Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 7 (2025) 2810

Journal of the Nigerian Society of Physical Sciences

Machine learning-based feature selection for ultra-high-dimensional survival data: a computational approach

Nahid Salma[a,b], Majid Khan Majahar Ali[a,*], Raja Aqib Shamim[a,c]

[a]School of Mathematical Sciences, Universiti Sains Malaysia, 11800, Pulau Penang, Malaysia
[b]Department of Statistics and Data Science, Jahangirnagar University, Savar, 1342, Dhaka, Bangladesh
[c]Department of Mathematics, University of Kotli, 11100, Azad Jammu and Kashmir, Pakistan

Abstract

Ultra-high-dimensional (UHD) survival data presents significant computational challenges in biomedical research, particularly in Renal Cell Carcinoma (RCC), where genomic complexity complicates risk assessment. Effective feature selection is crucial for identifying key biomarkers that improve RCC diagnosis, prognosis, and treatment. This study evaluates machine learning (ML)-based feature selection methods to address limitations in scalability, feature redundancy, and predictive accuracy in UHD RCC survival data. Gene expression data from 4,224 differentially expressed genes across 74 individuals was analyzed using LASSO, EN, Adaptive LASSO, Group LASSO, SIS, ISIS, SCAD, and SVM. Models were assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 values. SCAD demonstrated the best predictive performance (MSE: 529.00, RMSE: 23.00, R^2: 0.69), surpassing ISIS (R^2: 0.61), SIS (R^2: 0.60), and EN (R^2: 0.57). LASSO and Adaptive LASSO underperformed. SCAD identified 14 key genes—NCAM1, ATP1B3, NAT8, MT2A, GTF2F2, X4197, GUCY2C, SLC3A1, CRYZ, DES, MT1L, NFYB, PRKAR2B, and CLIP1—as potential RCC biomarkers. Gene interaction network analysis confirmed their role in RCC progression. Despite SCAD's strong performance, 31% of variability remained unexplained, underscoring the need for hybrid ML models. Combining deep learning approaches like CNNs with interpretable methods such as SCAD or Elastic Net could improve predictive accuracy and reveal complex gene interactions in RCC prognosis.This research supports SDG 3 (Good Health and Well-being) and SDG 9 (Industry, Innovation, and Infrastructure) by advancing precision medicine, early RCC detection, and biomedical data-driven innovations for improved clinical decision-making.

DOI:10.46481/jnsps.2025.2810

Keywords: Ultra-high dimension, Machine learning, Feature selection, Renal cell carcinoma, Survival data

Article History :
Received: 29 March 2025
Received in revised form: 22 May 2025
Accepted for publication: 23 May 2025
Available online: 08 June 2025

Communicated by: B. J. Falaye

1. Introduction

The rapid expansion of high-throughput computing and big data analytics has transformed computational biology and medical informatics. Modern biomedical research increasingly relies on ultra-high-dimensional (UHD) datasets, such as gene expression profiles, which contain thousands of features but are constrained by limited sample sizes. Processing such datasets presents significant computational challenges, including the curse of dimensionality, multicollinearity, high computational costs, and feature selection difficulties [1]. Traditional statistical methods struggle with these issues, often leading to overfitting, loss of relevant information, and reduced model generalization [2]. Machine learning (ML)-based feature selection

*Corresponding author: Tel. No.: +60-14-9543-405.
Email address: majidkhanmajaharali@usm.my (Majid Khan Majahar Ali)

has emerged as a powerful solution, improving dimensionality reduction, computational efficiency, and predictive accuracy in survival analysis [3, 4].

Despite these advancements, several computational challenges remain unresolved. Scalability is a major limitation, as increasing data dimensionality leads to excessive computational costs and memory inefficiencies [5]. Feature redundancy and correlation further complicate selection, as models like LASSO tend to favor one variable from correlated groups while discarding others, potentially omitting crucial information [6]. While deep learning models and ensemble learning techniques enhance predictive performance, they often lack interpretability, which is crucial for biomedical applications where model decisions must be explainable [7]. Furthermore, no standardized benchmarking framework exists for hybrid and ensemble-based feature selection methods, making it difficult to determine the optimal approach for UHD survival data [8]. Additionally, small biomedical datasets are prone to overfitting, and although regularization techniques help mitigate this risk, they can introduce biases that limit generalizability [5]. Lastly, the computational intensity of ML-based feature selection methods remains a major bottleneck, making real-time or streaming data analysis impractical [9].

To evaluate these computational challenges in a real-world biomedical context, we use Renal Cell Carcinoma (RCC) as a case study. RCC, the most common form of kidney cancer, accounts for 90% of kidney tumors [10] and 5% of all cancers worldwide [11]. In 2020, it contributed to 430,000 new cases and 179,368 deaths globally, with U.S. projections estimating 81,800 new cases and 14,890 deaths annually by 2025 [12–16]. RCC is often detected incidentally, but prognosis depends heavily on tumor stage and metastasis, with a five-year survival rate as low as 12% in advanced cases [15, 16]. The disease originates in the kidney's urine-producing tubules and frequently spreads to other organs, complicating treatment [17, 18]. Familial RCC, which accounts for 2–3% of cases, further increases the risk among first-degree relatives [19]. Genetic mutations in VHL [20, 21], c-MET [16], and PBRM-1 influence tumor growth, treatment response, and prognosis [22]. As RCC incidence continues to rise with advancements in diagnostics and imaging, there is a growing need for improved predictive models and targeted therapies [1].

Several ML-based approaches have been explored to enhance RCC diagnosis, prognosis, and survival prediction. Some studies have combined filtering methods (e.g., XGBoost, GBM, Rpart) with wrapper techniques (e.g., mRMR, RFE, Boruta) to identify gene signatures [7]. Others have applied RNA-seq data and Cox regression to construct survival risk scores [9] or leveraged CT scan-based AI models for RCC prognosis prediction [4]. Additional approaches include ML models for tumor classification [23, 24], texture-based RCC grading [24], and RCC post-surgical outcome prediction [5]. While these studies demonstrate the potential of ML in RCC research, they often lack scalability, standardized evaluation metrics, and real-time feasibility, leaving uncertainty about the most effective computational strategy for survival analysis.

This study presents a comprehensive benchmarking framework that systematically evaluates ML-based feature selection methods for UHD survival data, using RCC as an experimental case study. We apply and compare multiple ML-driven feature selection methods—including LASSO, Elastic Net (EN), Adaptive LASSO, Group LASSO, SCAD, SVM, SIS, and ISIS—to identify the most effective approach for reducing dimensionality and improving survival prediction models. The evaluation is based on key performance metrics such as Sum of Square Error (SSE), Mean Square Error (MSE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

By systematically benchmarking these methods, this research aims to enhance the efficiency and reliability of ML-driven feature selection in UHD datasets, addressing challenges in big data processing, feature redundancy, model generalization, and real-time feasibility. This work aligns with the United Nations Sustainable Development Goals (SDGs), particularly SDG 3 (Good Health and Well-being) and SDG 9 (Industry, Innovation, and Infrastructure). Enhancing computational feature selection methods will contribute to early disease prediction, precision medicine, and improved digital healthcare infrastructure, fostering advancements in both biomedical informatics and global healthcare innovation.

2. Methodology

2.1. Flowchart of the study

In order to achieve our goal, we meticulously created a study plan, which we adhered to religiously (Figure 1). Shortly after, the investigation began by importing an RCC dataset with the goal of extracting the optimal feature selection method of UHD RCC data. A few simple descriptive analyses, such as frequency analysis and chi-square testing, were performed on the dataset in order to gain a better understanding of it. After that, the UHD RCC gene-expression data was pre-processed and divided into a 70:30 ratio, with the remaining 30% being utilized to validate all methodologies and the remaining 70% being used for training. Eight machine learning feature selection techniques (LASSO, EN, Adaptive LASSO, Group LASSO, SCAD, SVM, SIS and ISIS) were subsequently applied to the UHD data and 10-fold cross-validation. SSE, MSE, RMSE, and R^2 were used to assess each of the ten feature selection techniques' performances. After selecting the best feature selection methodology, the most effective way was used to extract significant features. GeneMANIA was used to examine the gene interaction of the relevant RCC genes that were extracted.

2.2. Data description

The gene expression dataset, sourced from the R tool "kidpack," comprises 4224 differentially expressed gene entries across 74 individuals at risk for kidney cancer. It includes renal tumor samples with diverse histological types, grades, stages, chromosomal abnormalities, and survival data. Multiple tumor samples were pooled to establish a hybridization reference. Accessible via E-DKFZ-1 < ArrayExpress < BioStudies < EMBL-EBI, the dataset tracks survival outcomes, indicating whether a patient has died (1) or remains alive (0).

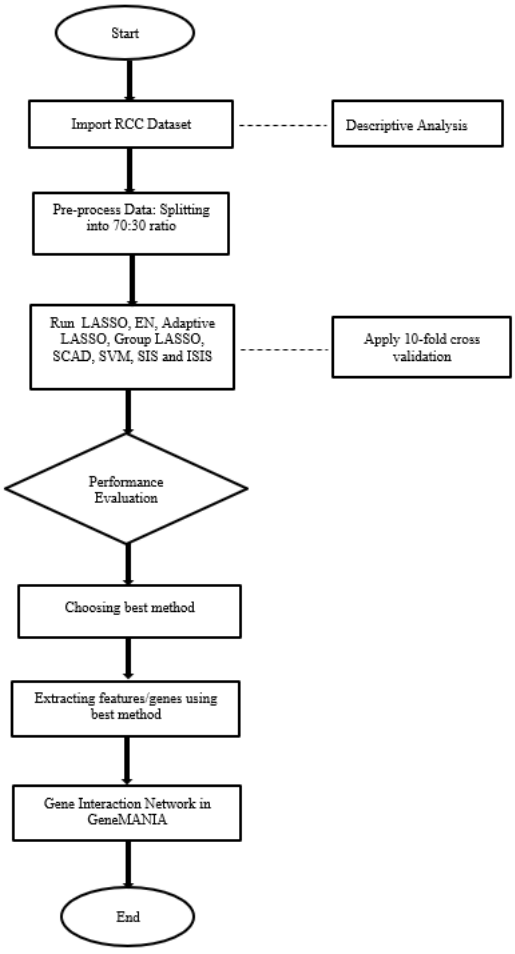

Figure 1: Overall methodology of the study.

2.3. Data preprocessing

The dataset was then split into training (70%) and validation (30%) sets, ensuring that all methodologies were evaluated on an independent test set to mitigate overfitting.

2.4. Feature selection methods

The choice of feature selection methods in this study was guided by the specific challenges associated with ultra-high-dimensional (UHD) genomic data, where the number of predictors (4,224 genes) far exceeds the number of observations (74 individuals). In such scenarios, it is critical to employ methods that can reduce dimensionality, control overfitting, and improve model generalizability. To this end, we applied a diverse set of eight feature selection techniques: LASSO, Elastic Net (EN), Adaptive LASSO, Group LASSO, SCAD, SVM, SIS, and ISIS. LASSO was employed due to its widely recognized ability to perform variable selection and regularization through L1 penalization, encouraging model sparsity [25, 26]. Elastic Net was selected to address multicollinearity by combining L1 and L2 penalties, which is especially useful in genomic data where many genes are highly correlated [27]. Adaptive LASSO extends the basic LASSO approach by introducing data-adaptive weights to enhance selection consistency and achieve oracle properties [22]. Group LASSO was included to allow for selection at the group level, reflecting potential biological structures such as gene pathways [28]. SIS and ISIS were used as efficient screening techniques. SIS filters variables based on their marginal correlation with the response, significantly reducing the dimensionality prior to model fitting [29, 30]. ISIS enhances SIS by iteratively refining variable selection to capture joint effects, improving robustness in identifying relevant predictors [30, 31]. Particular emphasis was placed on SCAD, a penalized regression method with a non-convex penalty that effectively balances sparsity and estimation accuracy. SCAD addresses limitations of LASSO by avoiding over-shrinkage of large coefficients, thus preserving the influence of truly significant predictors. Its oracle properties and reduced bias make it particularly suited for small-sample, high-dimensional contexts like ours [22, 28]. SVM was included for comparative purposes due to its strong performance in high-dimensional classification tasks, particularly through the maximization of decision boundaries. Although not a feature selection method in the traditional sense, linear SVMs provide interpretable coefficient estimates that can be used to identify important features. We acknowledge that SVM is sensitive to outliers, a known limitation [32, 33].

2.4.1. Least absolute shrinkage and selection operator (LASSO)

For the purpose of doing cross-validation analysis utilizing the L1 norm as a penalty mechanism and identifying important variables, the collected data were analyzed by the LASSO [25, 34]. In this big-data age shrinkage techniques are becoming more and more popular in biosystems due to their advantageous features for selecting factors and regularization. The LASSO technique computes coefficients by increasing log-partial likelihood while modifying the tuning value to control the penalty parameter. It is simple to state Lasso's general meaning as follows:

Let $\{1x_i, y_i\}, i = 1, 2, \ldots, N$ denote a sample of N Independent and Identically Distributed (IID) randomly generated vectors. Where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ signifies the row vector of data with respect to the p-explanatory factors of the i-th sample element and $1x_i \in R^p$ where $y_i \in R$ implies the corresponding respond vector. Thus, the LASSO algorithm estimator's overall shape is as follows:

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

where λ stands for both the multiplier of Lagrange and the penalty component. The matrix forms of the formula mentioned above can be further expressed as follows:

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

Y corresponds to the final results' column vectors of size $n \times 1$, X designates the matrix $n \times p$ holding the pertinent variables that

were discovered to be of concern, and, correspondingly, $\|\cdot\|_1$, $\|\cdot\|_2$ indicate the L1 and L2 vector norms.

2.4.2. Elastic net (EN)

Although LASSO performs well for many different variable selection problems, it breaks down when there are noticeably more predictors (p) than samples [19]. An improved version of LASSO was first proposed by [25] for managing strong correlations, provided that the maximum sample size for the total number of predictor variables selected is not exceeded and that there are strong correlations between several sets of variables: Elastic Net (EN) methodology. The EN uses correction phases L1-LASSO and L2-right, self-identifying the factors, and performs continuous shrinkage to improve forecasting accuracy. This technique removes irrelevant variables while keeping all the larger fish (important covariates) intact, much like a stretchy fishing net. This is the definition of the Elastic Net Estimator:

$$J(\beta, \lambda_1, \lambda_2) = \sum_{j=1}^{p} \left[\lambda_1 |\beta_j| + \lambda_2 \beta_j^2\right]. \tag{3}$$

The equation uses Lasso (left part of the above equation) and Ridge penalties (right part of the above equation) for sparse variables and highly correlated features, promoting average computation in linear models like regression or classification approaches [35]:

$$\begin{aligned}\hat{\beta}^{\text{elastic net}} =& (1+\lambda_2) \arg\min_{\beta \in R^p} \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\beta)^2 \right.\\ &\left. + \lambda_1 \sum_{j=1}^{p} |\beta_j^2| + \lambda_2 \sum_{j=1}^{p} \beta_j^2\right].\end{aligned} \tag{4}$$

The parameters of the elastic net, denoted by λ_1 and λ_2, necessitate an optimal ratio instead of just one value, frequently culminating in the combined value of the two parameters. To obtain an estimation of the elastic-net coefficient, the regression loss function is minimized using the elastic-net cost.

$$\sum_{j=1}^{p} \left[\alpha|\beta_j| + (1-\alpha)\beta_j^2\right] \leq \kappa, \tag{5}$$

where κ represents the additional α parameter, that makes up the sum of λ_1 and λ_2.

2.4.3. Adaptive LASSO

The Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive LASSO) is a regression method that enhances the standard LASSO by assigning adaptive weights to penalize coefficients differently. It retains the ability of LASSO to perform variable selection and regularization but mitigates some of its shortcomings, such as inconsistent variable selection when true coefficients are small [22]. Let us assume a linear regression model:

Let us assume a linear regression model:

$$y = X\beta + \epsilon, \tag{6}$$

where $y \in \mathbb{R}^n$: Response vector (dependent variable), $X \in \mathbb{R}^{n\times p}$ Design matrix with n observations and predictors, Coefficient vector, ϵ: Error term, assumed to be $N(0, \sigma^2)$.

The Adaptive LASSO estimator solves the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|\right), \tag{7}$$

where $\|y - X\beta\|_2^2 = \sum_{i=1}^{n}(y_i - X_i\beta)^2$: Residual sum of squares, $\lambda > 0$: Regularization parameter controlling the trade-off between the fit and penalty; $w_j > 0$: Adaptive weights for each coefficient β_j , computed as: $w_j = \frac{1}{|\hat{\beta}_j^{int}|^\gamma}$. With $\hat{\beta}_j^{int}$ being an initial estimate of β_j (e.g., from ordinary least squares or Ridge regression), and $\gamma > 0$ is a tuning parameter.

2.4.4. Group LASSO

Group LASSO is a regularization technique designed for scenarios where predictors (features) are naturally organized into predefined groups. Instead of penalizing individual coefficients as in standard LASSO, Group LASSO applies penalties at the group level. This approach ensures that entire groups of predictors are either selected or excluded together, making it suitable for problems where predictors within a group are correlated or have a shared interpretation. In group LASSO, the predictors are partitioned into G groups, $\{G_1, G_2, \ldots, G_G\}$ such that each group G_k corresponds to a subset of indices of β (e.g., group G_k contains indices $G_k\{j_1, j_2, \ldots, j_{|G_k|}\}$). The Group LASSO optimization problem is defined as [22]:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda \sum_{k=1}^{G} \|\beta_{Gk}\|_2\right). \tag{8}$$

Here $\|y - X\beta\|_2^2 = \sum_{i=1}^{n}(y_i - X_i\beta)^2$: Residual sum of squares, $\lambda > 0$: Regularization parameter controlling the trade-off between the fit and penalty, $\|\beta_{Gk}\|_2 = \sqrt{\sum_{j\in G_k}\beta_j^2}$: l_2 norm of the coefficients in group G_k, β_{Gk} : subsectors of β corresponding to the indices in group G_k. The key difference from standard LASSO lies in the penalty term. By summing the l_2-norms of the coefficient groups, Group LASSO encourages sparsity at the group level rather than individual coefficients [22, 28].

2.4.5. Smoothly clipped absolute deviation (SCAD)

SCAD is a regularization method designed to address some limitations of LASSO, such as excessive bias in large coefficients and inconsistent variable selection. SCAD achieves this by using a nonconvex penalty function that reduces bias for large coefficients while still maintaining sparsity for smaller ones. It is particularly useful for high-dimensional data and variable selection tasks. The SCAD estimator is obtained by solving the optimization problem [22, 28]:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{2n}\|y - X\beta\|_2^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\right). \tag{9}$$

Where $P_\lambda(|\beta_j|)$: SCAD penalty function, which depends on $\lambda > 0$:(regularization parameter) and $a > 2$ (a parameter controlling the nonconvexity). The SCAD penalty function is defined as:

$$P_\lambda(t) = \begin{cases} \lambda t, & \text{if} \quad 0 \le t \le \lambda, \\ -\frac{t^2 - 2a\lambda t + \lambda^2}{2(a-1)} & \text{if} \quad \lambda < t \le a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if} \quad t > a\lambda, \end{cases} \tag{10}$$

where $t = |\beta_j|$.

2.4.6. Support vector machines (SVM)

SVM are popular machine learning algorithms for classification and regression tasks. When applied to survival data, SVMs are adapted to handle censored data and predict survival times or risks. Survival data involves observations with two components: the survival time (T) and the event indicator δ, where $\delta = 1$ indicates the event (e.g., death, failure) and $\delta = 0$ indicates censoring. SVMs for survival analysis combine ideas from regression and survival models by creating a hyperplane or decision boundary in the feature space, optimized based on survival-specific loss functions. Let, $\{(x_i, T_i, \delta_i)\}_{i=1}^n$ represent the survival data, where: $x_i \in \mathbb{R}^p$: Feature vector for the i-th observation, T_i: Observed survival or censoring time, and δ_i is the Event indicator. The goal is to estimate a function $f(x)$ that relates the features x to the survival outcome [32, 33].

2.4.7. Sure independence screening (SIS)

SIS is a two-step process for selecting meaningful covariates for high-dimensional data [29, 30, 36]. This approach consists of two steps, which are as follows: Phase one: The preliminary screening stage in which marginal utilities are used to roughly examine the main consequences. In the second phase, known as the selection stage, variables and parameter estimation are done using a penalized regression with LASSO penalty. These are the steps that make up SIS:

1. Let's start by assuming that the sample size for the data acquisition $\{(x_i, y_i), i = 1, \ldots, n\}$ is n and $x \in R$. The following method can be used to calculate the covariate-specific marginal benefits $X_i, i = 1, 2, \ldots, p$:$L_m = \min_{\beta_0, \beta_j} \frac{1}{n} \sum L(y_i\beta_0 + x_m\beta_m)$; where $L(\cdot, \cdot)$ is the representation of a generalized loss function. More simply, fit p bivariate models, like the generalized linear model (GLM), to determine the marginal benefits.
2. To find the utility, the partial likelihood of each parameter is maximized in the following way:

$$U_m = \max_{\beta_m} \left(\sum_{i=I} \delta x_{im} \beta_m - \sum_{i=I} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp(x_{im}\beta_m) \right\} \right), \tag{11}$$

where $R(y_i)$ hazards are established before the occurrence y_i. $x_i m$ is the m-th factor is an indicator of censorship that present among the p components. Assemble the covariates according to this marginal utility in an ascending sequence. Thus, the most important covariate is the one with the minimal marginal value, U for the feature or variable.
3. According to their respective marginal utilities, arrange the indicators in order of chronology. Hence, based on the feature or variable's marginal utility, the most significant predictors are those with the lowest L_j.
4. Describe the first few d attributes. The formula $d = \lfloor n/\log(n) \rfloor$ is widely used, where $\lfloor \cdot \rfloor$ is the floor function. $\tilde{A}$ is thus defined as a subset of components that have been pre-approved.
5. The final step in the SIS technique is to estimate the model's parameters of the regression with repercussions as the following demonstrates:

$$\left(\hat{\beta}_0, \hat{\beta}_m\right) = \arg \min_{(\beta_0, \beta_m) \in R^{d+1}} \frac{1}{n} \sum_{i=1}^{n} L\left(i, \beta_0 + x_{i,M^*}\beta_{M^*}\right) + \sum_{j \in M^*} \lambda\left(|\beta_j|\right), \tag{12}$$

where $x_{i,M^*} \in R_d$ indicates that the sub-vector returned $x_i \in R_p$ through $d < p$ specified variables M^*. $\lambda\left(|\beta_j|\right)$ stands for the LASSO penalty, and [30] explains why the approach's name (SIS) makes sense. When d is sufficiently large, there is a good likelihood that the first sorting phase of the procedure stated will choose all the important predictors. The penalized equation of LASSO, that additionally evaluates the main effects of the other covariates, is used in the following stage to choose the variables.

2.4.8. Iterative sure independence screening (ISIS)

One major problem with the SIS approach is that if elements are disregarded in the first round, they won't be found in the subsequent one. Stated differently, if a marker is jointly connected with the outcome but marginally unconnected, or if a predictor is simultaneously uncorrelated but has a stronger peripheral association to the result over specific relevant components in the section [29, 37]. Introduced by [38], ISIS is a continuous SIS technique meant to fortify SIS and deal with the previously listed problems. The following is a summary of the ISIS technique's workflows, per [30, 31, 39–42]:

1. Using the SIS technique, all statistically significant factors are retrieved with a likelihood of one. Nevertheless, when several significant factors are only weakly uncorrelated by the response, ISIS approach is used [39, 43].
2. While an index $\hat{I}_1$ is selected using the Sure independence screening procedure, the iterative SIS uses a penalty-based choosing of features step to produce regression parameter estimations β_{i1}. The variation in the positive portion of $\hat{\beta}_{i1}$ affects the estimate $\hat{M}_1$ in $\hat{I}_1$. The following describes the coefficient's m conditional utility, assuming

that M is not part of the covariate.

$$U_{m|\hat{M}_1} = \max_{\beta_m \beta_{\hat{M}}} \left(\sum_{i=1}^{n} \delta_i \left(x_{im}\beta_m + x^T_{\hat{M}_1,i}\beta_{\hat{M}_1} \right) - \sum_{i=1}^{n} \delta_i \log \left(\sum_{j \in R(y_i)} \exp \left(x^T_{jm}\beta_m + x^T_{\hat{M}_1,j}\beta_{\hat{M}_1} \right) \right) \right). \tag{13}$$

3. Using the following formula, we calculate each factor's marginal utility in order to apply the second SIS step within this ISIS step.
4. Utilizing penalized regression, researchers minimize the previously described equation to determine the parameters of the model. The outcome of using the penalized regression approach is as follows: a median model that bears a striking resemblance to the actual model

$$\begin{aligned} &- \sum_{i=1}^{n} \delta_i \left(x^T_{\hat{M}_1 \cup \hat{I}_{2,i}} \beta_{\hat{M}_1 \cup \hat{I}_2} \right) \\ &+ \sum_{i=1}^{n} \delta_i \log \left\{ \sum_{j \in R(y_i)} \exp \left(x^T_{\hat{M}_1 \cup \hat{I}_{2,j}} \beta_{\hat{M}_1 \cup \hat{I}_{2,j}} \right) \right\} \\ &+ \sum_{m \in \hat{M}_1 \cup \hat{I}_2} P_\lambda(\beta_j). \end{aligned} \tag{14}$$

The magnitudes of $\beta_{\hat{M}_1 \cup \hat{I}_2}$ that are bigger than zero result in a smaller subset $\hat{M}_2$ of the selected factors.

5. Finally, we execute steps 3 and 4 when we reach the set or D's specified set, that is, $\left(\hat{M}_j = \hat{M}_{j-1}\right)$.

2.5. Performance evaluation

Assessing the accuracy and performance of a model is a fundamental aspect of regression analysis. In this study, Mean Square Error (MSE), Sum of Square Errors (SSE), Root Mean Square Error (RMSE), and Coefficient of determination or R-squared are employed as key evaluation metrics to determine the model's reliability and predictive accuracy. These metrics facilitate the comparison of different regression models, helping to identify the one that best fits the data while achieving the desired level of prediction accuracy. Generally, lower values of MSE, SSE, and RMSE indicate higher prediction accuracy [44], whereas a higher R-squared value suggests a better fit between the model and the data. Following the guidelines outlined by Arsad [45], where R-squared values are classified into specific ranges—85%-100% as very good, 70%-85% as good, 50%-70% as reasonably good, 30%-50% as reasonably bad, 15%-30% as bad, and 0%-15% as very bad—this study uses these classifications to assess the quality of the models, reflecting a decreasing ability to explain data variability.

The formulas for these evaluation metrics are presented in Table 1, where y_i represents the actual observations, $\hat{y}_i$ denotes the predicted values, $\bar{y}$ is the mean of all observations, and n represents the total number of observations. To further enhance the reliability of our results, we implemented 10-fold cross-validation, ensuring that model performance is not influenced by data partitioning biases.

2.6. Gene interaction network analysis using GeneMANIA

GeneMANIA (http://genemania.org) is a versatile and easy-to-use website designed to help generate hypotheses about gene function, analyze gene lists, and prioritize genes for functional assays. When provided with a list of query genes, GeneMANIA identifies genes with similar functions by leveraging a wide range of genomics and proteomics data. It assigns weights to each functional genomic dataset based on its predictive value for the query. Additionally, GeneMANIA can predict gene functions by finding genes that are likely to share similar functions with a single query gene based on their interactions. Researchers utilize GeneMANIA to identify related genes, visualize interaction networks, and prioritize candidate genes for further investigation. This tool aids in uncovering potential gene associations that may contribute to disease mechanisms and therapeutic development [46, 47].

3. Results and discussion

3.1. Demographic data

Table 2 illustrates the demographic and clinical characteristics of the patient cohort and their association with survival status. The analysis reveals notable differences in clinical and demographic characteristics between patients who were deceased ("Dead") and those who were alive ("Alive") at the time of the study. Tumor type showed a significant association with survival status. Clear cell RCC (ccRCC) was the most prevalent type among deceased patients, accounting for 88.89%, compared to 69.05% among alive patients. Chromophobe RCC (chRCC) was absent among deceased patients but constituted 7.14% of alive patients. Papillary RCC (pRCC) was also less common among deceased patients (11.11%) compared to alive patients (23.81%). These differences were statistically significant with a p-value of <0.001. Sex distribution did not show a significant association with survival. Among deceased patients, males constituted 61.11%, while females accounted for 38.89%. Among alive patients, males were slightly more prevalent at 71.42%, with females at 28.58%. The p-value for this comparison was 0.706, indicating no significant difference in survival based on sex. Tumor grade also did not show a statistically significant association with survival status. Deceased patients primarily had Grade 2 tumors (72.22%), with the remainder being Grade 3 (27.78%) and none in Grade 1. In contrast, alive patients were distributed across all grades, with Grade 2 being most frequent (73.81%), followed by Grade 1 (19.05%) and Grade 3 (7.14%). The p-value for this comparison was 0.084. Clinical stage demonstrated a strong association with survival. The majority of deceased patients were in Stage 4 (83.33%), with smaller proportions in Stages 3 (11.11%) and 1 (5.56%). Alive patients had a more even distribution across stages, with 28.57% in Stage 1, 14.29% in Stage 2, 42.85% in Stage 3, and 14.29% in Stage 4. These differences were highly significant, with a p-value of <0.001. Disease progression was another factor significantly associated with survival. All deceased patients experienced disease progression (100%), whereas alive patients were predominantly

Table 1: Equations for performance metrics (as in Ref. [44]).

Used Metrics	Description
Mean Square Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
Sum of Square Error (SSE)	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
Root Mean Square Error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{MSE}$
R-squared	$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$

non-progressive (59.52%), with 38.10% showing progression and 2.38% having missing progression data. This difference was statistically significant, with a p-value of <0.001. Tumor size and age were not significantly associated with survival. The mean tumor size was slightly larger in deceased patients (7.78 cm, SE = 0.824) compared to alive patients (7.11 cm, SE = 0.53), but this difference was not significant (p-value = 0.901). Similarly, the mean age of deceased patients was 61 years (SE = 2.79), comparable to 61.93 years (SE = 1.81) for alive patients, with a p-value of 0.519. Finally, survival time showed a highly significant difference between the groups. Deceased patients had a much shorter mean survival time of 12.61 years (SE = 4.25) compared to 24.17 years (SE = 2.69) for alive patients, with a p-value of <0.001. This emphasizes the critical role of factors like tumor type, clinical stage, and disease progression in influencing patient outcomes, highlighting the importance of early diagnosis and effective management in RCC. The findings of this study are consistent with those of previous studies [2, 7, 9, 23, 24].

3.2. Performance evaluation

This study seeks to identify the most effective machine learning feature selection method for ultra-high-dimensional (UHD) RCC survival data, with a particular focus on related genes. A range of advanced feature selection techniques—including LASSO, Elastic Net (EN), Adaptive LASSO, Group LASSO, SCAD, SVM, SIS, and ISIS—were applied and thoroughly evaluated to assess their performance in improving RCC prognosis and patient outcomes.

Table 3 presents a comparative analysis of various feature selection methods applied to the RCC dataset, evaluating their performance across four key metrics: Sum of Squared Errors (SSE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). Among these methods, SCAD demonstrates the highest performance, achieving the lowest SSE (39,146.00), MSE (529.00), and RMSE (23.00), along with the highest R^2 value of 0.69. This indicates that SCAD is the most effective in capturing variance within the dataset, explaining 69% of the variability in RCC-related features [45]. ISIS follows closely, with SSE (42,846.74), MSE (579.01), RMSE (24.06), and R^2 (0.66), demonstrating strong predictive capabilities, benefiting from an integrated selection process that enhances accuracy [30, 36]. SIS also performs well, achieving SSE (46,250.00), MSE (625.00), RMSE (25.00), and an R^2 of 0.64, highlighting its effectiveness in selecting informative features while maintaining relatively low error values. However, while SIS is computationally efficient, it may overlook important feature interactions [22].

In contrast, ISIS addresses this limitation by iteratively refining feature selection, capturing complex dependencies, and improving predictive accuracy [44]. Elastic Net (EN) and LASSO show competitive performance, with EN achieving SSE (48,304.72), MSE (652.78), RMSE (25.55), and R^2 (0.57), while LASSO follows with SSE (48,400.44), MSE (654.06), RMSE (25.59), and R^2 (0.53). These results suggest that EN slightly outperforms LASSO by reducing errors and explaining more variance. This is evident from the fact that as the data dimensionality increases, LASSO's performance declines [27]. Moreover, since genes are often interrelated, the dataset comprises 4,224 differentially expressed genes, which may lead to multicollinearity. A previous study [46] found that LASSO's effectiveness is significantly impacted by strong correlations among both relevant and irrelevant features. In such cases, Elastic Net (EN) offers a solution [27]. By incorporating both L1 (LASSO) and L2 (ridge) penalties, EN identifies important variables while maintaining continuous shrinkage, enhancing predictive accuracy. As a result, EN outperformed LASSO in this scenario. Adaptive LASSO exhibits moderate performance, with SSE (48,849.62), MSE (660.13), RMSE (25.69), and R^2 (0.54), offering slight improvements over standard LASSO by assigning adaptive penalties to features [22]. Group LASSO, however, performs less effectively, with SSE (53,263.01), MSE (719.77), RMSE (26.83), and an R^2 of 0.54, indicating higher error rates and a reduced ability to explain variance. SVM achieves an SSE of 49,393.01, an MSE of 667.43, an RMSE of 25.83, and an R^2 of 0.55, performing comparably to EN and Adaptive LASSO [22].

The consistently low R^2 values across all methods suggest the presence of outliers in the dataset, which is common, as one to ten percent of an actual dataset may contain outliers, according to Ref. [1, 48]. Outliers are data points that deviate significantly from the overall pattern of the data and can disproportionately influence statistical models, potentially distorting results [44, 49]. Given the presence of 4,224 genes, multicollinearity is also expected, as ultra-high-dimensional datasets

Table 2: Demographic characteristics of patients and association with their status.

Characteristics		Patient's Status		p-value
		Dead	Alive	
Type				
	ccRCC	16 (88.89)	29 (69.05)	
	chRCC	0 (0)	3 (7.14)	<0.001
	pRCC	2 (11.11)	10 (23.81)	
Sex		n (%)	n (%)	
	Male	11 (61.11)	30 (71.42)	0.706
	Female	7 (38.89)	12 (28.58)	
Grade		n (%)	n (%)	
	1	0 (0)	8 (19.05)	0.084
	2	13 (72.22)	31 (73.81)	
	3	5 (27.78)	3 (7.14)	
Clinical Stage		n (%)	n (%)	
	1	1 (5.56)	12 (28.57)	
	2	0 (0)	6 (14.29)	<0.001
	3	2 (11.11)	18 (42.85)	
	4	15 (83.33)	6 (14.29)	
Progress		n (%)	n (%)	
	0	0 (0)	25 (59.52)	
	1	18 (100)	16 (38.10)	<0.001
	NA	0	1 (2.38)	
Tumor Size	(cm)	Mean = 7.78 SE= 0.824	Mean=7.11 SE= 0.53	0.901
Age	(year)	Mean = 61 SE = 2.79	Mean = 61.93 SE =1.81	0.519
Survival Time	(year)	Mean =12.61 Se = 4.25	Mean =24.17 SE =2.69	<0.001

Table 3: Performance of feature selection methods of RCC.

Methods	SSE	MSE	RMSE	R^2
LASSO	48400.44	654.06	25.59	0.53
EN	48304.72	652.78	25.55	0.57
Adaptive LASSO	48849.62	660.13	25.69	0.54
Group LASSO	53263.01	719.77	26.83	0.54
SCAD	39146.00	529.00	23.00	0.69
SVM	49393.01	667.43	25.83	0.55
SIS	46250.00	625.00	25.00	0.64
ISIS	42846.74	579.01	24.06	0.66

often exhibit correlations between features [44, 48, 50]. Multicollinearity refers to a situation in which two or more predictor variables in a model are highly correlated, making it difficult to determine the individual contribution of each variable [38, 51]. The behavior of these methods varies significantly in the presence of outliers and multicollinearity. LASSO is particularly sensitive to outliers, as its L1 penalty tends to heavily penalize coefficients associated with extreme values. It also struggles with multicollinearity, selecting only one feature from correlated groups, which may result in the loss of important information [48, 51, 52]. Elastic Net (EN) is more robust to multicollinearity due to its combination of L1 and L2 penalties, though it remains somewhat influenced by outliers [49, 50]. Adaptive LASSO offers improved handling of multicollinearity by assigning different penalties to features but remains vulnerable to outliers [22, 28]. Group LASSO addresses multicollinearity by selecting entire groups of correlated variables, though outliers within any group can lead to incorrect selection [53, 54]. SCAD is less sensitive to outliers compared to LASSO, as its non-convex penalty reduces bias, and it better handles multicollinearity by allowing larger coefficients for relevant features [28, 32, 53, 55]. SVM are highly sensitive to outliers, as they can drastically shift the decision boundary, and multicollinearity can reduce the model's ability to distinguish between classes [33, 38]. SIS is less robust to outliers and may miss key interactions between correlated features, whereas Integrated ISIS, which combines SIS with methods like LASSO or EN, improves robustness but still inherits some weaknesses [48, 50, 56].

Overall, methods such as Elastic Net, SCAD, and Group LASSO handle multicollinearity more effectively, while LASSO and SVM are more affected by outliers, requiring careful tuning to mitigate these challenges [22, 28, 32, 54, 56]. Based on the findings of this study, SCAD emerges as the top-performing method for reducing prediction errors and explaining variance in the RCC dataset, though its relatively low explanatory power limits its suitability for ultra-high-dimensional

survival data. ISIS and SIS also perform well, but further refinement is needed to enhance accuracy.

3.3. Selected genes identified by all considered methods

Table 4 presents the genes selected by the considered methods, listing both the number of selected genes and their corresponding names.

LASSO selects a larger set of eight genes, including NCAM1, PRKAR2B, CLIP1, NFYB, GUCY2C, MT1L, DES, and ATP1B3. EN stands out by selecting the largest set of 27 genes, such as PFDN5, NCAM1, GTF2H3, RHOG, ZNF148, PRKAR2B, STK17A, LUM, MT1X, CLIP1, ANXA13, NFYB, GUCY2C, SRSF11, SLC3A1, ALG2, ACAA2, MT1L, MT2A, DES, SLC6A3, NAT8, ATP1B3, OVGP1, and ZNF783. Adaptive LASSO selects 13 genes, including RPA3, FGF1, RHOG, CDKL1, CLIP1, GUCY2C, TUBA3C, PNMA2, DES, NAT8, ATP1B3, and PPP4C. Group LASSO, on the other hand, selects 10 genes, focusing on NCAM1, PRKAR2B, MT1X, ANXA13, NFYB, MT2A, MT1L, DES, NAT8, and ATP1B3. SCAD selects 9 genes, including NCAM1, ATP1B3, NAT8, MT2A, GTF2F2, GUCY2C, SLC3A1, CRYZ, and MT1L. The SIS method selects 20 genes, covering a broad range such as CTNNB1, RPA3, ROR1, RAC2, CDKL1, GMFB, FGF5, E2F6, PAX2, NFYB, GUCY2C, HGD, MT2A, IL2RG, DCI, COX4, FGFR3, and others. ISIS selects 7 genes, including NCAM1, MITF, OSBP, APOB, FGFR3, RHOG, and GUCY2C. Lastly, the SVM method selects 6 genes with identifiers such as GATA3, CDKN2C, RPLP1, KIAA0281, and EST (Expressed Sequence Tag). This variation in selected genes demonstrates the differences in sensitivity and selection criteria across methods, highlighting how each approach captures distinct features in the dataset, particularly when handling high-dimensional survival analysis in RCC.

3.4. Overlapping genes across the feature selection methods

The heatmap (Figure 2) visualizes gene selection consistency across various Feature Selection Methods, with Gene Names on the Y-axis and methods on the X-axis. It employs a cool-warm color scale, where darker red hues indicate the presence (1) of a gene in multiple methods, reflecting higher agreement and potential biological significance in RCC prognosis. Conversely, lighter shades or white represent absence (0) or selection by fewer methods, suggesting lower consensus or method-specific relevance.

Notably, genes such as NCAM1, ATP1B3, NAT8, GUCY2C, CLIP1, DES, and MT2A exhibit strong red coloration across several methods, reinforcing their consistent selection and highlighting their potential as robust prognostic markers or therapeutic targets. The methods LASSO, EN, Group LASSO, and SCAD show dense red patterns, indicating substantial overlap in gene selection, which may reflect methodological similarities or shared sensitivity to key biological signals. In contrast, SVM, shown in white across all genes, did not select any overlapping genes, suggesting distinct selection criteria or lower sensitivity to these signals. Additionally, scattered red cells in methods like SIS and Adaptive LASSO reflect more individualized selection behavior. This color distribution effectively visualizes how different methods converge on key genes while also revealing method-specific differences. The intensity and clustering of red shades help identify the most consistently selected genes, which may serve as robust prognostic markers or therapeutic targets in RCC.

The most overlapping genes across multiple feature selection methods, as indicated by the heatmap, include NCAM1, ATP1B3, NAT8, GUCY2C, DES, MT2A, CLIP1, PRKAR2B, NFYB, GTF2F2, SLC3A1, and MT1L. NCAM1 is the most consistently identified gene, being selected by LASSO, EN, Group LASSO, SCAD, and ISIS, highlighting its significant role in RCC progression. Similarly, ATP1B3 appears across LASSO, EN, Adaptive LASSO, Group LASSO, and SCAD, emphasizing its importance in various cancer types. NAT8, GUCY2C, and DES also show strong overlap, being selected by multiple methods such as EN, Adaptive LASSO, Group LASSO, and SCAD, underlining their potential as robust biomarkers. MT2A and CLIP1, although slightly less recurrent, still demonstrate significant selection across methods like EN, Group LASSO, and SCAD, suggesting their relevance in RCC. These overlapping genes, identified through the heatmap, are key candidates for further research and may serve as critical prognostic markers or therapeutic targets in RCC.

3.5. Final Selected Genes and Their Connection With RCC

Based on the above analysis, this study considered the genes selected by the best-performing method, SCAD, along with the most overlapping genes across different methods, identifying 14 key genes: NCAM1, ATP1B3, NAT8, MT2A, GTF2F2, X4197, GUCY2C, SLC3A1, CRYZ, DES, MT1L, NFYB, PRKAR2B, and CLIP1. These genes play crucial roles in various biological processes, including cell adhesion, ion transport, transcription regulation, and cancer progression, with several serving as prognostic markers in RCC and other malignancies. NCAM1 (X134), a cell adhesion molecule in the immunoglobulin superfamily, is essential for cell interactions, migration, and immune regulation, activating key signaling pathways such as MAPK and PI3K. It has been implicated in glioblastoma multiforme and pancreatic adenocarcinoma [57]. ATP1B3 (X3671), a subunit of Na+/K+-ATPase, is vital for maintaining ion gradients and cell excitability, making it significant in multiple cancers, including cervical, liver, and glioblastoma [58]. NAT8 (X3218), primarily expressed in the kidney and liver, is associated with cell adhesion and tissue-specific expression and has been identified as a prognostic marker in kidney renal clear cell carcinoma (KIRC) [59]. MT2A (X2663), a member of the metallothionein family, plays a role in metal homeostasis, detoxification, and oxidative stress response, and it has been linked to multiple cancers, including glioblastoma, lung, liver, pancreatic, and RCC [60]. GTF2F2 (X1247) is involved in RNA polymerase II transcription initiation and serves as a prognostic marker in RCC, cervical, and liver cancers [61]. However, limited information is available for the variable X4197 in the dataset for which its gene name could not be identified, suggesting the need for further investigation. GUCY2C (X1728), a transmembrane receptor involved

Table 4: Selected genes by considered methods.

Methods	NGS	Selected Variables (Genes)	Gene Name
LASSO	8	X134, X524, X1364, X1654, X1728, X2465, X3206, X3218, X3671	NCAM1, PRKAR2B, CLIP1, NFYB, GUCY2C, MT1L, DES, NAT8, ATP1B3
EN	27	X128, X134, X317, X358, X491, X524, X936, X1037, X1136, X1364, X1558, X1654, X1728, X1838, X2027, X2189, X2235, X2335, X2397, X2465, X2663, X2679, X3206, X3215, X3218, X3671, X3696, X3890	PFDN5, NCAM1, GTF2H3, RHOG, ZNF148, PRKAR2B, STK17A, LUM, MT1X, CLIP1, ANXA13, NFYB, GUCY2C, SRSF11, SLC3A1, ALG2, ACAA2, No info, No info, MT1L, MT2A, No info, DES, SLC6A3, NAT8, ATP1B3, OVGP1, ZNF783
Adaptive LASSO	13	X67, X168, X358, X428, X1364, X1728, X1740, X1887, X2389, X3206, X3218, X3671, X3951	RPA3, FGF1, RHOG, CDKL1, CLIP1, GUCY2C, TUBA3C, PNMA2, No info, DES, NAT8, ATP1B3, PPP4C
Group LASSO	10	X134, X936, X1364, X1654, X1728, X2465, X2663, X3206, X3218, X3671	NCAM1, PRKAR2B, MT1X, ANXA13, NFYB, MT2A , MT1L, DES, NAT8, ATP1B3
SCAD	9	X134, X3671, X3218, X2663, X1247, X4197, X1728, X2027, X146	NCAM1, ATP1B3, NAT8, MT2A, GTF2F2, No info, GUCY2C, SLC3A1, CRYZ
SIS	20	X29, X67, X135, X252, X428, X687, X690, X1352, X1499, X1654, X1728, X2017, X2036, X2465, X2664, X2745, X3004, X3181, X3405, X3946	CTNNB1, RPA3, ROR1, RAC2, CDKL1, GMFB, FGF5, E2F6, PAX2, NFYB, GUCY2C, No info, HGD, MT2A, IL2RG, DCI, COX4, No info, FGFR3, No info
ISIS	7	X134, X1156, X1866, X2249, X2831, X3341, X358	NCAM1, MITF, OSBP, APOB, FGFR3, No info, RHOG
SVM	6	X133, X1246, X2026, X2662, X3217, X3670	GATA3, CDKN2C, No info, RPLP1, KIAA0281, EST (Expressed Sequence Tag)

in ion transport regulation, has mutations associated with hereditary diarrheal disorders and serves as a prognostic indicator in liver hepatocellular carcinoma [62]. SLC3A1 (X2027), a transporter of amino acids in the renal tubule, has mutations linked to cystinuria and functions as a prognostic marker in both clear cell and papillary RCC subtypes [63]. Lastly, CRYZ (X146), a crystallin protein with enzymatic activity, plays a role in maintaining eye lens transparency and possesses NADPH-dependent quinone reductase activity, making it a prognostic marker in glioblastoma and RCC [64]. This analysis underscores the biological significance of these genes in RCC progression, emphasizing their potential as prognostic markers and therapeutic targets.

The most overlapping genes across multiple feature selection methods, as indicated by the heatmap, include NCAM1, ATP1B3, NAT8, GUCY2C, DES, MT2A, and CLIP1. NCAM1 is the most consistently identified gene, being selected by LASSO, EN, Group LASSO, SCAD, and ISIS, highlighting its significant role in RCC progression have emerged as significant prognostic markers in various cancers, including RCC. Each gene plays a distinct role that contributes to disease progression and could serve as a potential target for therapeutic intervention. The roles of NACM1, NAT8, GUCY2C have already been discussed. Another critical gene, CLIP1 (CAP-Gly Domain Containing Linker Protein 1), is involved in microtubule dynamics and intracellular trafficking, making it a noteworthy prognostic indicator in RCC [65]. Similarly, TRMT1L (tRNA Methyltransferase 1 Like), which is associated with tRNA modification and affects protein translation, has been linked to RCC, further emphasizing its relevance [66]. Finally, DES (Desmin), a type III intermediate filament protein crucial for muscle integrity, has shown promise as a prognostic marker in various cancers, including bladder urothelial carcinoma, glioblastoma multiforme, kidney renal papillary cell carcinoma, and lung squamous cell carcinoma [67]. The repeated identification of NCAM1 across multiple methods underscores its potential significance in RCC progression, while NAT8 and CLIP1 also emerge as important markers in RCC. The association of TRMT1L and DES with several cancers further supports their relevance.

Previous studies on RCC and other healthcare disorders often overlooked the use of ML models. For example, [68, 69]

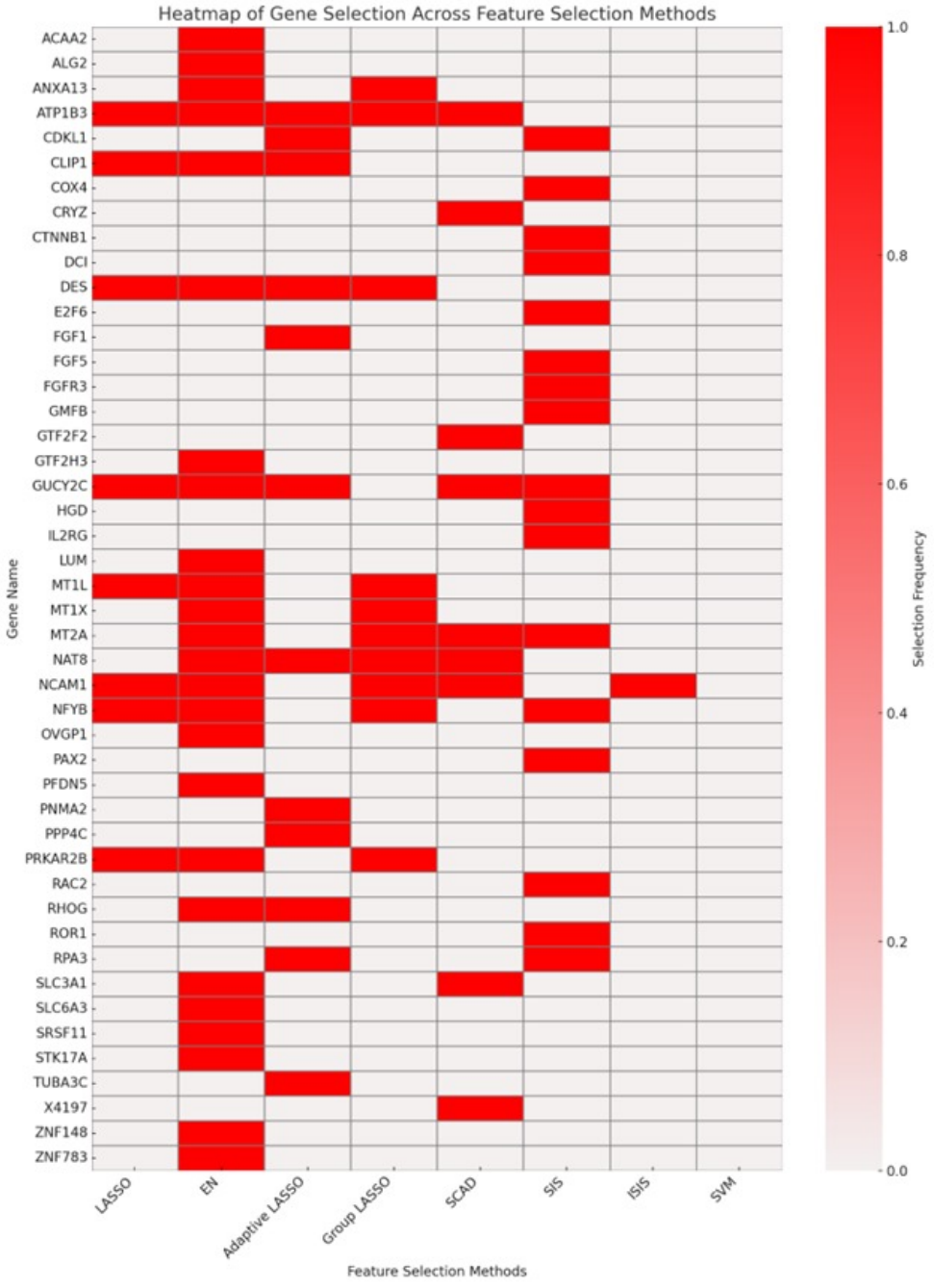

Figure 2: Heatmap of overlapping genes across feature selection methods.

identified 32 genes via microarray analysis without incorporating machine learning, while [70] pinpointed six hub genes (e.g., SUCLG1, PCK2, GLDC) linked to reduced survival in RCC patients. Berglund et al. (2020) identified nine genes, six of which had prior associations with RCC. These studies, although valuable, relied on conventional models that lacked the computational depth required for ultra-high-dimensional datasets. More recent efforts have incorporated machine learning: [9] utilized a random forest algorithm to identify a five-gene signature, and [42] applied eight machine learning models to discover a 13-gene signature. An ensemble-based approach by [8] identified just two genes, NOP2 and NSUN5, while [25] used LASSO-SVM RFE to pinpoint four key genes (ACPP, ANGPTL4, SCNN1G, SLC22A7) in KIRC-related datasets. Notably, [44] identified 49 genes using RLF-ISIS, though without accounting for outliers, demonstrating that the relevance of selected genes is often more critical than the total number identified. However, unlike previous studies, none of them utilized as many machine learning approaches as we did in this study, where we applied a total of eight different methods.

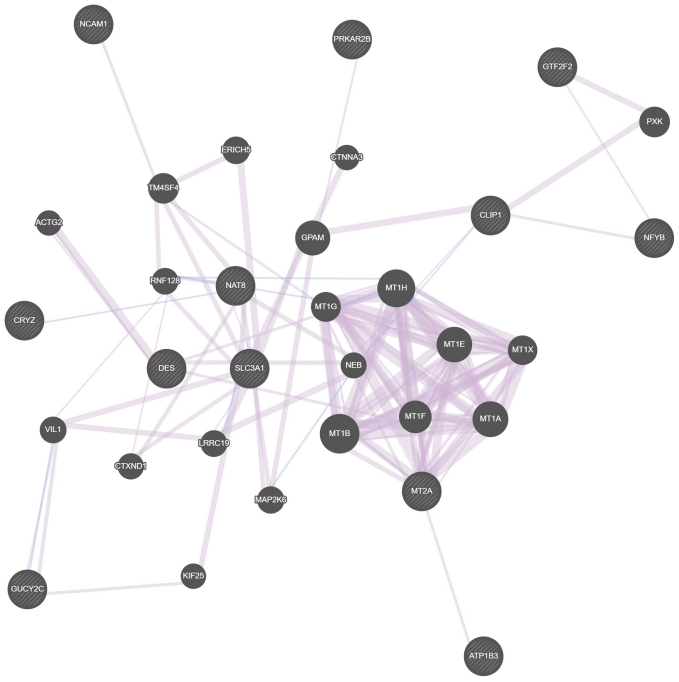

Figure 3: Gene interaction network of 14 SCAD-selected genes and common genes across all methods.

3.6. Gene interaction network analysis

Further analysis using GeneMANIA revealed interactions among these 14 genes, showcasing their interconnectedness through co-expression, co-localization, genetic and physical interactions, pathway involvement, and shared protein domains. Figure 3 presents the GeneMANIA-derived interaction network, highlighting the nine genes identified by SCAD alongside the common genes selected across all feature selection methods. In this network, nodes represent gene names, while edges depict their functional relationships based on various interaction types. The visualization reinforces the potential functional relationships of these selected genes in RCC development and progression. This study suggests that these genes play a direct and indirect role in RCC development and progression. Their consistent identification across multiple methods highlights their potential significance in RCC pathology and therapeutic strategies, underscoring the need for further research to elucidate their individual and collective roles as potential therapeutic targets.

Finally, it can be concluded that this study highlights 14 genes—NCAM1, ATP1B3, NAT8, MT2A, GTF2F2, X4197, GUCY2C, SLC3A1, CRYZ, DES, MT1L, NFYB, PRKAR2B, and CLIP1—identified through the SCAD method and an overlapping gene list, suggesting their direct and indirect involvement in RCC development and progression.

4. Conclusion

This study systematically evaluated ML-based feature selection methods for UHD survival data, using RCC as a case study. By benchmarking LASSO, Elastic Net (EN), Adaptive LASSO, Group LASSO, SCAD, SVM, SIS and ISIS, we identified SCAD as the most effective method, achieving the best predictive performance (MSE: 529.00, RMSE: 23.00, R^2: 0.69). However, SCAD left 31% of data variability unexplained, highlighting the need for hybrid ML models that integrate multiple feature selection techniques to improve gene selection, prediction accuracy, and RCC prognosis. This study identified 14 key genes—NCAM1, ATP1B3, NAT8, MT2A, GTF2F2, X4197, GUCY2C, SLC3A1, CRYZ, DES, MT1L, NFYB, PRKAR2B, and CLIP1—as potential RCC biomarkers. Gene interaction network analysis confirmed their involvement in RCC progression, reinforcing their relevance for future biomarker discovery and clinical applications. Additionally, tumor type, stage, and progression significantly influenced survival outcomes. Clear cell RCC (ccRCC) was the most prevalent among deceased patients, while chromophobe RCC (chRCC) and papillary RCC (pRCC) exhibited better survival rates.

While this study provides valuable insights, several limitations remain. One major challenge is the reliance on publicly available datasets, which may have constraints in sample size, diversity, and data quality. These factors can limit the generalizability of the findings due to potential biases in patient demographics and tumor characteristics. Moreover, the R^2 values suggest room for improvement, indicating the potential benefit of more complex models or additional features. Additionally, the study focused on a specific set of feature selection techniques, and exploring alternative approaches may yield better results. Future research should prioritize the development of hybrid ML frameworks that combine the strengths of classical statistical models with advanced deep learning techniques. In particular, Convolutional Neural Networks (CNNs) hold promise for modeling complex, nonlinear interactions in ultra-high-dimensional gene expression data. Although originally developed for image processing, CNNs can be adapted to analyze structured gene expression matrices, capturing spatial or correlated patterns among genes that conventional methods may miss. When integrated with interpretable models such as SCAD or Elastic Net, CNNs can contribute to hybrid pipelines that not only improve predictive accuracy but also preserve biological interpretability, which is crucial in clinical genomics. Moreover, ensemble learning approaches—which combine multiple feature selection strategies with deep learning-based representation learning—can further enhance model stability and generalizability. These hybrid frameworks offer a promising path toward capturing the full complexity of RCC-related genomic variability, improving biomarker discovery, and advancing personalized prognosis models in oncology.

Importantly, this research makes a meaningful contribution to Sustainable Development Goal (SDG) 3: Good Health and Well-being by enabling more precise and individualized risk assessment in patients with RCC. Through the identification of key genetic biomarkers associated with disease progression, clinicians are better equipped to personalize treatment plans, track patient responses with greater accuracy, and ultimately improve survival outcomes. In parallel, the study advances SDG 9: Industry, Innovation, and Infrastructure by

demonstrating the practical integration of sophisticated machine learning techniques—particularly interpretable models like SCAD—into biomedical data analysis pipelines. This represents a pivotal move toward the development of AI-driven clinical decision support systems capable of managing and extracting insights from ultra-high-dimensional genomic data. Moreover, embedding SCAD within a broader machine learning framework paves the way for scalable, reproducible analytic workflows that can be deployed across diverse healthcare environments. As these computational tools evolve and are integrated with electronic health records and large-scale genomic repositories, they hold the potential to transform research outputs into real-world clinical protocols. In doing so, this study not only addresses a complex challenge in oncology but also contributes to the evolution of a data-driven, patient-centered healthcare ecosystem aligned with global sustainability and innovation goals.

Data availability

The gene expression dataset can be obtained from the R tool "kidpack" and also available at https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-DKFZ-1.

Acknowledgment

I am thankful that University Sains Malaysia permitted me to attend this prominent university. The outstanding administrative team at the School of Mathematical Sciences is another thing for which I am immensely grateful, as they assisted me overcoming every obstacle I encountered while doing this study.

References

[1] J. Rahnenführer, R. De Bin, A. Benner, F. Ambrogi, L. Lusa, A. L. Boulesteix, E. Migliavacca, H. Binder, S. Michiels, W. Sauerbrei & L. McShane, "Statistical analysis of high-dimensional biomedical data: a gentle introduction to analytical goals, common approaches and challenges", BMC medicine **21** (2023) 182. https://doi.org/10.1186/s12916-023-02858-y.

[2] X. Han & D. Song, "Using a machine learning approach to identify key biomarkers for renal clear cell carcinoma", International Journal of General Medicine **15** (2022) 3541. https://doi.org/10.2147/IJGMS351168.

[3] P. Lambin, E. Rios-Velazquez, R, Leijenaar, S. Carvalho, R. G.van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker & H. J. Aerts, "Radiomics: extracting more information from medical images using advanced feature analysis", Eur J Cancer **48** (2012) 441. https://doi.org/10.1016/j.ejca.2011.11.036.

[4] M. Mahootiha M, H. A. Qadir, J. Bergsland & I. Balasingham, "Multimodal deep learning for personalized renal cell carcinoma prognosis: Integrating CT imaging and clinical data", Computer Methods and Programs in Biomedicine **244** (2024) 107978. https://doi.org/10.1016/j.cmpb.2023.107978.

[5] S. W. Oh, S. S. Byun, J. K. Kim, C. W. Jeong, C. Kwak, E. C. Hwang, S. H. Kang, J. Chung, Y. J. Kim, Y. S. Ha & S. H. Hong, "Machine learning models for predicting the onset of chronic kidney disease after surgery in patients with renal cell carcinoma", BMC Medical Informatics and Decision Making **24** (2024) 85. https://doi.org/10.1186/s12911-024-02473-8.

[6] N. P. Singh, R. S. Bapi & P. K. Vinod, "Machine learning models to predict the progression from early to late stages of papillary renal cell carcinoma", Computers in biology and medicine **100** (2018) 92. https://doi.org/10.1016/j.compbiomed.2018.06.030.

[7] P. Terrematte, D. S. Andrade, J. Justino, B. Stransky, D.S. de Araújo & A. D. Dória Neto, "A novel machine learning 13-gene signature: improving risk analysis and survival prediction for clear cell renal cell carcinoma patients", Cancers **14** (2022) 2111. https://doi.org/10.3390/cancers14092111.

[8] Z. Xin, R. Lv, W. Liu, S. Wang, Q. Gao, B. Zhang & G. Sun, "An ensemble learning-based feature selection algorithm for identification of biomarkers of renal cell carcinoma", PeerJ Computer Science **10** (2024) 1768. https://doi.org/10.7717/peerj-cs.1768.

[9] Y. Zhan, W. Guo, Y. Zhang, Q. Wang, X. J. Xu & L. Zhu, A five-gene signature predicts prognosis in patients with kidney renal clear cell carcinoma. Computational and Mathematical Methods in Medicine **1** (2015) 842784. https://doi.org/10.1155/2015/842784.

[10] H. Liu, Y. Luo, S. Zhao, J. Tan, M. Chen, X. Liu, X. & W. Zhong, "A reactive oxygen species–related signature to predict prognosis and aid immunotherapy in clear cell renal cell carcinoma", Frontiers in Oncology **13** (2023) 1202151. https://doi.org/10.3389/fonc.2023.1202151.

[11] T. Ebru, O. P. Fulya, A. Hakan, Y. C. Vuslat, S. Necdet, C. Nuray & O. Filiz, "Analysis of various potential prognostic markers and survival data in clear cell renal cell carcinoma", International Braz J Urol **43** (2017) 440. https://doi.org/10.1590/S1677-5538.IBJU.2015.0521.

[12] R. L. Siegel, K. D. Miller, N. S. Wagle & A. Jemal, "Cancer statistics", CA Cancer J Clin **73** (2023) 17. https://doi.org/10.3322/caac.21763.

[13] A. C. Society, "What is kidney cancer?". [Online]. https://www.cancer.org/cancer/types/kidney-cancer/about/what-is-kidney-cancer.html.

[14] World Cancer Research Fund International. "Kidney Cancer Statistics", 2023. [Online]. https://www.wcrf.org/cancertrends/kidney-cancer-statistics/.

[15] S. A. Padala, A. Barsouk, K. C. Thandra,K. Saginala, A. Mohammed, A. Vakiti & A. Barsouk, "Epidemiology of renal cell carcinoma", World journal of oncology **11** (2020) 79. https://doi.org/10.14740/wjon1279.

[16] B. Ljungberg, S. C. Campbell, H. Y. Cho, D. Jacqmin, J. E. Lee, S. Weikert & L. A. Kiemeney, "The epidemiology of renal cell carcinoma", European urology **60** (2011) 615. https://doi.org/10.1016/j.eururo.2011.06.049.

[17] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal & F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA: a cancer journal for clinicians **71** (2021) 209. https://doi.org/10.3322/caac.21660.

[18] A. Znaor, J. Lortet-Tieulent, M. Laversanne, A. Jemal & F. Bray, "International variations and trends in renal cell carcinoma incidence and mortality", European urology **67** (2015) 519. https://doi.org/10.1016/j.eururo.2014.10.002.

[19] H. Chamlal, A. Benzmane & T. Ouaderhman, "Elastic net-based high dimensional data selection for regression", Expert Systems with Applications **244** (2024) 122958. https://doi.org/10.1016/j.eswa.2023.122958.

[20] S. Bajaj, D. Gandhi, D. Nayar & A. Serhal, "Von Hippel–Lindau disease (VHL): characteristic lesions with classic imaging findings", Journal of Kidney Cancer and VHL **10** (2023) 23. https://doi.org/10.15586/jkcvhl.v10i3.293.

[21] S. Nabi, E. R. Kessler, B. Bernard, T. W. Flaig & E. T. Lam, "Renal cell carcinoma: a review of biology and pathophysiology", F1000Research **7** (2018) 29568504. https://doi.org/10.12688/f1000research.13179.1.

[22] D. Hou, W. Zhou, Q. Zhang, K. Zhang & J. Fang, "A comparative study of different variable selection methods based on numerical simulation and empirical analysis", PeerJ Computer Science **9** (2023) e1522. https://doi.org/10.7717/peerj-cs.1522.

[23] F. Li, M. Yang ,Y. Li, M. Zhang , W. Wang , D. Yuan & D. Tang, "An improved clear cell renal cell carcinoma stage prediction model based on gene sets", BMC Bioinformatics **21** (2020) 232. https://doi.org/10.1186/s12859-020-03543-0.

[24] I. Alnazer, O. Falou, P. Bourdon, T. Urruty, R. Guillevin, M. Khalil, A. Shahin, C. Fernandez-Maloigne, "Usefulness of computed tomography textural analysis in renal cell carcinoma nuclear grading", Journal of Medical Imaging **9** (2022) 054501. https://doi.org/10.1117/1.JMI.9.5.054501.

[25] R. Tibshirani, "Regression shrinkage and selection via the LASSO", Journal of the Royal Statistical Society Series B: Statistical Methodology **58** (1996) 267. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[26] H. Zhou & H. Zou, "The nonparametric box–cox model for high-dimensional regression analysis", Journal of Econometrics **239** (2024) 105419. https://doi.org/10.1016/j.jeconom.2023.01.025.

[27] H. Zou & T. Hastie, "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society Series B: Statistical Methodology **67** (2005) 301. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[28] A. Mendez-Civieta, M. C. Aguilera-Morillo & R. E. Lillo, "Adaptive sparse group LASSO in quantile regression", Advances in Data Analysis and Classification **15** (2021) 547. https://doi.org/10.1007/s11634-020-00413-8.

[29] D. F. Saldana & Y. Feng, "SIS: an R package for sure independence screening in ultrahigh-dimensional statistical models", Journal of Statistical Software **83** (2018) 1. https://doi.org/10.18637/jss.v083.i02.

[30] J. Fan & J. Lv, "Sure independence screening for ultrahigh dimensional feature space", Journal of the Royal Statistical Society Series B: Statistical Methodology **70** (2008) 903. https://doi.org/10.1111/j.1467-9868.2008.00674.x.

[31] A. Domingo-Relloso, Y. Feng, Z. Rodriguez-Hernandez, K. Haack, S. A. Cole, A. Navas-Acien & J. D. Bermudez, "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 **193** (2024) 1010. https://doi.org/10.1093/aje/kwae006.

[32] W. Wang, J. Liang, R. Liu, Y. Song & M. Zhang, "A robust variable selection method for sparse online regression via the elastic net penalty", Mathematics **10** (2022) 2985. https://doi.org/10.3390/math10162985.

[33] M. Baldomero-Naranjo, L. I. Martínez-Merino & A. M. Rodríguez-Chía, "A robust SVM-based approach with feature selection and outliers' detection for classification problems", Expert Systems with Applications **178** (2021) 115017. https://doi.org/10.1016/j.eswa.2021.115017.

[34] B. Lu, F. Wang, S. Wang, J. Chen, G. Wen & R. Fu, "Improvement of motor imagery electroencephalogram decoding by iterative weighted Sparse-Group Lasso", Expert Systems with Applications **238** (2024) 122286. https://doi.org/10.1016/j.eswa.2023.122286.

[35] A. Spooner, E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor & H. Brodaty, "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction", Scientific reports **10** (2020) 20410. https://doi.org/10.1038/s41598-020-77220-w.

[36] S . Sartori, *Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets*, Ph.D. dissertation, Department of Statistical Sciences, Università degli Studi di Milano, Milan, Italy, 2011. https://air.unimi.it/bitstream/2434/153099/6/phd_unimi_R07738.pdf.

[37] J. Fan, R. Samworth & Y. Wu, "Ultrahigh dimensional feature selection: beyond the linear model", The Journal of Machine Learning Research **10** (2009) 2013. https://doi.org/0.5555/1577069.1755853.

[38] R. Derraz, F. M. Muharam, K. Nurulhuda, N. A. Jaafar & N. K. Yap, "Ensemble and single algorithm models to handle multicollinearity of UAV vegetation indices for predicting rice biomass", Computers and Electronics in Agriculture **205** (2023) 107621. https://doi.org/10.1016/j.compag.2023.107621.

[39] A. Araveeporn, "The penalized regression and penalized logistic regression of lasso and elastic net methods for high-dimensional data: a modelling approach", Trans. Innov. Sci. Technol **3** (2022) 28. https://doi.org/10.9734/bpi/ist/v3/1695B.

[40] T. Xiong, Y. Wang & C. Zhu, "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** (2024) 4. https://doi.org/10.1186/s13000-023-01414-9.

[41] A. Ghosh, M. Jaenada & L. Pardo, "Robust adaptive variable selection in ultra-high dimensional linear regression models", Journal of Statistical Computation and Simulation **94** (2024) 571. https://doi.org/10.1080/00949655.2023.2262669.

[42] R. Madadjim, T. An & J. Cui, "MicroRNAs in Pancreatic Cancer: Advances in Biomarker Discovery and Therapeutic Implications", International Journal of Molecular Sciences **25** (2024) 3914. https://doi.org/10.3390/ijms25073914.

[43] A. Bhattacharjee, J. Dey & P. Kumari, "A combined iterative sure independence screening and Cox proportional hazard model for extracting and analyzing prognostic biomarkers of adenocarcinoma lung cancer", Healthcare Analytics **2** (2022) 100108. https://doi.org/10.1016/j.health.2022.100108.

[44] N. Salma, A. H. M. Al-Rammahi & M. K. M. Ali, "A novel feature selection method for ultra high dimensional survival data", Malaysian Journal of Fundamental and Applied Sciences **20** (2024) 1149. https://doi.org/10.11113/mjfas.v20n5.3665.

[45] Z. Arsad, Chapter 2: Multiple linear regression. Regression analysis, School of Mathematical Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia, 2023. [Online]. https://math.usm.my/images/pdf/RegressionLogistic_MacApr.pdf.

[46] M. Franz, H. Rodriguez, C. Lopes, K. Zuberi, J. Montojo, G. D. Bader & Q. Morris, "GeneMANIA update 2018", Nucleic acids research **46** (2018) W60. https://doi.org/10.1093/nar/gky311.

[47] H. Eo, "Discovery of novel genetic alteration using meta-analysis of colorectal cancer", International Journal of High School Research **6** (2024) 38. https://doi.org/10.36838/v6i1.7.

[48] A. H. AL-Rammahi & T. R. Dikheel, "Freund's model with iterated sure independence screening in Cox proportional hazard model", In AIP Conference Proceedings, Al-Samawa, Iraq, 2022, 060009. https://doi.org/10.1063/5.0093464.

[49] H. M. Nayem, S. Aziz & B. G. Kibria, "Comparison among Ordinary Least Squares, Ridge, Lasso, and Elastic Net Estimators in the Presence of Outliers: Simulation and Application", International Journal of Statistical Sciences **24** (2024) 25. https://doi.org/10.3329/ijss.v24i20.78212.

[50] A. H. Al-Rammahi & T. R. Dikheel, "Sure independent screening elastic net for ultra-high dimensional survival data", AIP Conference Proceedings, Al-Samawa, Iraq, 2021, 040001. https://doi.org/10.1063/5.0069137.

[51] K. Enwere, E. Nduka & U. Ogoke, "Comparative analysis of ridge, bridge and lasso regression models in the presence of multicollinearity", IPS Intelligentsia Multidisciplinary Journal **3** (2023) 1. https://doi.org/10.54117/iimj.v3i1.5.

[52] R. Muthukrishnan & C. K. James, "The effect of multicollinearity on feature selection", Indian Journal of Science and Technology **17** (2024) 3664. https://doi.org/10.17485/IJST/v17i35.1876.

[53] J. Pannu & N. Billor, "Robust group-Lasso for functional regression model", Communications in statistics-simulation and computation **46** (2017) 3356. https://doi.org/10.1080/03610918.2015.1096375.

[54] C. Shang, H. Ji, X. Huang, F. Yang & D. Huang, "Generalized grouped contributions for hierarchical fault diagnosis with group Lasso", Control Engineering Practice **93** (2019) 104193. https://doi.org/10.1016/j.conengprac.2019.104193.

[55] F. Khan & O. Albalawi, "Analysis of fat big data using factor models and penalization techniques: a Monte Carlo simulation and application", Axioms **13** (2024) 418. https://doi.org/10.3390/axioms13070418.

[56] J. Fan & J. Lv, "Sure independence screening", Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd., Hoboken, NJ, USA, 2018, pp. 1–8. https://doi.org/10.48550/arXiv.math/0612857.

[57] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000149294-NCAM1/2025/ (accessed 08 March 2025).

[58] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000069849-ATP1B3/2025/ (accessed 08 March 2025).

[59] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000144035-NAT8/2025/ (accessed 08 March 2025).

[60] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000125148-MT2A/2025/ (accessed 08 March 2025).

[61] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000188342-GTF2F2/2025/ (accessed 08 March 2025).

[62] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000070019-GUCY2C/2025/ (accessed 08 March 2025.

[63] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000138079-SLC3A1/2025/ (accessed 08 March 2025) .

[64] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000116791-CRYZ/2025/ (accessed 08 March 2025).

[65] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000130779-CLIP1/2025/ (accessed o8 March 2025).

[66] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000121486-TRMT1L/2025/ (accessed 08 March 2025).

[67] The Human Protein Atlas. [Online]. https://www.proteinatlas.org/ENSG00000175084-DES/ (accessed 08 March 2025).

[68] M. A. Climent, J. Muñoz-Langa, L. Basterretxea-Badiola & C. Santander-Lobera, "Systematic review and survival meta-analysis of real-world evidence on first-line pazopanib for metastatic renal cell carcinoma", Critical Reviews in Oncology/Hematology **121** (2018) 45. https://doi.org/10.1016/j.critrevonc.2017.11.009.

[69] D. Chicco, M. J. Warrens & G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation", Peerj computer science **7** (2021) e623. https://doi.org/10.7717/peerj-cs.623.

[70] A. P. Brady, "Error and discrepancy in radiology: inevitable or avoidable?", Insights Imaging **8** (2017) 171. https://doi.org/10.1007/s13244-016-0534-1.