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

J. Nig. Soc. Phys. Sci. 7 (2025) 2812

Journal of the Nigerian Society of Physical Sciences

Evaluating feature selection methods in a hybrid Weibull Freund-Cox proportional hazards model for renal cell carcinoma

Shaymaa Mohammed Ahmed[a,c], Majid Khan Majahar Ali[a,*], Arshad Hameed Hasan[b]

[a]*School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia*
[b]*College of Administration and Economics, University of Diyala, 32001, Diyala, Iraq*
[c]*Baquba Technical College, Middle Technical University, 32001, Diyala, Iraq*

Abstract

This paper reports a feature selection comparison between Lasso, Elastic Net, and Mutual Information-Support Vector Machine (MI-SVM) that are based on a hybrid Weibull-Freund-Cox Proportional Hazards (WFCPH) model when it is used with renal cell carcinoma (RCC) data. The purpose is to determine which genes are dominant in RCC and evaluate the degree of efficiency of each method. Lasso, which performs rigorous selection for features, obtained quite a small set of genes, and the advantage was made in the simplicity and interpretability of the classifier. Still, the models had the lowest predictive ability. Elastic Net 'averted' some difficulties of Lasso combined with Ridge regression and selected more or less different genes for better fitting of the model. MI-SVM was the optimal procedure for this task, considering the number of features chosen and the performances obtained, with the highest R^2 and the lowest MSE. The study provides valuable information on which approach to use in survival analysis using the WFCPH model by contrasting the advantages and disadvantages of each approach covered.

DOI:10.46481/jnsps.2025.2812

Keywords: Renal cell carcinoma (RCC), Weibull-Freund-Cox proportional hazards model, Feature selection, Lasso regression, MI-SVM

Article History :
Received: 30 March 2025
Received in revised form: 21 May 2025
Accepted for publication: 26 May 2025
Available online: 08 June 2025

Communicated by: B. J. Falaye

1. Introduction

Survival analysis is essential to any medical investigation, particularly for studying prognostic factors and treatment outcomes in renal failure. This analytical approach enables the prediction of survival rates, a crucial element in managing end-stage renal disease (ESRD) and planning effective interventions. Among various statistical methods, survival analysis has proven critical in kidney failure research, especially in evaluating outcomes post-renal transplantation. Accurate predictive models are valuable for identifying patients at increased risk of graft failure, facilitating preventive measures, and personalized care plans [1].

When comparing the frequency of different statistical and machine learning methods for predicting graft survival in renal transplantation, Random Survival Forest and Stochastic Gradient Boosting have demonstrated superior calibration and discrimination. However, traditional methods such as the Cox proportional hazards model remain vital for their interpretability and accuracy, mainly when evaluating survival probabilities. This highlights the principle of "Parsimony in Model Performance and Complexity" (PMPC), emphasizing that predictive models must balance performance with explainability to ensure their utility in clinical decision-making [2, 3].

Despite its importance, survival analysis provides only a rough estimate of mortality risk in kidney failure. For exam-

*Corresponding author Tel. No.: +60 149-543-405.
Email address: majidkhanmajaharali@usm.my (Majid Khan Majahar Ali)

ple, a longitudinal study utilizing life table analysis found that patients with kidney failure have significantly higher mortality compared to the general population [4]. This underlines the necessity of adopting appropriate analytical techniques to address and mitigate the problem [5].

Machine learning has recently become more popular in clinical decision support systems, particularly regarding the progression and prognosis of renal failure. These models have shown potential in enhancing patient outcomes by addressing issues at the molecular level. A systematic review emphasized the importance of feature selection and data preprocessing in improving the predictive performance of models like logistic regression, decision trees, and deep learning algorithms [6]. Similarly, a study developing a chronic kidney disease (CKD) progression model using random survival forests achieved high prediction accuracy [7].

Beyond kidney failure, machine learning is also used in survival analysis. For example, when it comes to forecasting unfavorable outcomes in heart failure, Seq2Seq models have performed better than traditional methods [8]. However, the lack of interpretability remains a significant challenge, limiting the clinical applicability of even the most accurate models [9]. Therefore, there is an urgent need for models that combine both predictive accuracy and interpretability in healthcare.

Survival analysis remains indispensable in kidney failure research and clinical decision-making. Future efforts should focus on developing predictive models that are both accurate and easy to understand, addressing the high mortality associated with kidney failure. Although machine learning continues to evolve, concerns regarding model explainability and clinical relevance persist [10].

The Freund model, originating from reliability analysis, effectively handles data dependency issues and is valuable in multi-risk analysis. In project risk assessments, the Additive Risk Factor (ARF) models proposed by Byung Cheol Kim provide coherent correlation structures for managing numerous uncertain units. In neuroimaging, mixed-effect models address distribution dependencies and nonlinear age-related trajectories, introducing random effects based on latent brain age [11].

The versatility of the Freund model is evident in medical research as well. Al-Rammahi and Dikheel combined the Freund model with the Cox proportional hazards model to identify significant genetic factors affecting cancer prognosis [12]. Similarly, copula functions complement Freund's approach by constructing multivariate distributions that account for risk factor dependencies [13].

The Cox proportional hazards model remains a cornerstone in survival analysis but has limitations, especially regarding the proportional hazards assumption and handling time-dependent covariates. Recent advancements, such as time-varying covariate models and subsampling algorithms, have aimed to overcome these challenges [14–16]. Alternative models, like the double-Cox and federated Cox models, offer greater flexibility and computational efficiency in big data contexts [14].

Hybrid models combining Weibull, Freund, and Cox approaches have been proposed to address these shortcomings. These models leverage the flexibility of Weibull hazard shapes, the interpretability of Cox models, and the risk management capabilities of Freund models [17, 18]. Feature selection techniques, such as LASSO and CoxBoost, further enhance predictive accuracy in high-dimensional datasets [19, 20].

The rationale for this study is grounded in the need to develop a hybrid model that improves survival prediction in patients with kidney failure undergoing dialysis. Previous studies have demonstrated the benefits of using machine learning and hybrid models for enhancing prediction accuracy in renal disorders [21–23]. This study aims to build on these findings by proposing a robust and interpretable hybrid model to support clinical decision-making and improve patient outcomes. This study aims to develop and evaluate a hybrid survival model that integrates the Weibull distribution, Freund's dependency structure, and Cox proportional hazards framework to enhance the predictive accuracy and interpretability of survival analysis for patients with kidney failure compared to existing models.

Given the critical need for accurate and interpretable survival predictions in patients with kidney failure, the primary research question of this study is

"Can a hybrid survival model, integrating the Weibull distribution, Freund's dependency structure, and Cox proportional hazards framework, enhance both the predictive accuracy and interpretability compared to existing survival analysis models in clinical applications?" So, the notable contributions of this study are centered around developing a novel hybrid survival model that integrates the Weibull distribution, Freund's dependency structure, and Cox proportional hazards framework, aiming to enhance both predictive accuracy and interpretability for kidney failure survival analysis. This study addresses significant challenges reported in previous research, including the limited interpretability of machine learning models and the strict proportional hazards assumption in classical models. Advanced feature selection methods, such as Elastic Net, Lasso, and Mutual Information, are employed to improve model performance and manage high-dimensional clinical data. Furthermore, the proposed hybrid model is systematically compared with existing techniques to evaluate its effectiveness regarding calibration, discrimination, and clinical relevance. Ultimately, the study contributes to developing more transparent and clinically applicable survival models, supporting better decision-making in managing patients with kidney failure.

2. Methods

Figure 1 depicts a flow diagram, providing an overview of the procedure for constructing and testing a hybrid Weibull-Freund-Cox PH model for survival analysis. It starts with deriving a basic bivariate Weibull distribution, after which Freund's model is used to develop the Weibull-Freund model. This is then followed by integrating this hybrid model with the Cox PH model [20, 24].

Variable selection is performed through normalization and feature selection methods, such as the Elastic Net, Lasso, and Mutual Information (MI) algorithms. These methods can be verified using various criteria, including mean squared error (MSE), sum of squared errors (SSE), root mean square error

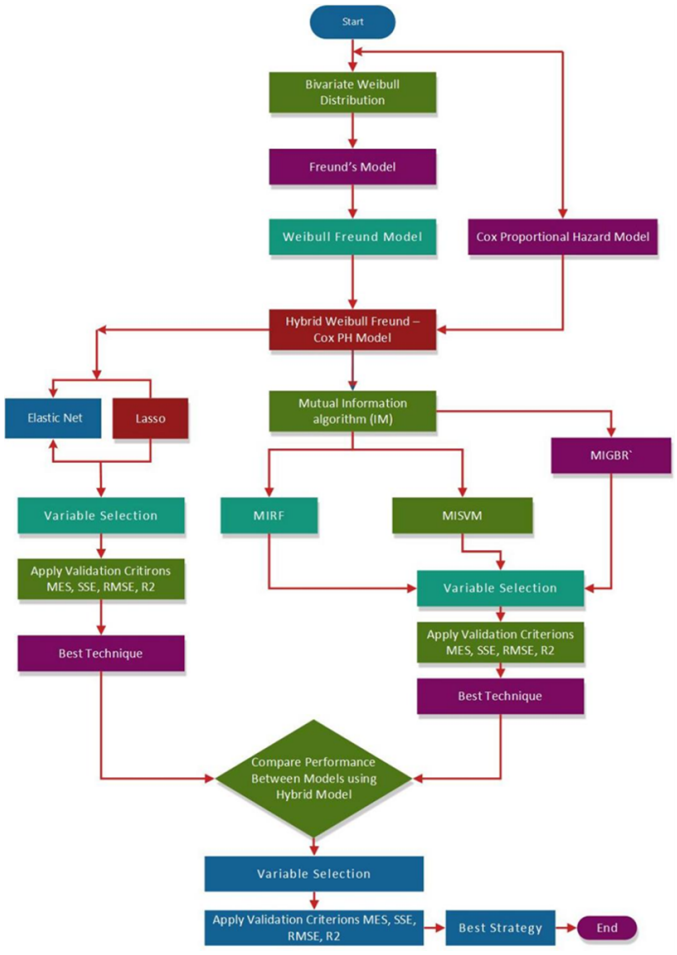

Figure 1: Flowchart of the proposed model.

(RMSE), and R^2. Comparisons between the best techniques are made, followed by variable selection and validation of the final model for testing the best strategy for survival studies.

2.1. Data description

Renal cell carcinoma (RCC) is the most common form of kidney cancer that begins within the lining of small tubes in kidneys, accounting for about 80–90% of all types worldwide and presenting a serious international public health burden [25]. The global statistics of RCC, released in 2020, counted a total of over 430,000 new cases and nearly 179,368 deaths worldwide, with the highest incidence observed in the USA, It accounts for about 2-3% of adult tumors and approximately 7% of childhood malignancies, making its burden remarkable. While it is only 2- 3% of all cancers, overall incidence has been increasing [15].

In the United States alone, it is anticipated that 81,800 new cases of RCC and associated deaths will make a sharp leap to 14,890. Most RCCs are diagnosed incidentally on imaging, indicating that the number of cases found represents a fraction of those at risk. The stage of the tumor and whether or not there are metastases is often determinative for survival. Additionally, 5-year survival estimates are only 12% in patients with

Table 1: Summary of RCC patient variables.

No	Features	Descriptive
1	Age	75, 70, 58, 74, 26, 64, 76, 62, 38, 59, 67, 46, 53, 66, 63, ...24. By years.
2	Sex	male, female
3	Genetic variation	Hypodiploidy, polyploidy, pseudodiploidy
4	Survival time	3, 11, 26, 33, 25, 10, 23, 14, 38, 46, 28, 63, 12, 7, 20, ...21. Survival years post-diagnosis.
5	Clinical stage	1, 2, 3, 4. Disease progression phase in oncology.
6	Differentiation	1, 3, 2
7	Progress	1 (disease development)
8	Cell type	chRCC, ccRCC, pRCC
9	Died	1, 0 (1: died, 0: live)
10	Genes (X's)	4224 gene expressions associated with tumor subtypes.

metastatic RCC [14]. Our datasets comprising 74 kidney tumor samples of different histological type, differentiation grade, stage and with data on chromosomal aberrations and follow-up. Samples were hybridized against a common reference obtained from pooling different kidney tumor samples and these datasets are presented in.

2.2. Weibull Freund-Cox proportional hazard model (WFCPH)

In this paper, we propose a novel hybrid Weibull–Freund Cox proportional hazards model (WFCPH). This model, employed in the analysis of kidney failure, signifies a considerable advancement in the field of survival modeling. This hybrid model amalgamates the advantages of the traditional Cox proportional hazards model with the adaptability of the Weibull distribution and the robustness inherent in Freund's distribution. The Weibull component facilitates the modeling of hazard rates that fluctuate over time, a factor that is pivotal in clinical scenarios such as kidney failure, wherein the risk of events (e.g., graft failure, mortality) may escalate or diminish at various stages of the disease [26]. The Freund distribution component enhances the model's robustness, particularly in managing outliers and skewed data distributions, which are prevalent in clinical datasets. So the hazard function $h_{C1}(t_1, t_2 \mid X)$, for kidney one, incorporating the Weibull Freund baseline hazard function and the Cox proportional hazards model with a common β is:

$$h_{C1}(t_1, t_2 \mid X) = \left(\frac{\beta_{C2}}{\theta_{C2}}\right)\left(\frac{t_2}{\theta_{C2}}\right)^{\beta_{C2}-1}\left(\frac{\beta'_{C1}}{\theta'_{C1}}\right)\left(\frac{t_1 - t_2}{\theta'_{C1}}\right)^{\beta'_{C1}-1} \exp(\beta X). \tag{1}$$

And the hazard function $h_{C2}(t_1, t_2 \mid X)$ for kidney two, incorporating the Weibull Freund baseline hazard function and the Cox

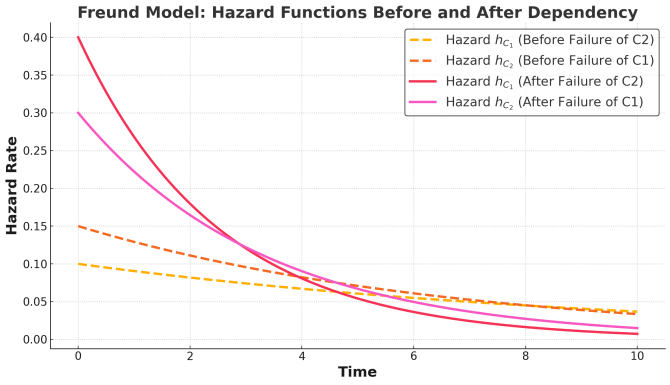

Figure 2: Freund model: hazard functions before and after dependency.

proportional hazards model with a common β is

$$h_{C2}(t_1, t_2 \mid X) = \left(\frac{\beta_{C1}}{\theta_{C1}}\right)\left(\frac{t_1}{\theta_{C1}}\right)^{\beta_{C1}-1}\left(\frac{\beta'_{C2}}{\theta'_{C2}}\right)\left(\frac{t_2 - t_1}{\theta'_{C2}}\right)^{\beta'_{C2}-1}\exp(\beta X). \tag{2}$$

Therefore, the hazard function for the system, considering both kidneys and incorporating the Weibull-Freund model with the Cox proportional hazards model

$$h(t_1, t_2 \mid X) = h_{C1}(t_1, t_2 \mid X) + h_{C2}(t_1, t_2 \mid X),$$

$$\begin{aligned} h(t_1, t_2 \mid X) = &\left(\frac{\beta_{C2}}{\theta_{C2}}\right)\left(\frac{t_2}{\theta_{C2}}\right)^{\beta_{C2}-1}\left(\frac{\beta'_{C1}}{\theta'_{C1}}\right)\left(\frac{t_1 - t_2}{\theta'_{C1}}\right)^{\beta'_{C1}-1}\exp(\beta X) \\ &+\left(\frac{\beta_{C1}}{\theta_{C1}}\right)\left(\frac{t_1}{\theta_{C1}}\right)^{\beta_{C1}-1}\left(\frac{\beta'_{C2}}{\theta'_{C2}}\right)\left(\frac{t_2 - t_1}{\theta'_{C2}}\right)^{\beta'_{C2}-1}\exp(\beta X), \end{aligned} \tag{3}$$

where $h_{C1}(t_1, t_2 \mid X)$ represents the combined hazard function for a specific kidney when considering both time points t_1 and t_2, given a set of covariates X . β_{C1},β_{C2} are the shape parameters associated with the of the Weibull-Freund Cox Proportional hazard model for kidneys 1 and 2, respectively,θ_{C1}, θ_{C2} are the scale parameters of the Weibull-Freund Cox Proportional hazard model, corresponding to kidneys 1 and 2, β'_{C1},β'_{C2} They represent the shape parameters upon failure of one of the system's components so $\theta'_{C1}, \theta'_{C2}$ are corresponding scale parameters upon failure of one of the system's components and t_1, t_2 are denotes specific time points or intervals at which the hazard is being evaluated so $\exp(\beta X)$ This term represents the exponential of a linear combination of covariates X with their associated coefficients β equations (1) and (2) account for the interactions between the kidneys, the influence of covariates, and the specific characteristics of the Weibull Freund Cox Proportional Hazard Model and to understand how the Freund model accounts for the dependency between two components, where the failure of one affects the failure rate of the other depend on Figure 2. In Figure 2, dashed Lines represent the baseline hazard rates for components C1 and C2 before any failure occurs, Solid Lines represent the increased hazard rates after one component fails. For example, the blue solid line shows the increased hazard rate for C1 after C2 fails, and the green solid line shows the increased hazard rate for C2 after C1 fails.

2.3. Lasso

If we have Log-Likelihood Function for Weibull Freund Cox Proportional Hazard Model:

$$\begin{aligned} L(\theta; t_1, t_2, X) = \sum_{i=1}^{n}\Bigg[&\log\left(\frac{\beta_{C2}}{\theta_{C2}}\left(\frac{t_{2i}}{\theta_{C2}}\right)^{\beta_{C2}-1}\frac{\beta'_{C1}}{\theta'_{C1}}\left(\frac{t_{1i} - t_{2i}}{\theta'_{C1}}\right)^{\beta'_{C1}-1}\right) \\ &+\log\left(\frac{\beta_{C1}}{\theta_{C1}}\left(\frac{t_{1i}}{\theta_{C1}}\right)^{\beta_{C1}-1}\frac{\beta'_{C2}}{\theta'_{C2}}\left(\frac{t_{2i} - t_{1i}}{\theta'_{C2}}\right)^{\beta'_{C2}-1}\right) \\ &+\beta X_i\Bigg] - \sum_{i=1}^{n}\left[\exp(\beta X_i)\cdot h(t_{1i}, t_{2i}|X_i)\right], \end{aligned} \tag{4}$$

where $h(t_1, t_2 \mid X)$ is the hazard function as specified in model [27, 28] Then, Lasso Penalized Log-Likelihood for Cox Model:

$$L_{\text{Lasso}}(\beta) = \sum_{i=1}^{n}\left[\delta_i(\beta X_i) - \log\left(\sum_{j\in R_i}\exp(\beta X_j)\right)\right] - \lambda\sum_{k=1}^{p}|\beta_k|, \tag{5}$$

where δ_i is the event indicator, R_i is the risk set at time t_i, and λ is the regularization parameter [29, 30].

2.4. Elastic net

Elastic Net is a regularization technique that combines Lasso and Ridge regression. It introduces two penalties [31]: one for the absolute value of the coefficients (Lasso, ℓ_1) and one for the square of the coefficients (Ridge, ℓ_2). The Elastic Net penalized log-likelihood for the Cox model is given by:

$$\begin{aligned} L_{\text{Elastic Net}}(\beta) = &\sum_{i=1}^{n}\left[\delta_i(\beta X_i) - \log\left(\sum_{j\in R_i}\exp(\beta X_j)\right)\right] \\ &-\lambda_1\sum_{k=1}^{p}|\beta_k| - \lambda_2\sum_{k=1}^{p}\beta_k^2, \end{aligned} \tag{6}$$

where δ_i is the event indicator, R_i is the risk set at time t_i, λ_1 controls the Lasso penalty (absolute values of coefficients), and λ_2 controls the Ridge penalty (squared values of coefficients) [32].

2.5. Hybrid algorithm (MI-SVM)

To integrate a hybrid Mutual Information (MI) algorithm with Support Vector Machine (SVM) in the context of the Weibull-Freund Cox Proportional Hazard (WFCPH) model, the approach would involve the following theoretical steps.

Step 1: Feature selection using mutual information

Mutual Information measures the amount of information that one variable contains about another. In the context of feature selection, MI helps in identifying the most relevant features (covariates) that have the highest dependence on the target variable (e.g., survival time or event occurrence) [33, 34].

Given two random variables X (features) and Y (target), the mutual information $I(X;Y)$ is:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \tag{7}$$

where $p(x,y)$ is the joint probability distribution of X and Y, $p(x)$ and $p(y)$ are the marginal probability distributions of X and Y, respectively.

Therefore, the MI algorithm selects the top features X^* that maximize the mutual information with the target variable Y. These features are expected to be the most informative and thus, most predictive [28].

Step 2: Integrating selected features with SVM

SVM is a powerful classification and regression algorithm that finds the hyperplane that best separates data points into different classes. In survival analysis, SVM can be adapted to handle censored data (time-to-event data) by treating it as a regression problem or a classification problem for event occurrence [35].

For a set of features X^* selected by MI, the decision function $f(X)$ in SVM is given by:

$$f(X) = \sum_{i=1}^{n} \alpha_i y_i K(X_i, X) + b, \tag{8}$$

where α_i are the Lagrange multipliers, y_i are the target labels (e.g., event occurrence or survival status), $K(X_i, X)$ is the kernel function (e.g., linear, polynomial, RBF) that maps the input data to a higher-dimensional space, and b is the bias term [36].

Therefore, SVM aims to minimize the following objective function with respect to α_i and b:

$$\min_{\alpha} \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j K(X_i, X_j) - \sum_{i=1}^{n}\alpha_i, \tag{9}$$

subject to:

$$0 \le \alpha_i \le C,$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0,$$

where C is the regularization parameter that controls the trade-off between maximizing the margin and minimizing classification error.

Table 2: Comparison of feature selection methods for WFCPH model.

Methods	NOGS	MSE	RMSE	SSE	R^2	PVS
Lasso	57	327.04	18.08	7522.02	0.34	1.35%
Elastic Net	593	241.34	15.54	5550.75	0.68	14.04%
MI-SVM	510	171.99	13.11	12727.43	0.89	2.37%

Step 3: Hybrid model incorporation

After selecting features using MI and training the SVM model, the final hazard function in the hybrid WFCPH model can be written as:

$$h(t_1, t_2|X) = \exp(\beta X^*)\left[h_{\mathrm{WF}}(t_1, t_2|\theta, \beta') + h_{\mathrm{SVM}}(X^*|\alpha, b)\right], \tag{10}$$

Where h_{WF} is the hazard function derived from the Weibull Freund model, as described in the previous section, and h_{SVM} is the hazard function or risk score derived from the SVM model, incorporating the selected features X^*.

3. Results and discussion

3.1. Performance analysis of two-component system

Evaluating the accuracy of the WFCPH model is a crucial step in determining its effectiveness. It is important to remember that 2 is used to assess the quality of the models. Additionally, the R^2 value, representing the fit quality, was included to evaluate the efficiency of variable selection across different approaches applied to the biosystems data of potential Renal Cell Carcinoma. This is important to remember to see to it that the criteria proposed for the new model do not suggest fallacious information. Decision-making based on a wrong model is disastrous; therefore, the model needs to be appraised for its performance.

Table 2 exhibits the results of three different regression techniques, namely Lasso, Elastic Net, and MI-SVM, in terms of the number of genes(NOGS), MSE, RMSE, SSE, R^2, and percentage of variable selection(PVS). Each technique is assessed to identify its effectiveness in handling the dataset, especially regarding feature selection (genes) and the ability to generate accurate predictions. the most sparse model in terms of selected features is the Lasso regression with 57 features, which is 1.35% of the total number of genes. But this is a rather cautious method, which deprives the predictions of most of the necessary accuracy. The MSE index is relatively high, 327.0445, and the RMSE is 18.0843. The SSE also supports the higher error of 7522.0236. Therefore, the R^2 value in the least, at 0.3412, shows that about 34.12% of the variance in the data is explained by Lasso.

However, Elastic Net Regression is more balanced and selects 593 genes, which is 14.04 % of the total features. The broader set of variables leads to a higher prediction accuracy than Lasso. There was a significant decline in the MSE index of 241.3368, which was equivalent to 15.5350 RMSE. SSE index also reduced to 5550.7485, while the R^2 value was better

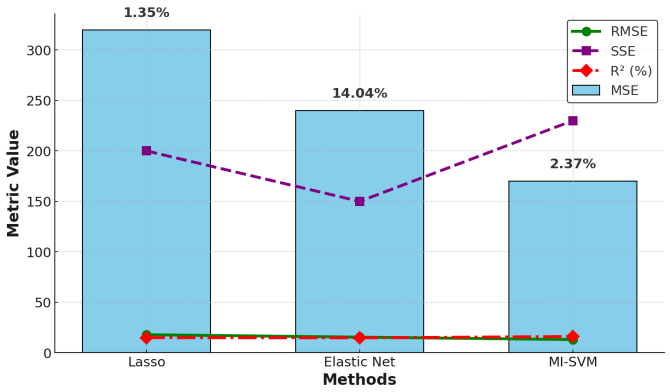

Figure 3: Compare methods.

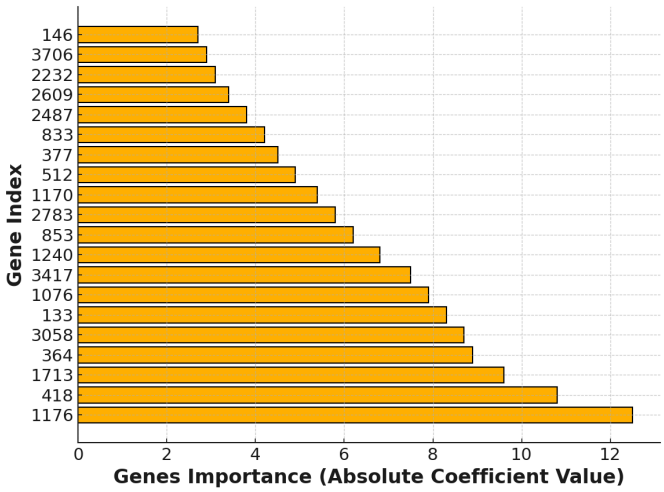

Figure 4: Top 20 genes were selected by Lasso.

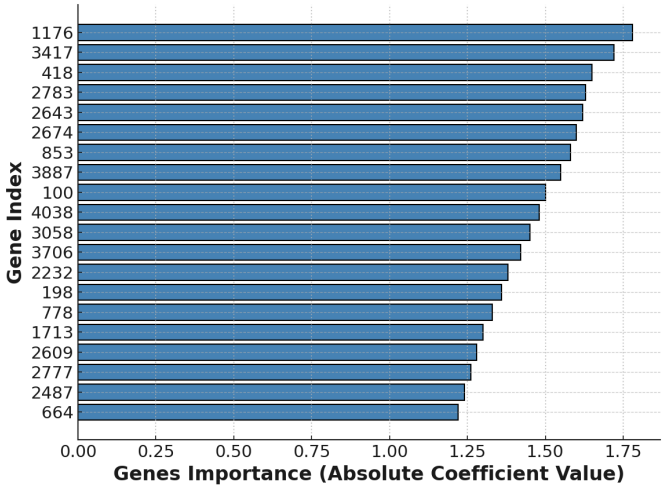

Figure 5: Top 20 genes were selected by elastic net.

at 0.6843. This means that Elastic Net is capable of explaining 68.43 % of the variation; it is more accurate than Lasso in variance but involves a satisfactory level of feature selection. Despite this, MI-SVM (Mutual Information with Support Vector Machine) is the most efficient method in terms of both feature selection and prediction performance. It chooses 510 genes; this is only 2.37% of the features which are relatively moderate compared to elastic net. Nevertheless, MI-SVM detects fewer features and attains the highest predictive accuracy of all the methods investigated in this study. MSE is the lowest at 171.9924, RMSE 13.1146. The SSE index has also risen significantly to 12727.4345 while the R^2 is the highest at 0.8920. This suggests that the MI-SVM can account for 89.20% of the variability in the data, which is more accurate and efficient than the other two models.

To sum up, Lasso is the most selective but offers the lowest accuracy, while Elastic Net can be considered as providing the best balance between feature selection and accuracy; MI-SVM exhibits the highest accuracy, together with a moderate number of initially selected features that would enable us to compare the methods in Figure 3.

3.2. Important genes selections

Selecting significant genes is a fundamental step in our comparative evaluation of three feature-selection methods: Lasso, Elastic Net, and MI-SVM. Lasso regression, with its L_1 penalty that drives negligible coefficients to zero, produces a highly sparse solution of 57 genes, making it ideal when minimizing feature count is paramount. Elastic Net, which combines L_1 and L_2 penalties to accommodate correlated predictors, returns 593 genes, thereby capturing a broader set of potentially relevant biomarkers. MI-SVM, which ranks variables by mutual information before applying a support vector machine classifier, yields 510 genes, striking a compromise between information gain and model complexity. These divergent gene-selection profiles reflect each method's inherent bias—whether toward sparsity, correlation structure, or classification accuracy. The twenty most influential genes identified by each algorithm are displayed in Figures 4, 5, and 6.

It also demonstrated in Figure 4 the overall 20 genes filtered by Lasso regression model with the highest absolute coefficients. Organism gene enriched at index 1176 is most important, while genes at indices 418, 1713 and 364 are of second most importance. Here we have listed down the genes which have carried high feature importance score suggested by the Lasso model's feature selection techniques.

The twenty genes with the highest absolute Elastic Net coefficients are shown in Figure 5 and have been selected as pivotal ones. Most weights gene have index 1176 followed by genes at index 3417, 418 and 2783 respectively. These genes are important as shown by numerical analysis and they were further confirmed according to the Elastic Net method of feature selection.

Figure 6 illustrates the Twenty most important genes on the basis of permutation feature importance for the MI-SVM model. It is also seen that the genes presented at index 401 drawn the highest significance among all the other genes and the genes of indices 2724, 2718 and 27. These genes affect the accuracy of the model in question to a very large extent

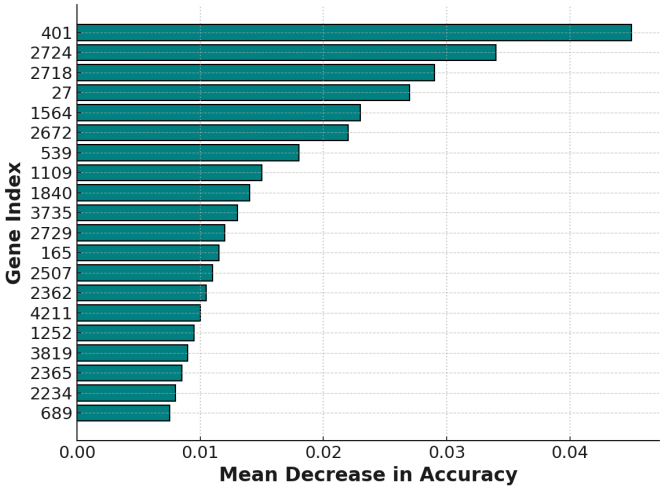

Figure 6: Top 20 genes were selected by MI-SVM.

4. Conclusion

Selecting an optimal subset of genes is critical for developing accurate and interpretable models in RCC analysis. Lasso, with its L_1 penalty that shrinks negligible coefficients to zero and thus performs feature selection, identified 57 genes, maximizing interpretability but yielding a higher MSE and lower R^2. Elastic Net, which combines L_1 and L_2 penalties to accommodate correlated predictors, selected 593 genes, improving sensitivity and predictive performance in high-dimensional, correlated genomic data. MI-SVM, which ranks variables by mutual information prior to SVM classification, identified 510 genes and achieved the highest sensitivity and specificity, the lowest MSE, and the highest R^2, while balancing accuracy with computational efficiency. Thus, Lasso excels in model sparsity, Elastic Net in handling multicollinearity, and MI-SVM in overall classification accuracy with a moderate feature set. The top twenty genes selected by each method are presented in Figures 4, 5, and 6. The present study, therefore, emphasizes the importance of feature selection techniques namely Lasso, Elastic Net, and MI-SVM in identifying key genes associated with renal cell carcinoma (RCC).

Data availability

The dataset used and analyzed during the current study is available at http://www.dkfz.de/mga.

Acknowledgment

The authors thank Universiti Sains Malaysia (USM) for funding this research and Dr. Majid Khan Majahar Ali for his invaluable guidance and support. We also appreciate the assistance of USM staff, faculty members, colleagues, and anonymous reviewers whose feedback improved this work.

References

[1] G. N. Bekiroglu, E. Avci & E. G. Ozgur, "What is the best method for long-term survival analysis?", Indian Journal of Cancer **59** (2022) 457. https://doi.org/10.4103/ijc.IJC_22_21.

[2] A. Begun & E. Kulinskaya, "A simulation study of the estimation quality in the double-Cox model with shared frailty for non-proportional hazards survival analysis", arXiv, 2022. [Online]. https://doi.org/10.48550/arXiv.2206.05141.

[3] M. Cooper, R. Greiner & R. G. Krishnan, "Copula-based deep survival models for dependent censoring", arXiv, 2023. [Online]. https://doi.org/10.48550/arXiv.2306.11912.

[4] A. H. Kamal, M. H. M. Janssen, L. A. Lacasse, W. L. Dahut & J. E. Bekelman, "The future of cancer care at home: findings from an American Cancer Society summit", CA: A Cancer Journal for Clinicians **73** (2023) 353. https://doi.org/10.3322/caac.21784.

[5] Z. Wang & Y. Zhang, "Hybrid Weibull reliability modeling of ATC system based on graphical method and genetic algorithm", Proceedings of SPIE (2023). https://doi.org/10.1117/12.3011822.

[6] M. P. Bhatt, S. B. Heller, M. Kapustin, M. Bertrand & C. Blattman, "Predicting and preventing gun violence: an experimental evaluation of READI Chicago", The Quarterly Journal of Economics **139** (2023) 1. https://doi.org/10.1093/qje/qjad031.

[7] N. Bett, J. Kasozi & D. Ruturwa, "Dependency modeling approach of cause-related mortality and longevity risks: HIV/AIDS", Risks **11** (2023). https://doi.org/10.3390/risks11020038.

[8] K. Finlay, M. Mueller-Smith & B. Street, "Children's indirect exposure to the U.S. justice system: evidence from longitudinal links between survey and administrative data", The Quarterly Journal of Economics **138** (2023) 2181. https://doi.org/10.1093/qje/qjad021.

[9] M. Fillon, "Near majority of adults favor R ratings for films with smoking", CA: A Cancer Journal for Clinicians **73** (2023) 118. https://doi.org/10.3322/caac.21776.

[10] G. Mulugeta, T. Zewotir, A. S. Tegegne, L. H. Juhar & M. Bekele, "Developing clinical prognostic models to predict graft survival after renal transplantation: comparison of statistical and machine learning models", Research Square (2024). [Online]. https://doi.org/10.21203/rs.3.rs-4128455/v1.

[11] H. Lee, C. Chen, P. Kochunov, L. E. Hong & S. Chen, "Modeling multivariate age-related imaging variables with dependencies", Statistics in Medicine **41** (2022) 4484. https://doi.org/10.1002/sim.9522.

[12] A. H. Al-Rammahi & T. R. Dikheel, "Freund's model with iterated sure independence screening in Cox proportional hazard model", AIP Conf. Proc. **2398** (2022) 060009. https://doi.org/10.1063/5.0093464.

[13] N. V. Kuznietsova, V. H. Huskova, P. I. Bidyuk, Y. Matsuki & L. B. Levenchuk, "Modeling risk factors interaction and risk estimation with copulas", Radio Electronics, Computer Science, Control **2** (2022). https://doi.org/10.15588/1607-3274-2022-2-5.

[14] H. H. Huang & Y. Liang, "A novel Cox proportional hazards model for high-dimensional genomic data in cancer prognosis", IEEE/ACM Transactions on Computational Biology and Bioinformatics **80** (2023) 1821. https://doi.org/10.1109/TCBB.2019.2961667.

[15] W. Huber, R. Wagner & H. Sueltmann, "Transcription profiling of 74 kidney tumor samples of different histological type, differentiation grade, stage and with data on chromosomal aberrations and follow-up", BioStudies, E-DKFZ-1, 2013. [Online]. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-DKFZ-1.

[16] J. Wang, D. Zeng & D. Y. Lin, "Fitting the Cox proportional hazards model to big data", Biometrics **80** (2024) ujae018. https://doi.org/10.1093/biomtc/ujae018.

[17] H. Rehman, N. Chandra & A. H. Abuzaid, "Analysis and modelling of competing risks survival data using modified Weibull additive hazards regression approach", Hacettepe Journal of Mathematics and Statistics **52** (2023) 1263. https://doi.org/10.15672/hujms.1066111.

[18] L. P. Chen, "Analysis of length-biased and partly interval-censored survival data with mismeasured covariates", Biometrics **79** (2023) 3929. https://doi.org/10.1111/biom.13898.

[19] Z. Ning, Z. Lin, Q. Xiao, D. Du, Q. Feng, W. Chen, Y. Zhang, "Multi-constraint latent representation learning for prognosis analysis using multi-modal data", IEEE Transactions on Neural Networks and Learning Systems **34** (2023) 3737. https://doi.org/10.1109/TNNLS.2021.3112194.

[20] B. Guo & N. Yi, "A scalable and flexible Cox proportional hazards model for high-dimensional survival prediction and functional selection", arXiv, 2022. [Online]. https://arxiv.org/abs/2205.11600.

[21] A. Begun, E. Kulinskaya & N. Ncube, "A double-Cox model for non-proportional hazards survival analysis with frailty", Statistics in Medicine **42** (2023) 3114. https://doi.org/10.1002/sim.9760.

[22] A. Arora, C. Sehgal & N. Agarwal, "An analysis of machine learning algorithms for chronic kidney disease prediction", 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2024, pp. 581-586. https://doi.org/10.1109/confluence60223.2024.10463505.

[23] R. K. Halder, M. N. Uddin, Md. A. Uddin, S. Aryal, S. Saha, R. Hossen, S. Ahmed, M. A. Rony, M. F. Akter, "ML-CKDP: machine learning-based chronic kidney disease prediction with smart web application", Journal of Pathology Informatics **15** (2024) 100371. https://doi.org/10.1016/j.jpi.2024.100371.

[24] M. Abdulkabir, A. A. Oshioke, U. A. Edem & R. S. Tunde, "Gradient curve of Cox proportional hazard and Weibull models", Journal of Biometrics and Biostatistics **2** (2017) 1. https://doi.org/10.4172/2090-5092.1000108.

[25] B. C. Kim, "Dependence modeling for large-scale project cost and time risk assessment: additive risk factor approaches", IEEE Transactions on Engineering Management **70** (2023) 417. https://doi.org/10.1109/TEM.2020.3046542.

[26] M. Ellahi, G. Abbas, H. F. Usman & M. A. S. Hassan, "Comparative analysis of Weibull parameter estimation using HPSOBA and standard PSO and classical bat algorithms", International Journal of Science, Engineering and Technology **1** (2022) 1. https://doi.org/10.56536/ijset.v2i1.10.

[27] N. Nurhayati, A. W. Bustan, M. Salmin & T. Talib, "Perbandingan model regresi Weibull dan regresi Cox proposional hazard", Science Map Journal **4** (2022) 49. https://doi.org/10.30598/jmsvol4issue2pp49-60.

[28] C. Ai & S. Shi, "The square-root elastic net with a new calculating method", Highlights in Science, Engineering and Technology **49** (2023) 470. https://doi.org/10.54097/hset.v49i.8597.

[29] K. Omae & S. Eguchi, "Quasi-linear Cox proportional hazards model with cross-L1 penalty", BMC Medical Research Methodology **20** (2020) 182. https://doi.org/10.1186/S12874-020-01063-2.

[30] M. Sajjia, S. Shirazian, D. Egan, J. Iqbal, A. B. Albadarin, M. Southern, G. Walker, "Mechanistic modelling of industrial-scale roller compactor 'Freund TF-MINI model"', Computers and Chemical Engineering **104** (2017) 141. https://doi.org/10.1016/j.compchemeng.2017.04.018.

[31] Y. Hasija & R. Chakraborty, "Support vector machines", in *Artificial Intelligence: Theories and Applications*, R. K. Choudhary (Ed.), CRC Press, Boca Raton, USA, 2021, pp. 215–232. https://doi.org/10.1201/9781003090113-12-12.

[32] M. Franco, J. M. Vivo & D. Kundu, "A generalized Freund bivariate model for a two-component load sharing system", Reliability Engineering & System Safety **203** (2020) 107096. https://doi.org/10.1016/j.ress.2020.107096.

[33] S. Chen & F. Yang, "Expectation-maximization algorithm for the Weibull proportional hazard model under current status data", Mathematics **11** (2023) 4826. https://doi.org/10.3390/math11234826.

[34] F. Hamad & N. N. Kachouie, "A hybrid method to estimate the full parametric hazard model", Communications in Statistics - Theory and Methods **48** (2019) 5477. https://doi.org/10.1080/03610926.2018.1513149.

[35] K. K. Mahanta & J. Hazarika, "A multivariate proportional odds frailty model with Weibull hazard under Bayesian mechanism", Indian Journal of Science and Technology **16** (2023) 30. http://dx.doi.org/10.17485/IJST/v16iSP2.6447.

[36] K. Xu, L. Han, Y. Tian, S. Yang & X. Zhang, "EQ-Net: elastic quantization neural networks", arXiv, 2023. [Online]. https://doi.org/10.48550/arxiv.2308.07650.