	2.012%

In addition, when integrating SIS with EN (EN-SIS) and with EN and Freund (ENF-SIS), ENF-SIS outperformed EN and EN-SIS (EN: MSE=587.67, SSE=43487.5, RMSE = 24.241, R$^{2\,=}$ 0.53; EN-SIS: MSE=600.15, SSE=44410.88, RMSE = 24.497, R^2 = 0.67; ENF-SIS: MSE=429.98, SSE=318 18.69, RMSE = 20.736, R^2 = 0.68). Since the foundation of the SIS method is correlation learning, which makes use of the observed relationship among the predictor and responder. Consequently, the accuracy of the suggested approaches (LF-SIS, EN-SIS, and ENF-SIS) improved when SIS was paired with LASSO and EN as opposed to using LASSO and EN alone. Again, since kidney is a parallel system, therefore it is necessary to keep in consideration that in the case of the parallel system, redundancy increases survival; hence, the failure of the system depends on the failure of each of its k components. Stated differently, the system continues to work even if one of its components fails. Furthermore, the Freund model is ideal for accurately illustrating such a scenario [21]. Because of this, the Freund model performed better when combined with SIS and EN (ENF-SIS outperformed). Additionally, the R^2 coefficient of determination values Clearly improved after adding SIS and EN to Freund. Nevertheless, based on the values (R^2), none of these methods is optimal for this dataset.

The picture was drastically altered when Freund Robust LASSO was merged with SIS, named RLF-SIS. Comparing RLF-SIS to other techniques under consideration (LASSO, EN, LF-SIS, EN-SIS, ENF-SIS, Table 3), its MSE, SSE, and RMSE values are significantly lower (RLF-SIS: MSE=342.009, SSE=25308.7, RMSE = 18.493). Vitally, irrespective all of the approaches taken into consideration, the correlation coefficient of measurement is R^2 = 0.710, resulting in is the highest. The improvement in every RLF-SIS performance index raises the possibility of an outlier in the data. Which is quite normal as noted by author [17]. According to the author, one to ten percent of a real dataset may contain outliers.

A significant flaw in the SIS approach is that factors will not be identified in the later round if they are ignored in the initial one. To put it another way, if a predictor is simultaneously uncorrelated yet has a larger peripheral association to the outcome over certain significant factors in the portion, or if a marker is marginally unconnected but jointly connected with the outcome [16, 39]. Iterative Sure Independence'Screening (ISIS), which was introduced by [16], is an ongoing SIS technique designed to strengthen SIS and address the aforementioned issues. Thus, at this point in the study, ISSIS was combined with EN- (EN-ISIS), LASSO (LF-ISIS) and Freund (ENF-ISIS, RLF-ISIS). Table 4 evaluates and summarizes the efficacy of the suggested and fundamental approaches. After merging ISIS, every recommended approach—LF-ISIS, RLF-ISIS, EN-ISIS, and ENF-ISIS—performed better than the basic LASSO and EN, with the exception of EN-ISIS. Though the value of MSE, SSE and RMSE for all approaches are quite low but Out of the all-suggested approaches, the RLF-ISIS technique yielded the least amount of generated error (RLF-ISIS: MSE = 342.009, SSE: 25308.7, RMSE =18.493). Moreover, according to the R^2 threshold, each method offers a credible approximation. Nevertheless, among the all options, the RLF-ISIS technique has the greatest R^2 score (R^2 = 0.71), meaning that 71% of the variation of the dependent variable has been accounted for in the data set. Given another 29 percent of the data remain unresolved for, the predicted outcome can be considered appropriate because it accurately covers the data [31, 14]. Once more, given that RLF-ISIS is superior to all other suggested approaches, it is advised that an outlier could exist in the dataset. Though, the score of R^2 for RLF-ISIS is highest among all considered approaches yet the value cannot be considered as an excellent one [22]. Accordingly, RLF-ISIS might not be the best option, but in ultra-high dimensional survival data, it can be thought of as a variable selection technique.

Table 4. Performance of LASSO, EN and Proposed methods (LF-ISIS, RLF-ISIS, EN-ISIS and ENF- ISIS) as a feature selection technique

No.	Methods	NGS	MSE	SSE	RMSE	R^2	PVS
1	**LASSO**	23	726.33	53748.6	26.950	0.54	0.54%
2	**EN**	92	587.67	43487.5	24.241	0.53	2.17%
3	**LF-ISIS**	38	429.982	31818.7	20.736	0.665	0.9%
4	**RLF-ISIS**	49	342.009	25308.7	18.493	0.710	1.2%
5	**EN-ISIS**	35	652.81	52205.01	26.560	0.67	0.828%
6	**ENF-ISIS**	69	429.98	31818.69	20.736	0.68	1.633%

Table 5 summarizes the study's total results, which were gathered from all suggested and considered methods. Out of the all approaches, the RLF-ISIS and RLF-SIS technique yielded the least amount of error (MSE = 342.009, SSE: 25308.7, RMSE =18.493). From the RCC data, the ENF-ISIS method found 69 worthy variables. As a result, the percentage of factors chosen for the RLF-ISIS and RLF-SIS procedures is 1.609% as well as 1.160%, accordingly, whereas the RLF-ISIS methodology selected 49 key predictors. This is a result of the reasoning for expanding the SIS [18]. Furthermore, there are commonalities in the assessment parameters (MSE = 429.982, SSE = 31818.69, RMSE = 20.736, and R^2 = 0.680) between the ENF-SIS, ENF-ISIS, and LF-ISIS approaches. The quantity of genes that are significant and discovered from the RCC data varies throughout the three methods, nonetheless.

According to the R^2 threshold, each method offers a credible approximation. Nevertheless, among the all options, the RLF-ISIS and RLF-SIS technique has the greatest R^2 score (R^2 = 0.71), meaning that 71% of the variation of the dependent variable has been accounted for in the data set. They were succeeded by the ENF-ISIS and ENF-SIS method, with a value of R^2 (R^2 = 0.68), after which came the EN-SIS and EN-ISIS approach, that offers a value of R^2 = 0.67. The least R^2 coefficients have been produced by the EN and LASSO, with 0.53 and 0.54 respectively. Even after producing least error by the method RLF-ISIS and RLF-SIS, the R^2 score still not higher enough [22, 47] to select these two methods as the best.

Table 5. Performance of LASSO, EN and all Proposed methods (LF-SIS, LF-ISIS, RLF-SIS, RLF-ISIS, EN-SIS, EN-ISIS, ENF-SIS and ENF-ISIS)

No.	Methods	NGS	MSE	SSE	RMSE	R^2	PVS
1	**LASSO**	23	726.33	53748.6	26.950	0.54	0.54%
2	**EN**	92	587.67	43487.5	24.241	0.53	2.17%
3	**LF-SIS**	41	712.630	52734.6	26.695	0.673	1%
4	**RLF-SIS**	**68**	**342.009**	**25308.7**	**18.493**	**0.710**	1.6%
5	**EN-SIS**	40	600.15	44410.88	24.497	0.67	0.946%
6	**ENF-SIS**	85	429.98	318 18.69	20.736	0.68	2.012%
7	**LF-ISIS**	38	429.982	31818.7	20.736	0.665	0.9%
8	**RLF-ISIS**	**49**	**342.009**	**25308.7**	**18.493**	**0.710**	**1.2%**
9	**EN-ISIS**	35	652.81	52205.01	26.560	0.67	0.828%
10	**ENF-ISIS**	69	429.98	31818.69	20.736	0.68	1.633%

Despite the RLF-ISIS and RLF-SIS approaches providing the least amount of error, their respective R^2 scores are still insufficient to declare them the best. The possible reason could be the regression lines, as indicated by their relatively low R^2 values, don't fit the data very well because there are a lot of outliers. Although the RLF-ISIS and RLF-SIS strategy perform better than the other suggested approaches, as indicated by the validating model measures (MSE, SSE and RMSE), their lower R^2 value suggests that these two methods may not be the best choice. However, in ultra-high dimensional survival data, they can be considered as variable selection technique. Thus, additional study is required to create a more sophisticated variable selection method in this context.

Selected Genes by the Best Methods

The two optimal method RLF-SIS and RLF-ISIS has led to the discovery of 68 and 49 genes that affect RCC patients directly as well as indirectly. The name of these selected gene is presented in Table 6.

Table 6. Selected Genes by the best Method

Methods	NSG	Selected Genes
RLF-SIS	68	X28, X96, X102, X134, X151, X161, X172, X210, X347, X388, X419, X421, X422, X429, X438, X482, X527, X610, X641, X672, X768, X779, X834, X854, X977, X995, X1029, X1082, X1115, X1188, X1241, X1302, X1428, X1525, X1571, X1601, X1656, X1656, X1709, X1758, X1830, X1842, X1882, X1948, X2048, X2174, X2233, X2295, X2467, X2511, X2734, X2784, X2850, X2915, X2975, X3197, X3218, X3353, X3369, X3418, X3451, X3685, X3760, X3803, X3914, X3918, X3975, X4039, X4212.
RLF-ISIS	49	X28, X134, X161, X172, X315, X419, X421 X42, X466, X482, X527, X610, X689, X778, X854, X892, X977, X995, X1047, X1217, X1241, X1325, X1428, X152, X1601, X1656, X1728, X1830 X2048, X2074, X2295, X2467, X2511, X2734, X2784, X2850, X2975, X3218, X3353, X3451, X3666, X3685, X3803, X3918, X3947, X3975, X4039, X4154, X4212.

Each of the genes chosen using the RLF-ISIS and RLF-SIS method influences the progression of RCC. For instance, the autosomal dominating sickness caused by the VHL syndromes in the planned RLF-ISIS and RLF-SIS has a propensity for numerous neoplasms. The illnesses encompass intrinsic pancreas cancers, VHL, lymph node hemangioblastomas, renal cell carcinomas, pheochromocytomas, cysts, along with cystadenomas [37, 53, 72].

One important function of the AKR7A2(X689) gene is to shield organs such as the liver and kidneys from the harmful and cancer-causing effects of AFB1, a potent hepatocarcinogen [72]. The X977 gene influences movement and aggressiveness of endothelial cells alongside is linked to carcinogenesis. It might additionally have a role in angiogenesis. Furthermore, it contributes to viral infection, namely with the human cytomegalovirus [51]. The X4154 gene plays a major role in RCC and is implicated among both tumors that are benign and malignant [15, 37]. The X1241 gene initiates signaling which is in charge of cellular degeneration [60]. The X1525 gene is implicated under transcriptional control and is in charge of transcription regulating the expression of genes [44]. The gene X4212 is responsible for the transportation and regeneration of sphingolipids as well as cholesterol levels through the plasma membrane.

The majority of investigators in the area of healthcare disorders have persistently disregarded both two-component structures and have only used fundamental approaches of choosing variables. By suggesting a mixed approach rather than relying solely on one-component regression algorithms to identify important genes across two-component infrastructure, this study closed the gap. The accuracy of the study's findings and their superiority over those of previous investigations are shown by comparing them with the findings of other scholars. For example, [41] used a single component and no machine learning techniques to identify 32 genes associated with the development of kidney cancer thus showed the value of microarrays for carcinoma categorization that takes into account variations in the makeup of tissues amongst RCC varieties. Furthermore, [12] discovered 6 hub genes (SUCLG1, PCK2, GLDC, SLC12A1, ATP1A1, PDHA1) and saw a markedly shorter life expectancy period for RCC sufferers. Berglund *et al*. (2020) identified 9 genes (AURKA, AURKB, BIRC5, CCNE1, MKI67, MMP9, PLOD2, SAA1, and TOP2A) that exhibit distinct expression patterns. While the remaining genes were previously associated to various cancers, 6 of the validated genes—BIRC5, MK167, MMP9, PLOD2, and TOP2A—have historically been suspected of RCC.

According to [24], the LASSO model was constructed and the support vector machine (SVM) technique was utilized to obtain 8 genes which are associated with the progression of early RCC. This study beat previous research in discovering important genes linked to RCC, where the RLF-ISIS strategy found 49 RCC-relevant genes. The study employed a mixture of hybrid models and techniques for machine learning. The findings show that the suggested techniques can improve the capacity of choosing factors utilizing machine learning to successfully detect a certain number of significant predictors (genes) from ultra-high dimensional survival data.

Box-Plot of Predictor Variables

The study created some box-plots of predictors, which are shown in Figures 2, 3, and 4, to examine outliers in the dataset. The gene expressions of the genes X1-X160 are shown in Figure 2, the gene expressions of the predictor genes X161-260 are displayed in Figure 3, and the gene expression values of the genes X4000-X4224 are presented in Figure 4. Three figures are demonstrating the existence of outliers in the dataset—which the prior analysis had suggested (lower R^2 value). The ultra-high dimensional RCC dataset contains outliers, as this study confirms. One to ten percent of an actual dataset may contain outliers, according to author [17]. For instance, there are currently 28 specimens in the RNA-seq collection of triple negative breast cancer that have ambiguous labeling, making them outliers. [59]. Heterogeneity issues, such as samples drawn from several segments of the population, or mechanical problems with microarray research led to outliers [58]. For ultra-high-dimensional omics data, multiple feature selection techniques are available [60-64]. But there are no feature selection techniques available that take the issue of outliers as well as containing multiple compartments. Further study is needed to build an advanced hybrid feature selection technique in this particular setting.

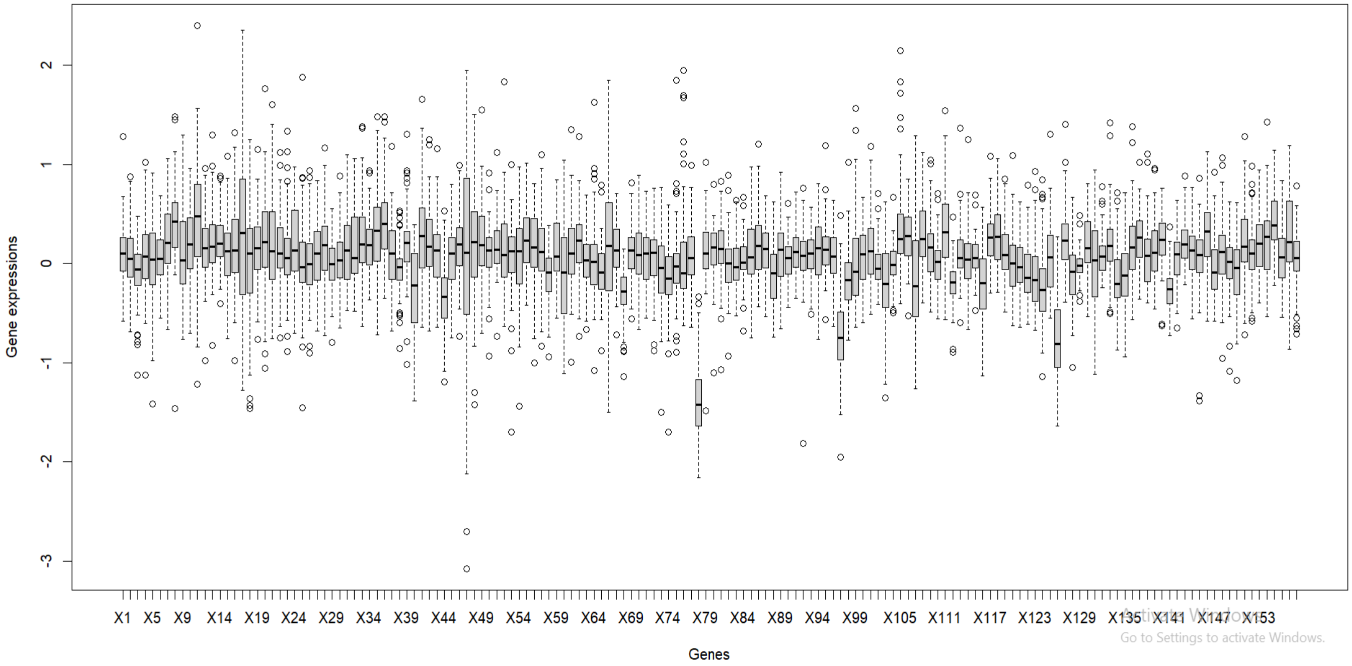

Figure 2. The gene expressions of the genes X1 to X160

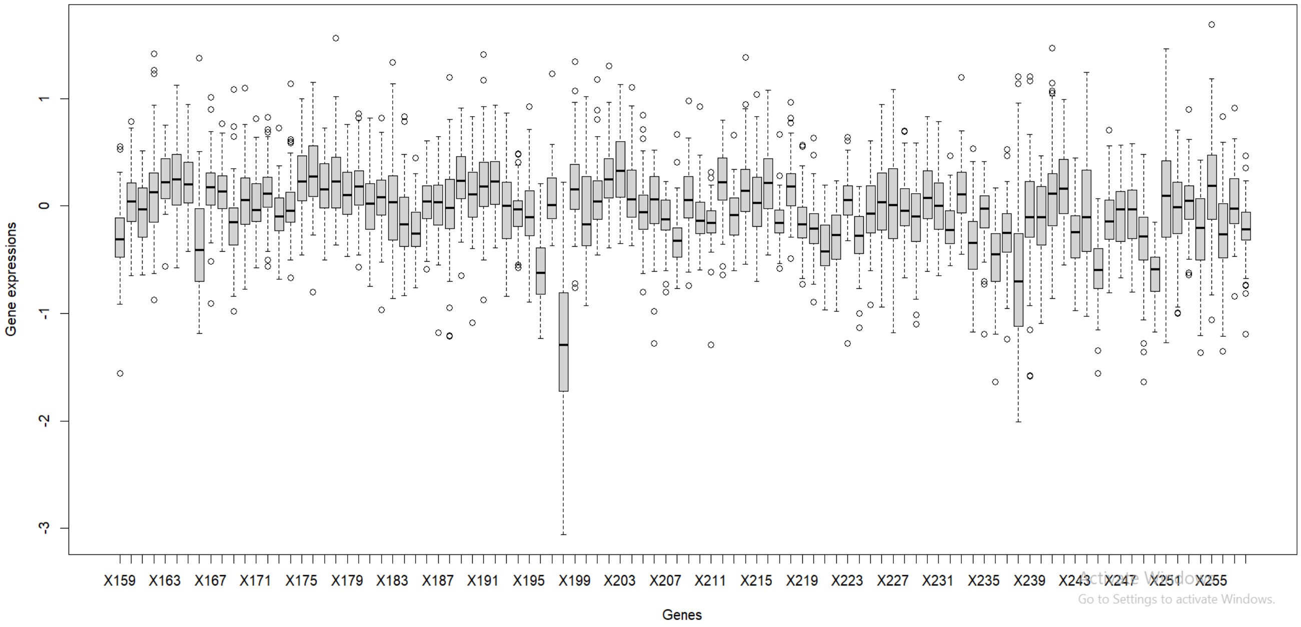

Figure 3. The expressions of the predictor genes X161 to X260

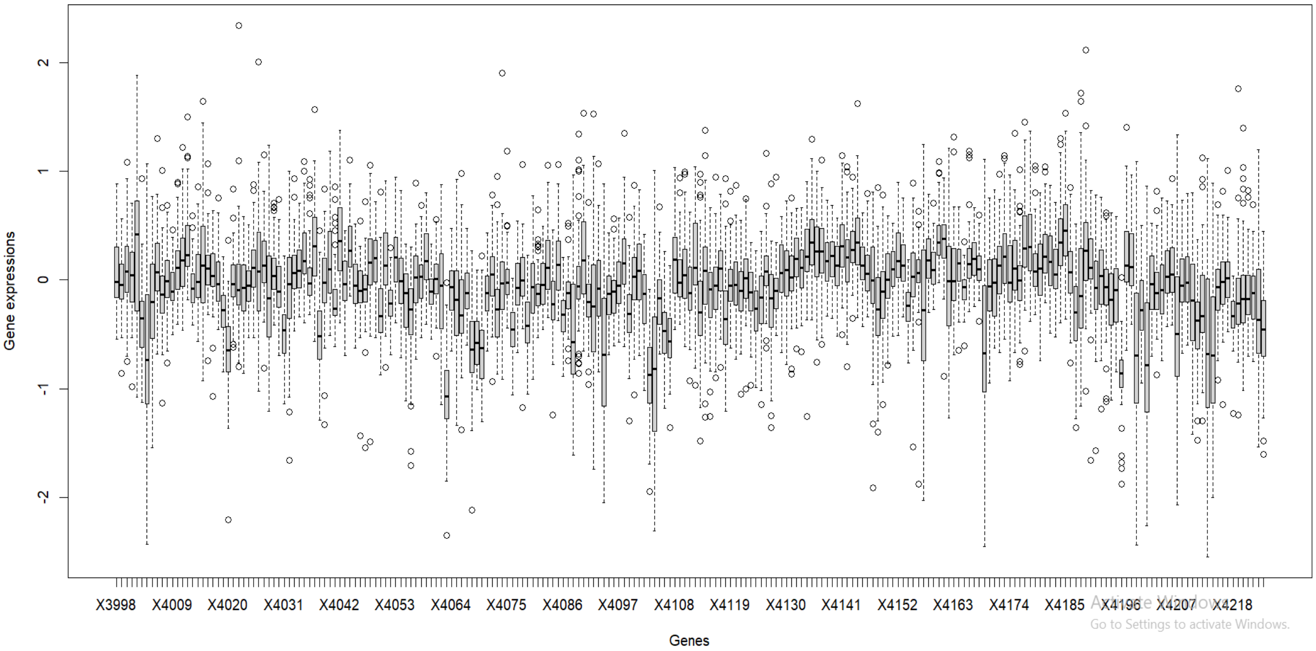

Figure 4. The gene expression values of the genes X4000 to X4224

Conclusion

Reliability of variable selection for actual clinical disease data is still one of the most difficult areas of biosystems study. A few genes are significant and connected to the illness. Conversely, certain genetic factors are unconnected or just slightly affect the illness under investigation. It is crucial to determine the affecting gene and how it relates to that illness. The study proposed eight approaches (LF-SIS, RLF-SIS, EN-SIS, ENF-SIS, LF-ISIS, RLF-ISIS, EN-ISIS, ENF-SIS) and also considered 2 fundamental methods (EN and LASSO) as a variable selection procedure for ultra-high dimensional survival data. The performance of these all approaches is evaluated based on MSE, SSE, RMSE and R^2 value. Although the RLF-ISIS and RLF-SIS methods yield the lowest error, their corresponding R^2 values are not high enough to qualify them as the best. The explanation for this could be that there are too many outliers in the data, which makes the regression lines fit the data poorly. The box-plot of some selected predictive genes confirms the presence of outliers in the dataset. However, this study examines the recommended approaches' capacity to find features for ultra-high dimensional data sets using variable selection approaches. The research can be widened later on by implementing competitive risk theory to a sequential and parallel structure, which makes up the foundation for the majority of intricate mechanical systems seen in manufacturing facilities. Finally, it can be concluded that although RLF-ISIS and RLF-SIS outperform other proposed approaches, they may, however, be regarded as a variable selection strategy but they might not be the optimal choice for ultra-high dimensional survival data with outliers. Outliers and multicollinearity of these genes create distorted or misleading results due to the behavior of genes among themselves. To date, there is still a gap in feature selection strategy for ultra-high-dimensional survival data with outliers and multi-compartment system. To handle this particular scenario, further research is needed to develop an advanced hybrid feature selection approach, focusing on deep learning strategies. For example, deep neural network (deep learning) can be used which is a part of artificial intelligence and is able to identify mistakes or deficits in the outcome and fix them without human interaction.

Conflict of Interest

The authors said they had no conflicting agendas.

Acknowledgement

I am appreciative that University Sains Malaysia gave me the opportunity to attend this esteemed university. Also, I am most thankful of the amazing administrative team at the School of Mathematics Sciences for their assistance in getting me beyond all of the challenges I had while conducting this research.

References

[1] AL-Rammahi, A. H., & Dikheel, T. R. (2021, October). Sure independent screening elastic net for ultra-high dimensional survival data. In *AIP Conference Proceedings* (Vol. 2404, No. 1). AIP Publishing.

[2] AL-Rammahi, A. H., & Dikheel, T. R. (2022, October). Freund's model with iterated sure independence screening in Cox proportional hazard model. In *AIP Conference Proceedings* (Vol. 2398, No. 1). AIP Publishing.

[3] Araveeporn, A. (2022). The penalized regression and penalized logistic regression of Lasso and elastic net methods for high-dimensional data: A modelling approach.

[4] Ba, Z., Xiao, Y., He, M., Liu, D., Wang, H., Liang, H., & Yuan, J. (2022). Risk factors for the comorbidity of hypertension and renal cell carcinoma in the cardio-oncologic era and treatment for tumor-induced hypertension. *Frontiers in Cardiovascular Medicine*, 9, 810262.

[5] Bhattacharjee, A., Dey, J., & Kumari, P. (2022). A combined iterative sure independence screening and Cox proportional hazard model for extracting and analyzing prognostic biomarkers of adenocarcinoma lung cancer. *Healthcare Analytics*, 2, 100108.

[6] Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n.

[7] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187—202.

[8] Chamlal, H., Benzmane, A., & Ouaderhman, T. (2024). Elastic net-based high dimensional data selection for regression. *Expert Systems with Applications*, 244, 122958.

[9] Cheon, S., Agarwal, A., Popovic, M., Milakovic, M., Lam, M., Fu, W., *et al.* (2016). The accuracy of clinicians' predictions of survival in advanced cancer: A review. *Annals of Palliative Medicine*, 5(1), 22—29.

[10] Zhang, L., Zhang, J., Gao, W., Bai, F., Li, N., & Ghadimi, N. (2024). A deep learning outline aimed at prompt skin cancer detection utilizing gated recurrent unit networks and improved orca predation algorithm. *Biomedical Signal Processing and Control*, 90, 105858.

[11] Cheng, X., & Wang, H. (2023). A generic model-free feature screening procedure for ultra-high dimensional data with categorical response. *Computer Methods and Programs in Biomedicine*, 229, 107269.

[12] Chen, Y., Gu, D., Wen, Y., Yang, S., Duan, X., Lai, Y., *et al.* (2020). Identifying the novel key genes in renal cell carcinoma by bioinformatics analysis and cell experiments. *Cancer Cell International*, 20, 1—16.

[13] Huang, J. W., Chen, Y. H., Phoa, F. K. H., Lin, Y. H., & Lin, S. P. (2024). An efficient approach for identifying important biomarkers for biomedical diagnosis. *Biosystems*, 105163.

[14] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, e623.

[15] Domingo-Relloso, A., Feng, Y., Rodriguez-Hernandez, Z., Haack, K., Cole, S. A., Navas-Acien, A., *et al.* (2024). Omics feature selection with the extended SIS R package: Identification of a body mass index epigenetic multi-marker in the Strong Heart Study. *American Journal of Epidemiology*, kwae006.

[16] Fan, J., Samworth, R., & Wu, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *The Journal of Machine Learning Research*, 10, 2013—2038.

[17] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348—1360.

[18] Fan, J., & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 70(5), 849—911.

[19] Feng, Y., & Wu, Q. (2022). A statistical learning assessment of Huber regression. *Journal of Approximation Theory*, 273, 105660.

[20] Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: A systematic review. *Diabetology & Metabolic Syndrome*, 13(1), 1—22.

[21] Freund, J. E. (1961). A bivariate extension of the exponential distribution. *Journal of the American Statistical Association*, 56(296), 971—977.

[22] Gujarati, D. N., Bernier, B., & Bernier, B. (2004). *Econométrie* (pp. 17—5). Brussels: De Boeck.

[23] Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A critical review of LASSO and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, 90(1), 118—145.

[24] Han, X., & Song, D. (2022). Using a machine learning approach to identify key biomarkers for renal clear cell carcinoma. *International Journal of General Medicine*, 3541—3558.

[25] Kong, S., Yu, Z., Zhang, X., & Cheng, G. (2021). High-dimensional robust inference for Cox regression models using desparsified Lasso. *Scandinavian Journal of Statistics*, 48(3), 1068—1095.

[26] Zhou, H., & Zou, H. (2024). The nonparametric Box–Cox model for high-dimensional regression analysis. *Journal of Econometrics*, 239(2), 105419.

[27] Huber, G. P. (1981). The nature of organizational decision making and the design of decision support systems.

MIS Quarterly, 1—10.

[28] Jaffe, S. (2015). Planning for US precision medicine initiative underway. *The Lancet*, 385(9986), 2448—2449.

[29] Sathasivam, S., Adebayo, S. A., Velavan, M., Yee, T. H., & Yi, T. P. (2024, January). Transmission of hepatitis B dynamics in Malaysia using modified SIS hybrid model with Euler and Runge-Kutta method. In *AIP Conference Proceedings* (Vol. 3016, No. 1). AIP Publishing.

[30] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, 8—17.

[31] Legates, D. R., & McCabe Jr, G. J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233—241.

[32] Liu, Z., Elashoff, D., & Piantadosi, S. (2019). Sparse support vector machines with l0 approximation for ultra-high dimensional omics data. *Artificial Intelligence in Medicine*, 96, 134—141.

[33] Lu, B., Wang, F., Wang, S., Chen, J., Wen, G., & Fu, R. (2024). Improvement of motor imagery electroencephalogram decoding by iterative weighted Sparse-Group Lasso. *Expert Systems with Applications*, 238, 122286.

[34] Mayer, D. G., & Butler, D. G. (1993). Statistical validation. *Ecological Modelling*, 68(1-2), 21—32.

[35] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care*, 24(1), 31—42.

[36] Mihaylov, I., Nisheva, M., & Vassilev, D. (2019). Application of machine learning models for survival prognosis in breast cancer studies. *Information*, 10(3), 93.

[37] Sartori, S. (2011). Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets.

[38] Salerno, S., & Li, Y. (2023). High-dimensional survival analysis: Methods and applications. *Annual Review of Statistics and Its Application*, 10, 25—49.

[39] Saldana, D. F., & Feng, Y. (2018). SIS: An R package for sure independence screening in ultrahigh-dimensional statistical

[40] Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & *Brodaty, H.* (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1), 1–10.

[41] Sültmann, H., Heydebreck, A. V., Huber, W., Kuner, R., Buneβ, A., Vogt, M., & *Poustka, A.* (2005). Gene expression in kidney cancer is associated with cytogenetic abnormalities, metastasis formation, and patient survival. *Clinical Cancer Research*, 11(2), 646–655.

[42] Shuch, B., Amin, A., Armstrong, A. J., Eble, J. N., Ficarra, V., Lopez-Beltran, A., & *Kutikov, A.* (2015). Understanding pathologic variants of renal cell carcinoma: Distilling therapeutic opportunities from biologic complexity. *European Urology*, 67(1), 85–97.

[43] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 58(1), 267–288.

[44] Vasilevsky, N. A., Matentzoglu, N. A., Toro, S., Flack IV, J. E., Hegde, H., Unni, D. R., & *Haendel, M. A.* (2022). Mondo: Unifying diseases for the world, by the world. *medRxiv*, 2022-04.

[45] Wang, H., & Li, G. (2017). A selective review on random survival forests for high dimensional data. *Quantitative Bio-science*, 36(2), 85.

[46] Xu, X., Liang, T., Zhu, J., Zheng, D., & Sun, T. (2019). Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*, 328, 5–15.

[47] Yarahmadi, M. N., MirHassani, S. A., & Hooshmand, F. (2024). Handling the significance of regression coefficients via optimization. *Expert Systems with Applications*, 238, 121910.

[48] Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

[49] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 67(2), 301–320.

[50] 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.

[51] Ryan, R. J., Nitta, M., Borger, D., Zukerberg, L. R., Ferry, J. A., Harris, N. L., & *Le, L. P.* (2011). EZH2 codon 641 mutations are common in BCL2-rearranged germinal center B cell lymphomas. *PLOS ONE*, 6(12), e28585.

[52] Wang, Z., Song, Q., Yang, Z., Chen, J., Shang, J., & Ju, W. (2019). Construction of immune-related risk signature for renal papillary cell carcinoma. *Cancer Medicine*, 8(1), 289–304.

[53] Walton, J., Lawson, K., Prinos, P., Finelli, A., Arrowsmith, C., & Ailles, L. (2023). PBRM1, SETD2 and BAP1—the trinity of 3p in clear cell renal cell carcinoma. *Nature Reviews Urology*, 20(2), 96–115.

[54] Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics—Simulation and Computation*, 46(8), 6261–6282.

[55] Sathasivam, S., Adebayo, S. A., Velavan, M., Yee, T. H., & Yi, T. P. (2024, January). Transmission of hepatitis B dynamics in Malaysia using modified SIS hybrid model with Euler and Runge-Kutta method. In *AIP Conference Proceedings* (Vol. 3016, No. 1). AIP Publishing.

[56] Xiong, T., Wang, Y., & Zhu, C. (2024). A risk model based on 10 ferroptosis regulators and markers established by LASSO-regularized linear Cox regression has a good prognostic value for ovarian cancer patients. *Diagnostic Pathology*, 19(1), 4.

[57] Ghosh, A., Jaenada, M., & Pardo, L. (2024). Robust adaptive variable selection in ultra-high dimensional linear regression models. *Journal of Statistical Computation and Simulation*, 94(3), 571–603.

[58] Madadjim, R., An, T., & Cui, J. (2024). MicroRNAs in pancreatic cancer: Advances in biomarker discovery and therapeutic implications. *International Journal of Molecular Sciences*, 25(7), 3914.

[59] Lopes, M. B., Veríssimo, A., Carrasquinha, E., Casimiro, S., Beerenwinkel, N., & Vinga, S. (2018). Ensemble

outlier detection and gene selection in triple-negative breast cancer data. *BMC Bioinformatics*, 19, 1–15.

[60] Yin, Q., Chen, W., Zhang, C., & Wei, Z. (2022). A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection. *Laboratory Investigation*, 102(10), 1064–1074.

[61] Cheng, X., & Wang, H. (2023). A generic model-free feature screening procedure for ultra-high dimensional data with categorical response. *Computer Methods and Programs in Biomedicine*, 229, 107269.

[62] Li, K., Wang, F., Yang, L., & Liu, R. (2023). Deep feature screening: Feature selection for ultra-high-dimensional data via deep neural networks. *Neurocomputing*, 538, 126186.

[63] Liu, Z., Elashoff, D., & Piantadosi, S. (2019). Sparse support vector machines with l0 approximation for ultra-high dimensional omics data. *Artificial Intelligence in Medicine*, 96, 134–141.

[64] Afshar, M., & Usefi, H. (2020). High-dimensional feature selection for genomic datasets. *Knowledge-Based Systems*, 206, 106370.

[65] Chamlal, H., Benzmane, A., & Ouaderhman, T. (2024). Elastic net-based high dimensional data selection for regression. *Expert Systems with Applications*, 244, 122958.

[66] Huang, C. (2021). Feature selection and feature stability measurement method for high-dimensional small sample data based on big data technology. *Computational Intelligence and Neuroscience*, 2021.

[67] Zambom, A. Z., & Matthews, G. J. (2021). Sure independence screening in the presence of missing data. *Statistical Papers*, 62, 817–845.

[68] Yi, G. Y., He, W., & Carroll, R. J. (2022). Feature screening with large-scale and high-dimensional survival data. *Biometrics*, 78(3), 894–907.

[69] Reese, R., Dai, X., & Fu, G. (2018). Strong sure screening of ultra-high dimensional categorical data. *arXiv Preprint arXiv:1801.03539*.

[70] Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35, 109–148.

[71] Li, F., Yang, M., Li, Y., Zhang, M., Wang, W., Yuan, D., & Tang, D. (2020). An improved clear cell renal cell carcinoma stage prediction model based on gene sets. *BMC Bioinformatics*, 21, 1–15.

[72] Sim, K. C., Han, N. Y., Cho, Y., Sung, D. J., Park, B. J., Kim, M. J., & *Han, Y. E.* (2023). Machine learning–based magnetic resonance radiomics analysis for predicting low-and high-grade clear cell renal cell carcinoma. *Journal of Computer Assisted Tomography*, 47(6), 873–881.

[73] Sofia, D., Zhou, Q., & Shahriyari, L. (2023). Mathematical and machine learning models of renal cell carcinoma: A review. *Bioengineering*, 10(11), 1320.

[74] Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, 89, 31–50.