Faculty of Commerce
Tanta University

The Scientific Journal of Business and Finance

Multidisciplinary Journal
Accounting, Management, Economics, Statistics, Mathematics and Insurance

Volume 45, Issue 2, June 2025

Print Issn: 1110-4716
Online Issn: 2682-4825

The Scientific Journal of Business and Finance

https://caf.journals.ekb.eg

Enhancing Lung Cancer Survival Prediction: A Comparative Study of Hybrid Cox-SVM and Logistic-SVM Models

Abdelreheem Awad Bassuny

Lecturer at the Higher Institute of A Management in Mahalla El-Kubra

Published online: June 2025.

To cite this article: Awad Bassuny,Abdelreheem. Enhancing Lung Cancer Survival Prediction: A Comparative Study of Hybrid Cox-SVM and Logistic-SVM Models, **The Scientific Journal of Business and Finance** 45, (2),341-376.

DOI: 10.21608/caf.2025.435002

***Corresponding author: dr-AbdelreheemBassuny@outlook.com**

Enhancing Lung Cancer Survival Prediction: A Comparative Study of Hybrid Cox-SVM and Logistic-SVM Models

Abdelreheem Awad Bassuny

the Higher Institute of A Management in Mahalla El-Kubra

Article History

Received 4April, 2025 ,, Accepted 15 , May, 2025, Available online June 2025.

Abstract

The research compares the survival performance of hybrid survival models that merge traditional statistical techniques with machine learning to forecast survival outcomes in lung cancer patients. The study compares a group of 165 patients from Tanta Oncology Institute and Kafr El-Sheikh Chest Hospital between 2020 and 2024 by contrasting individual models—Cox Regression, Logistic Regression, and Support Vector Machines (SVM)—with two hybrid models: Cox-SVM and Logistic-SVM. High-risk predictors such as smoking, occupation, age, treatment modalities, COVID-19 infection, and disease stage were found and modeled through Kaplan-Meier analysis for feature selection. The results show that the Cox-SVM hybrid model gives the best results among all the other models with a classification accuracy of 94.5%, sensitivity of 90%, specificity of 96%, and misclassification rate of 5.45%. The Logistic-SVM hybrid then comes into play with a 90.91% accuracy, and then the individual models (Cox: 80.6%, Logistic: 67.88%, SVM: 81.8%). Hybrid model performance is because Cox's hazard-based model combines SVM's capability to work with non-linear relationships and produce better predictive accuracy and clinical relevance. Despite small sample size limitations and omission of certain variables, these results suggest that hybrid approaches in survival analysis are a valuable resource in personalized medicine. These models must be validated with larger, more diverse datasets and other prognostic factors in future research.

Keywords: Survival Analysis؛ Hybrid Models؛ Kaplan-Meier؛ Machine Learning؛ Cox Regression.

1. Introduction

Survival analysis is a statistical specialty that deals with the time elapsed until a specific event occurs. It is becoming increasingly relevant across various fields such as medicine, engineering, and business. As data becomes increasingly complex, the discipline has experienced continuous growth in analytical techniques, and mixed models have proved to be powerful analytical tools. Such models integrate the characteristics of other models to become more comprehensive and accurate in data analysis, with characteristics of interdependencies and determinants of various types.

Hybrid models extend the limitations of conventional models, such as the proportionality assumption of hazards and distributional stability, by combining the strengths of multiple models into a single analytical model. Such flexibility allows for complicated interactions of multiple variables and their influences on survival to be modeled while controlling different levels of variance and heterogeneity in the data. In most disciplines, these models yield a better understanding of determinants of survival, enhance predictability, and enable better-informed decision-making. These models need to be used correctly through clear comprehension of their parameters and features to enable proper interpretation of the findings. Mixed models are a key and new development in the survival analysis armamentarium, an advance in the field. Numerous researchers are studying survival analysis as can be seen from recent studies in the field. Yakubu et al.'s (2022) study introduced a novel way to study heterogeneous survival data through a two-part mixed model: gamma-gamma and log-logistic-gamma. The authors conducted simulations to assess the performance of the model and utilized the expectation-maximization (EM) algorithm to estimate parameters. They checked consistency and stability with mean squared error (MSE) and root mean square error (RMSE) and compared the mixed model with one classical distribution under real data. They concluded that a log-logistic-gamma mixed model gave a better fit. As a result, these results indicate that mixed models are more appropriate for complex survival data with heterogeneity.

In addition, Cuthbert *et al*., (2022) likened a variety of survival models for predicting revision risk 8 years after total hip and knee arthroplasty. The research, based on over 400,000 procedures, established that Cox and flexible parametric models performed best in most situations, despite all approaches having similar discrimination. They also found poorer calibrated random survival forests for patients with total knee arthroplasty and discussed the incremental value of such advanced analyses over regression models, at least for predictive performance and variable importance scores.

Xiao *et al*., (2022) compared machine learning and Cox regression for breast cancer prognosis using a retrospective cohort of over 22,000 patients. The random survival forest (RSF) model slightly outperformed other models concerning discriminative ability, as demonstrated by the highest C-index. All models showed good calibration. The study identified important prognostic factors including TNM staging, neoadjuvant therapy, lymph node metastases, age, and tumor diameter.

In a different context, Al-Essa *et al*., (2023) investigated Gompertz models under competing risks and a generalized type-II hybrid censoring scheme, focusing on life tests of products from two production lines with two failure causes. They derived maximum likelihood estimators and Bayesian estimators using MCMC methods and assessed estimator performance with simulations and a real dataset. The results showed that informative Bayesian priors and bootstrap-t methods offered advantageous estimation.

Conversely, Cannes *et al*., (2023) introduced a conformal prediction technique that provides calibrated, covariate-dependent lower bounds for survival time predictions, suitable for any survival prediction model under Type-I right censoring. They show that their bounds, even when assuming conditionally independent censoring, possess a doubly robust characteristic where marginal coverage is roughly guaranteed if either the censoring mechanism or conditional survival function is accurately estimated. This finding was also confirmed through simulations and analysis of real COVID-19 data from the UK Biobank.

Lee (2023) provides an overview of the general application of simple survival analysis methods, i.e., Kaplan-Meier and Cox proportional hazards models, which are the pillars in medical research when considering time-to-event. The article points out the problem of handling censoring and varying observation times but states the non-parametric nature of Kaplan-Meier and the parametric nature of Cox regression, i.e., the assumption of proportional hazards. The article also details testing methods for this assumption, e.g., the log-minus-log plot, and suggests time-dependent Cox regression as a solution when proportionality is not met. This again emphasizes the value of good model selection and assumption testing to produce valid and coherent survival analysis. These methods, as mentioned, are the basis of statistical techniques on which newer techniques, like the hybrid techniques in your research, attempt to build and extend.

Lu *et al*., (2023) created a hybrid model that integrates CNN and RNN to forecast long-term survival in lung cancer screening participants who succumbed to cardiorespiratory issues. The CNN segment extracted features from CT scan images, while the RNN segment processed time-series data to provide a broader context. Different LSTM models were utilized to manage the irregularities in follow-up times. The integrated model achieved an AUC of 0.76, exceeding human performance in predicting cardiovascular mortality. The Cox Proportional Hazard model validated that including follow-up history enhanced survival predictions (IPCW C-index of 0.75). The results indicate that monitoring cardiorespiratory morbidity can be improved through longitudinal imaging.

Mandel *et al*., (2024) explored different survival analysis methods for predicting car part failures, developing a hybrid model to combine the strengths of Kaplan-Meier, Cox, random survival forest, and gradient boosting. They combined individual models through a weighted sum, and their hybrid model performed better as a predictor. They employed two datasets to understand the impact of the models as part of trying to enhance product reliability and user experience in-vehicle systems.

Germer *et al*., (2024) conducted a comparison of estimators for survival models in lung cancer patients, using the German Schleswig-Holstein cancer registry data. The study contrasted the Cox Proportional Hazards Regression model (CoxPH) with Random Survival Forests (RSF) and two neural network models (DeepSurv and TabNet). Results indicated that CoxPH with MissForest imputation was best when making use of UICC staging, while RSF was best when making use of TNM information. The study also employed explainability metrics to emphasize the importance of the stage of tumor progression and metastasis. This study extends prior work using a new dataset, applying TabNet to survival analysis, and quantifying the impact of imputation.

Salem *et al*., (2024) investigated statistical inference for a generalized progressive hybrid type-II censored Weibull model under competing risks. The study derived maximum likelihood, and Bayesian estimates for Weibull distribution parameters with different scales and a common shape parameter, using Markov Chain Monte Carlo techniques for Bayesian computations. The authors obtained estimates under squared error and linear exponential loss functions with independent gamma priors. Simulation studies and real-world examples were utilized to illustrate the theoretical developments.

In turn, we surveyed a lot of works describing the ongoing advances in survival analysis, from classical models to hybrids and modern techniques such as deep learning. All of these improvements present new channels for adjusting to complex data and coming to more accurate conclusions in various fields.

The current study is novel in presenting a new hybrid approach that blends the traditional Cox, logistic regression, and Kaplan-Meier models with the power of machine learning models (SVM and SVR) through kernel functions. This allows it to process complex and non-linear survival data in a superior manner. Compared to other studies that tend to focus on either traditional models or specific machine learning models, this study combines them to make sure both have their highest advantages and that they provide more flexible solutions than studies focusing on one type of model. Also, it outperforms the limitations of traditional models in assumptions of proportionality and develops models that can be flexible to varying data patterns. Finally, the study focuses on rigorous evaluation of the hybrid models that facilitate a better understanding of what they can achieve and how they can be applied to survival analysis.

2. Methodology

The paper utilizes a comparative approach to the assessment of predictive models for lung cancer patient survival using a combination of classical statistical methods and machine learning. It compares five models: Cox Regression, Logistic Regression, and Support Vector Machines (SVM), together with two hybrid models (Cox-SVM and Logistic-SVM). It follows the Kaplan-Meier method in the selection of variables, after which it conducts separate analyses for each of the models, where the conceptual backgrounds and primary mathematical representations are outlined for each.

2.1 Kaplan-Meier Method

Kaplan-Meier (KM) estimator, a non-parametric approach, calculates the survival probability over time with censored data in consideration. It is used here to identify important predictors by comparing survival curves across variable levels using the log-rank test. The survival function is given as:

$$\hat{S}(t_i) = \prod_{t_j \le t_i} \left(1 - \frac{d_j}{n_j}\right) \qquad (1)$$

where

dj is the number of events(e.g., deaths) at time t_j and n_j is the number at risk just before t_j. Significant variables ($p < 0.05$) are selected for subsequent modeling.

2.2 Cox Regression Mode

Cox Regression, a semi-parametric proportional hazards model, analyzes the effect of covariates on survival time. It assumes a baseline hazard modified by explanatory variables, expressed as:

$$h(t) = h_0\,(t) exp\,\beta_1 x_1 + \beta_2 x_2 + \cdots \dots\dots\dots\dots + \beta_p x_p \qquad (2)$$

where

$H_0(t)$: is the baseline hazard, and B_i are coefficients estimated via partial likelihood. Variable selection uses backward elimination, retaining significant predictors.

2.2.1. Morality Test of the Estimated Model:

Here, it is determined whether the independent variables (X_i) play a role in explaining the behavior of the dependent variable (Y_i). The overall fit of the model is assessed by calculating the coefficient of determination, which we can express in the following formula:

$$R^2 = 1 - exp\left[\frac{2}{n}\left(logL_0 - logL_f\right)\right] \qquad (3)$$

Where

R^2: Value of the coefficient of determination.

L_f: Represents the partial likelihood logarithm of the model containing all the variables.

L_0: Represents the partial likelihood logarithm of the model not containing any variables.

n: The total number of observations.

2.2.2.*Test the Significance of Estimated Parameters*

The second method of testing the significance of the model estimated is testing the significance of each parameter of the model estimated individually. Here, the Wald statistic is used most frequently. Notice that the Wald statistic follows a chi-squared distribution (x^2) with one degree of freedom and can be written as follows:

$$W^2 = \left(\frac{\hat{\beta}_j}{S.E_{\hat{\beta}_j}}\right)^2 \qquad (4)$$

Where

$S.E_{\hat{\beta}_j}$: Represents the standard error of the model parameters.

$\hat{\beta}_j$: Represents the estimated parameter of the model.

2.3 Logistic Regression

Logistic Regression models binary outcomes (e.g., survival vs. death) as a function of predictors, using the logit transformation:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (5)$$

Where p is the probability of the event, and coefficients B_i are estimated via maximum likelihood. It is suited for classification but does not account for time-to-event or censoring.

2.3.1 Goodness-of-fit test for logistic regression

The logistic regression model's fit is evaluated using several criteria, the most important being:

- Hosmer and Lemshow test:
  This test is used to examine whether there is a significant difference between the observed values and the predicted values of the model, using.

  $$R_L^2 = \frac{\chi^2}{-2\ln(l_0)}$$

  χ^2: Represents the difference between the logarithm of the likelihood in the case of the model without independent variables (with only the intercept) and the logarithm of the likelihood in the case of the model with all independent variables.
  Ln (l_0): Represents the logarithm of the likelihood for the model with only the intercept.
  R_L^2: Represents the proportion of reduction in the absolute value of the logarithm of the likelihood, and a measure of the improvement in goodness of fit due to the addition of independent variables. Its value ranges between zero and one.
  $R_L^2=0$: There is no benefit or effect of the independent variables on the dependent variable.
  $R_L^2=1$: There is a perfect benefit or effect of the independent variables on the dependent variable.
  R^2: The coefficient of determination for logistic regression (R^2) is used to test the strength of the logistic model, i.e., the proportion of the contribution of the influencing factors included in the estimated model on the dependent variable (response variable).

- . Cox & Snell R² coefficient (R_{CS}^2):

$$R_{cs}^2 = 1 - e^{\frac{2}{n}[Ln(l_f) - Ln(l_0)]}$$

- Nagelkerke R² coefficient (R_N^2):

$$R_N^2 = \frac{R_{cs}^2}{1 - e^{\frac{2}{n}[Ln(l_0)]}}$$

To estimate the importance of the coefficients in the logistic regression model, apart from the above-discussed Wald statistic, another measure that is reported to be better than the Wald statistic is partial R². This measure assesses the relative contribution of the independent variables in the model and is calculated based on the χ² value. A statistically significant χ² value signifies that the independent variable is statistically significant, and vice versa, thus justifying the inclusion of the variable in the final model.

2.4 Support Vector Machines (SVM)

SVM, a supervised learning algorithm, classifies data by finding an optimal hyperplane that maximizes the margin between classes. For non-linear data, a Radial Basis Function (RBF) kernel is applied:

$$K(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2) \qquad (6)$$

The decision function is:

$$y = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b) \qquad (7)$$

Where;

$_i$ am Lagrange multipliers,y_i are class labels, and b is the bias term. Parameters(γ, c) α

control flexibility and regularization

2.5 Hybrid Models

1. Cox-SVM: Combines Cox Regression's hazard-based variable selection with SVM's classification power. Significant variables from Cox are input into SVM with an RBF kernel $(\epsilon = 0.1, C = 1)$

2. Logistic-SVM: integrates Logistic Regression's predictor identification with SVM's non-linear classification, using the same kernel parameters.

These hybrids aim to enhance accuracy by leveraging statistical insights and machine learning adaptability.

2.6 Comparison Between Individual and Hybrid Models

The classification table is a cornerstone of model evaluation, summarizing predictive performance by comparing observed outcomes (e.g., survival or death, early or late stage) against model predictions. For this study, a binary classification framework is adopted: Class 1 (early-stage/survived) and Class 2 (late-stage/did not survive). The table structure is as follows:

classification table

Predicted \ Observed	Class 1 (Negative)	Class 2 (Positive)
Class 1 (Negative)	True Negative (TN)	False Negative (FN)
Class 2 (Positive)	False Positive (FP)	True Positive (TP)

1. **True Positives (TP):** Correctly predicted Class 2 cases (e.g., late-stage patients identified as such).

2. **True Negatives (TN):** Correctly predicted Class 1 cases (e.g., early-stage patients identified as such).

3. **False Positives (FP):** Class 1 cases incorrectly predicted as Class 2 (e.g., early-stage patients misclassified as late-stage).

4. **False Negatives (FN):** Class 2 cases incorrectly predicted as Class 1 (e.g., late-stage patients misclassified as early-stage).

Model Comparison Criteria and Metrics

1. Accuracy: The proportion of correct predictions, calculated as:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (8)$$

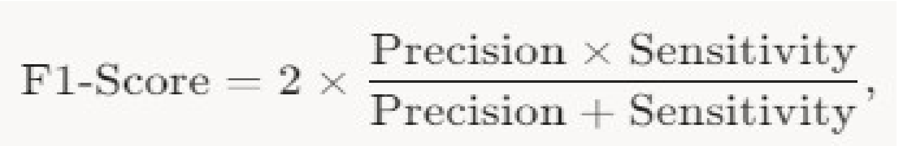
2. Sensitivity (Recall): The ability to detect Class 2 cases, crucial for identifying high-risk patients:

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \qquad (9)$$

3. Specificity: The ability to correctly identify Class 1 cases, minimizing false alarms:

$$\textbf{Specificity} = \frac{TN}{TN+FP} \qquad (10)$$

4. F1-Score: The harmonic mean of precision and sensitivity, balancing detection and reliability:



$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}},$$

Where

$$\textbf{Precision}= \frac{TP}{TP+FP} \qquad (11)$$

5. Misclassification Rate: The proportion of incorrect predictions, indicating error magnitude:

$$\textbf{Misclassification Rate} = \frac{FP+FN}{TP+TN+FP+FN} \qquad (12)$$

6. R² (where applicable): Variance explained by predictors in Cox and Logistic models, assessing explanatory power.

3. Applied study

This research aims to conduct a comparative study between individual models—namely, Cox regression, logistic regression, and support vector machines (SVM)—and hybrid models, specifically a hybrid Cox-SVM model and a hybrid logistic regression-SVM model. The purpose is to analyze survival time, and characterize, and predict outcomes for a sample of 165 lung cancer patients from the Tanta Oncology Institute and Kafr El-Sheikh Chest Hospital from 2020 to 2024. The analysis will be performed using statistical software packages including SPSS 28, Stata 15, and Stat Graphics 19. The research variables are described as follows:

Table 1: Description and Coding of Variables in the Lung Cancer Survival Analysis Study

Variable Type	Variable Name	Description	Coding
Dependent	Patient Survival Time (Months)	Time interval between the date of diagnosis and the date of death, measured in months.	
	Censoring Status	indicates whether the event (death) occurred or if the patient was censored.	0=Loss to Follow-up, 1 = Event Occurrence (Death)
Independent	X_1: Smoking	Categorical variable indicating whether the patient is a smoker.	1=Smoker, 2 = Non-Smoker
Independent	X_2: Family History	Categorical variable indicating the presence or absence of a family history of disease.	1=Family History Present, 2 = Family History Absent
Independent	X_3: Occupation	A categorical variable representing the patient's occupation.	1=Employee, 2 = Worker, 3 = Retired, 4 = Housewife, 5 = Child, 6 = Student
Independent	X_4: Gender	Patient's gender.	1=Male, 2 = Female
Independent	X_5: Age (Years)	The patient's age is in years at the time of diagnosis.	Continuous (measured in years(
Independent	X_6: Radiation Therapy Exposure	Whether the patient received radiation therapy.	1=Exposed to Radiation Therapy, 2 = Not Exposed to Radiation Therapy
Independent	X_7: Treatment Methods	The treatment methods used for the patient.	1=Surgery, 2 = Chemotherapy, 3 = Radiation, 4 = More than One Method
Independent	X_8: Social Status	Patient's marital status.	1=Married, 2 = Single
Independent	X_9: Consanguinity	Indicates if the parents of the patient were related.	1=Consanguinity Present, 2 = No Consanguinity
Independent	X_{10}: COVID-19 Infection	Whether the patient was infected with COVID-19.	1=Infected, 2 = Not Infected
Independent	X_{11}: Chronic Diseases	Presence of chronic diseases.	1=Chronic Diseases Present, 2 = No Chronic Diseases
Independent	X_{12}: Place of Residence	The patient's primary place of residence.	1=Urban, 2 = Rural
Independent	X_{13}: Stage of Disease	The stage of lung cancer at the time of diagnosis.	1=Stage I, 2 = Stage II, 3 = Stage III, 4 = Stage IV

For assessing the variables that are most significant and relevant for survival time, the Kaplan-Meier method and log-rank test are initially used. It takes into account the data by comparing each variable individually with the dependent variable, i.e., survival time. This is used before using any of the aforementioned statistical methods. The null hypothesis of the test is H_0: $h(t/x_i) = 0$, i.e., there are no statistically significant differences in survival functions at different levels of the variable (x_i). The alternative hypothesis is H_1: $h(t/x_i) \neq 0$, i.e., there are statistically significant differences in survival functions at different levels of the variable x_i. If yes, then the variable is included in the model. The results were as follows:

Table 2: Kaplan-Meier and Log-Rank Test Results

Variable	Levels	n	Events	Log-Rank	P-value
X_1(Smoking)	1= Smoker	190	155	12.78	0.00216
	2= Non-Smoker	70	23		
X_2(Family History)	1= Family History Present	127	66	2.23	0.451
	2= Family History Absent	133	40		
(X_3 Occupation)	1= Employee	89	15	15.278	0.004
	2=Worker	92	72		
	3== Retired	30	5		
	4= Housewife	20	7		
	5= Child	11	4		
	6= Student	18			
X_4(Gender)	1= Male	178	77	1.265	0.354
	2= Female	82	34		
X_5(Age)	≤60	102	73	8.057	0.382
	>60	156	84		
X_6(Radiation Therapy)	1= Exposed to Radiation Therapy	192	52	5.320	0.0263
	2= Not Exposed to Radiation Therapy	66	27		
X_7 Treatment Methods)	1= Surgery	79	25	4.894	0.0486
	2= Chemotherapy	83	53		
	3= Radiation	92	71		
	4= More than One Method	6	3		
X_8(Social Status)	1= Married	163	91	6.12	0.078
	2= Single	97	58		
(X_9Consanguinity)	1=Consanguinity Present	77	27	2.231	0.451
	2= No Consanguinity	183	63		
X_{10}(COVID-19 Infection)	1= Infected	123	97	10.850	0.046
	2= Not Infected	137	38		
X_{11} (Chronic Diseases)	1= Chronic Diseases Present	89	46	1.125	0.286
	2= No Chronic Diseases	171	96		
X_{12} (Place of Residence)	1= Urban	178	63	1.705	0.362
	2= Rural	82	22		
X_{13}(Stage of Disease)	1= Stage I	78	42	13.785	0.002
	2= Stage II	69	23		
	3= Stage III	81	48		
	4= Stage IV	32	29		

This article provides a statistical analysis of patient data of lung cancer using Kaplan- Meier, and Log-Rank tests for variables of significant influence on survival time. Kaplan-Meier estimates the survival function, which plots the probability of survival over time, and the Log-Rank test compares survival curves between groups, testing the null hypothesis of no difference. A P-value < 0.05 rejects this hypothesis, indicating significant differences in survival. Results show that smoking (X_1), occupation (X_3), radiation therapy (X_6), treatment methods (X_7), COVID-19 infection (X_{10}), and disease stage (X_{13}) significantly affect survival. Smoking is associated with reduced survival because of lung impairment and aggressive tumor types; occupation can be associated with exposure to carcinogens or access to healthcare; radiation therapy could represent advanced disease or side effects that affect survival; variations in treatment type could be due to differences in treatment efficacy by cancer type and stage; COVID-19 can worsen the condition of lung cancer patients; and later disease stage is associated with worse prognosis. Family history, gender, age, social status, consanguinity, chronic illness, and residence were not significantly influential, where small numbers or confounding factors may obscure effects. As an observational study, causal inferences are limited, selection biases are possible, and results must be interpreted clinically in the context of individual patient circumstances and treatment availability.

3.1 Cox – Regression

Cox regression is a method of survival analysis used for modeling the time to an event. It has among its major assumptions the requirement that the dependent variable be composed of two parts: the survival time and a binary (dichotomous) indicator variable. From variables that had been proven statistically significant from the results of prior studies and affect survival time via the Kaplan-Meier method, that is, seven variables: X_1 (Smoking), X_3 (Occupation), X_5 (Age), X_6 (Exposure to Radiotherapy), X_7 (Treatment Methods), X_{10} (COVID-19 Infection), and X_{13} (Disease Stage), we will examine the goodness of fit of the model.

Table 3: Goodness-of-Fit Test for the Cox Regression Model

Model	-2log likelihood	κ^2	Df	p-value
Block0	604.96	-	-	-
Block1	545	59.96	7	0.000

The goodness-of-fit of the Cox regression model was evaluated using the likelihood ratio test, as detailed in Table 3. This test compares the null model (Block0), with a -2log likelihood of 604.96, against the full model (Block1), which includes seven predictors—Smoking (X_1), Occupation (X_3), Age (X_5), Exposure to Radiotherapy (X6), Treatment Methods (X_7), COVID-19 Infection (X_{10}), and Disease Stage (X_{13})—yielding a -2log likelihood of 545. The resulting chi-square statistic of 59.96, with 7 degrees of freedom, and a p-value less than 0.001, indicates a statistically significant improvement in model fit. This confirms that these predictors collectively enhance the model's ability to explain variations in survival time, supporting their relevance in the survival analysis. When estimating a Cox regression model and using backward elimination, the variable 'Exposure to radiation therapy' (X_6) was excluded. The results are as follows:

Table :4 Results of the Backward Elimination Method

Variables	β_i	S.E	Wald	EXp(β)
X_1	1.16	0.339	11.789	3.189
X_3	-0.197	0.118	2.787	0.821
X_5	0.016	0.006	7.111	1.016
X_7	0.596	0.151	15.578	1.815
X_{10}	1.066	0.287	13.796	2.903
X_{13}	0.590	0.128	21.246	1.804
Block1				
-2log likelihood	χ^2	DF	p-value	
528.362	76.598	6	0.000	

Table 4, labeled "Results of the Backward Elimination Method," details a Cox regression analysis evaluating survival time among lung cancer patients, demonstrating strong overall model significance with a -2log likelihood of 528.362, 6 degrees of freedom (Df), and a p-value < 0.001, affirming that the included variables—X_1 (Smoking), X_3 (Occupation), X_5 (Age), X_7 (Treatment Methods), X_{10} (COVID-19 Infection), and X_{13} (Disease Stage)—collectively account for survival variability. The regression coefficients (βi) indicate pronounced positive effects from X_1 (β = 1.16, Exp(β) = 3.189), tripling the hazard rate for smokers, and X_{10} (β = 1.066, Exp(β) = 2.903), nearly tripling it for COVID-19 cases, alongside moderate increases from X_7 (β = 0.596, Exp(β) = 1.815) and X_{13} (β = 0.590, Exp(β) = 1.804). In contrast, X_5 (β = 0.016, Exp(β) = 1.016) shows a minimal per-year risk increase, while X_3 (β = -0.197, Exp(β) = 0.821) suggests a slight protective effect, though its significance is marginal (Wald = 2.787, $p \approx 0.05$).

In Table 4, Wald statistics underscore X_{13} (21.246) as the dominant predictor, followed by X_7 (15.578) and X_{10} (13.796), all highly significant ($p < 0.05$), with X_1 (11.789) and X_5 (7.111) also impactful, whereas X_3's lower Wald value (2.787) advises cautious interpretation. Standard errors remain tight (e.g., 0.006 for X_5, 0.339 for X_1), enhancing coefficient reliability, yet the exclusion of X_6 (Radiation Therapy) through backward elimination implies limited statistical contribution in this context, despite its potential clinical relevance. This model effectively ranks X_{13}, X_7, and X_{10} as primary survival influencers, with X_1 reinforcing smoking's critical role, though verifying the proportional hazards assumption and reassessing excluded variables like X_6 could further strengthen its applicability.

Cox Proportional Hazards Model Equation:

The hazard function can be represented as:

$h(t/x_i) = ho(t) \exp (1.16X_1 – 0 .197 X_3 + 0 .016X_5 + 0 .596X_7 + 1.066X_{10} + 0.590X_{13})$

The coefficient of determination is calculated as follows:

$$R^2_{cox} = 1 - \exp \frac{2}{n} [\ln Lf - \ln Lo]$$

1- exp (2/165)[528.362-604.96]=

=60.5%

An R-squared of 61.4% in the Cox model indicates that the independent variables explain 61.4% of the variance in survival time for lung cancer patients. A classification table was then created based on the hazard and survival functions, and the results were:

Table 5: classification table for Cox-Regression

	Observed Survived(0)	Observed Did Not Survive(1)	Correct percent
Low Predicted Risk (0)	73	20	77.4%
High Predicted Risk (1)	12	60	83.33%
Over percent	85	80	80.6%

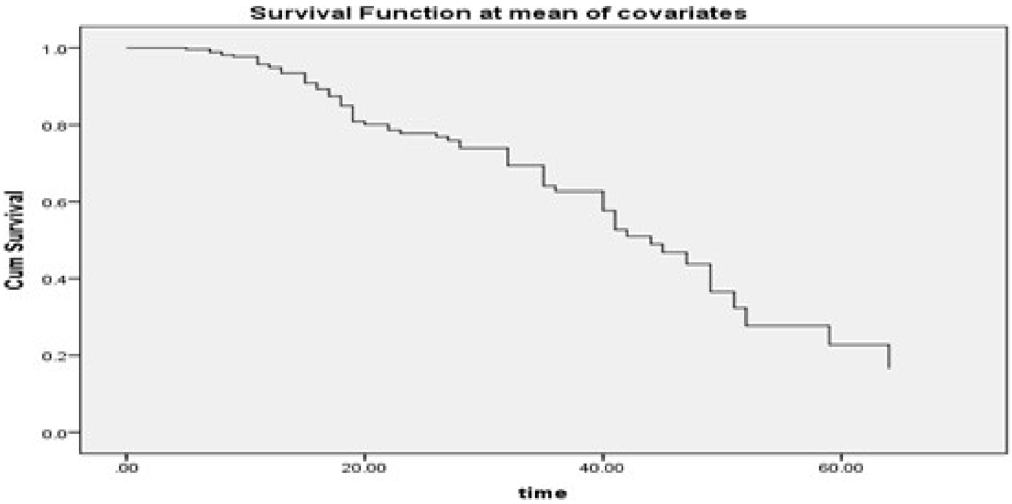
Based on Table 5, we can define the following classification outcomes: True Positives (TP) are the 60 individuals correctly predicted to not survive, found in the High Predicted Risk (1) row and Observed Did Not Survive (1) column; True Negatives (TN) are the 73 individuals correctly predicted to survive, located in the Low Predicted Risk (0) row and Observed Survived (0) column; False Positives (FP) are the 12 individuals incorrectly predicted to not survive (High Predicted Risk (1) but Observed Survived (0)); and False Negatives (FN) are the 20 individuals incorrectly predicted to survive (Low Predicted Risk (0) but Observed Did Not Survive (1).

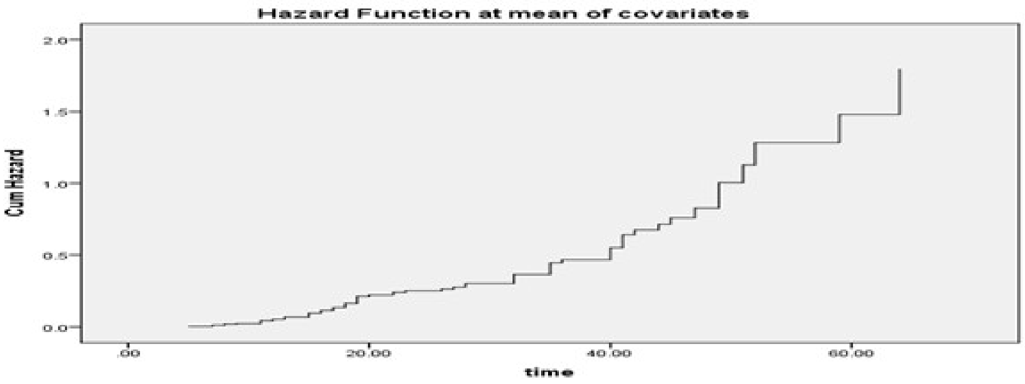
To calculate the model's performance, we compute different measures of performance. Sensitivity, or true positive rate, is a measure of the ability of the model to predict accurately those who did not survive, as calculated by the formula TP / (TP + FN) = 60 / (60 + 20) = 75%. It indicates that the model accurately predicted 75% of the passengers who did not survive. Specificity or true negative rate measures how accurately the model predicted the survivors. It is given by TN / (TN + FP), which is equal to 73 / (73 + 12) = 85.88%. The model correctly predicted the survivors with 85.88% accuracy. Accuracy, the proportion of correct classification of all instances overall, is (TP + TN) / (Total) = (60 + 73) / 165 = 80.6%, i.e., the model has correctly classified 80.6% of all instances. Finally, the misclassification rate, the proportion of misclassifications, is (FP + FN) / (Total) = (12 + 20) / 165 = 19.39%, i.e., the model has misclassified 19.39% of all instances.

These performance metrics provide valuable insights into the model's predictive capabilities but should be interpreted within the context of lung cancer survival. While the model demonstrates reasonably good accuracy (80.6%), sensitivity (75%), and specificity (85.88%), the relative importance of each metric depends on the clinical priorities. For instance, if early intervention for high-risk patients significantly improves outcomes, prioritizing sensitivity may be crucial, even at the cost of a slightly lower specificity. Conversely, if avoiding unnecessary interventions for low-risk patients is paramount, prioritizing specificity would be more appropriate. It's also important to note that the Cox model assumes proportional hazards and alternative variable selection techniques beyond the backward elimination used here might further optimize the model's predictive performance.

Based on the Wald statistic, a measure of the significance of each predictor variable, we can rank the importance of factors influencing survival time in this model. Disease Stage (X_{13}) emerges as the most significant predictor (Wald = 21.246), followed by Treatment Methods (X_7, Wald = 15.578), COVID-19 Infection (X_{10}, Wald = 13.796), and Smoking Status (X_1, Wald = 11.789). Age (X_5, Wald = 7.111) shows a moderate association, while Occupation (X_3, Wald = 2.787) exhibits the weakest association. It's important to remember that statistical significance, as reflected by the Wald statistic, doesn't always equate to practical importance, and the model-specific nature of these findings and potential multicollinearity among predictors should also be considered in the interpretation.

Regarding the behavior of the survival and hazard functions, we observe from the following figures:





The survival function, demonstrating a typical decreasing trend, illustrates the proportion of lung cancer patients surviving over time, while the hazard function reveals a mostly flat then suddenly increasing hazard of death, meaning the immediate risk of death is somewhat constant but then gradually increases throughout the study period.

3.2 logistic Regression

Logistic regression is advantageous due to its lack of assumptions regarding the normality of independent variables and the linear relationship between independent and dependent variables. Furthermore, it excels in classification and prediction tasks. The model fit of a logistic regression model for the given data is being evaluated based on independent variables that have demonstrated statistical significance in influencing survival time, as determined by Kaplan-Meier analysis. These variables are: Smoking (X_1), Occupation (X_3), Age (X_5), Radiotherapy Exposure (X_6), Treatment Methods (X_7), COVID-19 Infection (X_{10}), and Disease Stage (X_{13}). These variables were found to be statistically significant predictors of survival time**.** To assess the model's goodness-of-fit and its suitability for the data, the following will be examined:

Table 6: Goodness-of-Fit Test for the Logistic Regression Model

χ^2	Df	p-value
43.87	7	0.000

Table 6 reveals that the χ^2 value, representing the difference between the log-likelihood of the model with independent variables and the log-likelihood of the model with only the intercept (i.e., without independent variables), yields a P-value of 0.000. Since this P-value is less than 0.05, it indicates a good model fit and its suitability for the data. Consequently, we proceed to estimate the logistic regression model. Using the Backward Elimination Method, the variables Smoking (X_1), Occupation (X_3), Age (X_5), Treatment Methods (X_7), and Disease Stage (X_{13}) were found to be significant. Radiotherapy Exposure (X_6) and COVID-19 Infection (X10) were excluded due to their statistical insignificance. To test the validity of the new logistic model with five independent variables, the Hosmer-Lemeshow test was employed, and the results are as follows.

Table 7: Hosmer-Lemeshow test

χ^2	Df	sig
4.587	5	0.778

Table 7 presents the results of the Hosmer-Lemeshow test for assessing the goodness of fit of the logistic regression model. The test yielded a Chi-square value of 4.587 with 5 degrees of freedom and a P-value of 0.778, which is much greater than the conventional significance level of 0.05. This indicates that we fail to reject the null hypothesis, suggesting that the model fits the data well, as there are no significant differences between the observed and expected frequencies. Consequently, the estimated model, which includes the five independent variables—Smoking (X_1), Occupation (X_3), Age (X_5), Treatment Methods (X_7), and Disease Stage (X_{13})—can be reliably used for prediction and analysis purposes.

Table 8: Goodness-of-Fit of the Logistic Regression Model: R-Square Measures

log-likelihood -2	Cox 8 snell R square	Negelkereke R square
540.62	0.5415	0.5418

Table 8 presents the R-square measures for the logistic regression model, where the Cox & Snell R-square is 0.5415 and the Nagelkerke R-square is 0.5418. The Cox & Snell R-square indicates that the model explains approximately 54.15% of the variance in the dependent variable, though it has a theoretical limitation of not reaching 1. Corrected for this limitation, the Nagelkerke R-square suggests the model explains around 54.18% of the variance. These R-square values, suggesting a reasonably strong fit, must be considered with other diagnostics to evaluate the overall adequacy of the model for the specific data, accounting for variance left unexplained. The parameters of the logistic regression were estimated as follows:

Table 9 : Estimation of Logistic Regression Parameters**.**

Variables	β_i	S.E	Wald	Sig	EXp(β)
X_1	1.288	0.447	8.302	0.003	3.625
X_3	-0.619	0.187	10.957	0.001	0.538
X_5	0.019	0.008	5.641	0.035	1.019
X_7	0.461	0.215	4.597	0.031	1.585
X_{13}	0.576	0.187	9.488	0.002	1.778
constant	2.214	1.083	4.179	0.042	9.150

The logistic regression model in Table 9 reveals significant predictors of lung cancer survival, with coefficients (βi) indicating their impact. Smoking (X_1, $\beta = 1.288$, Wald = 8.302, $p = 0.003$) markedly increases the odds of non-survival by 3.625 times (Exp(β)), reflecting its strong influence, supported by a modest standard error (S.E = 0.447). Disease Stage (X_{13}, $\beta = 0.576$, Wald = 9.488, $p = 0.002$) follows with a 1.778-fold increase in odds per stage, its tight S.E (0.187) reinforcing estimate reliability. Occupation (X_3, $\beta = -0.619$, Wald = 10.957, $p = 0.001$) shows the highest statistical significance, reducing odds by 46.2% (Exp(β) = 0.538), suggesting a protective effect depending on category coding, though its interpretation requires context. Treatment Methods (X_7, $\beta = 0.461$, Wald = 4.597, $p = 0.031$) moderately elevates risk by 58.5% (Exp(β) = 1.585), while Age (X_5, $\beta = 0.019$, Wald = 5.641, $p = 0.035$) has a minimal per-year effect (1.9%, Exp(β) = 1.019), possibly due to limited age variance or confounding factors. The constant ($\beta = 2.214$, Wald = 4.179, $p = 0.042$) sets a baseline odds of 9.150, though its higher S.E (1.083) suggests some variability.

The model is highly statistically coherent, with all variables significant ($p < 0.05$) and low S.E values to enhance estimation confidence, except for the mild instability of the constant. Occupation (X_3) is the most significant (Wald = 10.957), followed by Disease Stage (X_{13}) and Smoking (X_1), as in clinical practice, while Age (X_5) and Treatment Methods (X_7) are less significant. The equation was a good fit for:

$$\text{logit}(p) = 2.214 + 1.288(X_1) - 0.619(X_3) + 0.019(X_5) + 0.461(X_7) + 0.576(X_{13})$$

Captures non-survival log-odds, with a strong predictive model (54.18% variance explained per Nagelkerke R^2). However, X_3's negative impact calls for further exploration of category-specific effects, and X_5's weak contribution suggests potential interaction effects (e.g., $X_1 \times X_{13}$) for future research. Overall, the model is successful in capturing prevailing survival determinants, though refinement in variable interactions and coding can make it more accurate. To analyze the goodness of fit of the model in classification, the following classification table was tabulated:

Table 10: Classification Table for Logistic Regression

	Observed Survived (0)	Observed Did Not Survive (1)	Correct Percent
Low Predicted Risk (0)	72	21	77.42%
High Predicted Risk (1)	32	40	55.56%
Overall Percent			67.88%

Table 10 presents the classification performance of the logistic regression model for lung cancer survival prediction, with 77.42% of low-predicted risk cases correctly classified (72 true negatives [TN] out of 93, with 21 false negatives [FN]) and 55.56% of high predicted risk cases correctly classified (40 true positives [TP] out of 72, with 32 false positives [FP]), yielding an overall accuracy of 67.88% (112/165). From this, we calculate sensitivity as TP / (TP + FN) = 40 / (40 +

21) ≈ 65.57%, indicating the model identifies 65.57% of those who did not survive; specificity as TN / (TN + FP) = 72 / (72 + 32) ≈ 69.23%, showing it correctly identifies 69.23% of survivors; precision as TP / (TP + FP) = 40 / (40 + 32) ≈ 55.56%, reflecting that 55.56% of predicted non-survivors did not survive; and the F1-score as 2 × (Precision × Sensitivity) / (Precision + Sensitivity) = 2 × (0.5556 × 0.6557) / (0.5556 + 0.6557) ≈ 0.6017 (60.17%), balancing precision.

3.3 Support vector machine

Support Vector Machine (SVM) was used to model the survival time for a group of 165 lung cancer patients and classify them according to variables determined to be significant by using the Kaplan-Meier method. The variables are smoking (x_1), occupation (x_3), age (x_5), radiation therapy exposure (x_6), treatment type (x_7), COVID-19 infection (x_{10}), and stage of disease (x_{13}). The calculation was carried out using the statistical package Stat Graphics 19 and the R programming language to determine the support vector function, using the Radial Basis Function (RBF) kernel and with tolerance on the error of e=0.1 and regularization parameter c=1. The outcomes of applying the SVM to classify the lung cancer patients into whether they are in the first group, which signifies an early stage, or the second group, which signifies the late stage, are as follows:

Table 11: Classification Results Using Support Vector Machine (SVM)

Actually	Group Size	Predicted Class 1	Predicted Class 2
Class 1	125	103	22
Class 2	40	8	32

Table 11 details the classification results of a Support Vector Machine (SVM) model with an RBF kernel, applied to 165 lung cancer patients split into early-stage (Class 1, n=125) and late-stage (Class 2, n=40) groups, using significant Kaplan-Meier variables. The model correctly classifies 103 out of 125 Class 1 patients and 32 out of 40 Class 2 patients, achieving an overall accuracy of (TP_1 + TP_2) / Total = (103 + 32) / 165 = 135 / 165 = 81.8%. For Class 1, sensitivity is TP_1 / (TP_1 + FN_1) = 103 / (103 + 22) = 82.4%, precision is TP_1 / (TP_1 + FP_1) = 103 / (103 + 8) = 92.8%, and F1-score is 2 × (0.928 × 0.824) / (0.928 + 0.824) = 87.3%. For Class 2, sensitivity is TP_2 / (TP_2 + FN_2) = 32 / (32 + 8) = 80%, precision is TP_2 / (TP_2 + FP_2) = 32 / (32 + 22) = 59.3%, and F1-score is 2 × (0.593 × 0.8) / (0.593 + 0.8) = 68.1%. Overall model metrics include macro-average sensitivity of (0.824 + 0.8) / 2 = 81.2%, macro-average precision of (0.928 + 0.593) / 2 = 76.05%, macro-average F1-score of (0.873 + 0.681) / 2 = 77.7%, and a misclassification rate of (FP + FN) / Total = (8 + 22) / 165 = 18.2%. Class 1 shows high precision, indicating reliable early-stage detection, while Class 2's lower precision reflects a higher false positive rate (22 out of 54), likely due to the dataset's imbalance favoring Class 1.

3.3.1 Support vectors

The support vectors in a Support Vector Machine (SVM) model represent a subset of observations—namely, the patients—closest to the separating hyperplane that divides the two groups within the data space, which, in this case, are Class 1 (early-stage) and Class 2 (late-stage). These support vectors are the critical data points that determine the position and orientation of the hyperplane, as the model aims to maximize the margin—the distance between the hyperplane and the nearest data points from each class—to ensure optimal classification of new observations. The identification of support vectors is based on the output of the following function: y_i ($w^T x_i$+b) where y_i is the actual label of the patient (e.g., +1 for Class 1 and -1 for Class 2), W is the weight vector defining the hyperplane's direction, X_i is the feature vector for patient I, and b is the bias term determining the hyperplane's position in space. Points satisfying the condition y_i (w^T X_i)=1 or -1 are deemed support vectors, as they lie directly on the margin's boundaries. As illustrated in the following table (presumed to contain additional details such as vector values or classifications), this formula calculates each patient's relative distance from the hyperplane, aiding in designating them as either support vectors or regular data points within their respective classes.

To elaborate further, in the context of Table 11, if we assume the model relies on variables such as smoking (X_1), age (X_5), and others(X_I) represents the set of values for these variables for each patient. The support vectors would be those records (patients) whose values are nearest to the boundary separating the two classes, directly influencing the final model's configuration. For instance, a patient in Class 1 with values very close to Class 2 might become a support vector, as it defines the minimal margin from Class 1's side. The same applies to Class 2. This process ensures the model focuses on the most challenging cases to distinguish, enhancing its ability to handle complex data, such as that of lung cancer patients, where variables like disease stage and treatment may intersect in non-linear ways. By prioritizing these boundary points, the SVM achieves robustness and precision, particularly in scenarios with intricate patterns, as seen in this medical application

1. Support vectors in the first group

Observation	X_1	X_3	X_5	X_6	X_7	X_{10}	X_{13}
5	0.6128	0.4856	2.5463	0.2553	0.9187	-0.2148	1.3695
7	-0.6725	1.0564	2.1780	-1.3111	1.2225	0.2856	0.7778
9	0.7263	0.2324	0.3645	-0.4527	-1.7894	0.9875	-1.7758
10	0.5638	0.4465	-1.7754	0.2223	0.5284	-0.7785	1.2221
11	-0.7125	-2.1250	3.7889	0.3778	0.4447	1.2223	-0.3189
15	-1.9231	0.9785	-0.4621	2.3695	1.9990	0.6555	0.6444
18	2.1350	0.6389	4.6890	0.8812	0.9998	0.4879	0.7585

19	0.3458	0.7831	1.2339	0.6345	0.8758	-1.8721	0.9634
25	1.6854	1.0233	-0.9967	1.9780	-0.7869	2.9640	1.6502
38	0.6890	0.8632	-0.7236	-2.7890	2.6890	0.7861	1.9877
65	-0.4685	1.2854	2.6500	0.6587	1.8882	0.9928	0.8787
83	2.1964	-0.2814	1.9601	0.5528	0.4452	0.3891	0.9631
95	-0.3962	0.7784	2.1145	0.5777	0.7778	0.4777	1.7888
101	1.0235	0.7564	-0.2478	-0.6669	0.6777	0.6891	0.4559
115	0.8647	1.6897	3.3368	1.7778	0.4568	-0.8881	-0.8847
119	1.0235	-.2545	0.6115	2.1450	-0.9639	1.8876	2.1987
124	-0.6458	2.7890	0.6298	1.8790	1.9821	0.6624	0.7788
131	-1.9645	0.6354	1.3698	-0.8890	0.5559	0.9991	0.4630
141	2.1369	0.3356	1.3335	0.4447	1.9630	-0.3352	2.5580
148	-0.1254	-0.2546	0.3564	0.8702	-2.3690	1.9967	-0.8445
155	2.1879	-0.4586	-0.4559	0.9639	0.6669	0.5881	0.9652
162	-2.0147	0.6354	2.8889	0.9987	0.8025	-0.4586	1.6350

From the previous table, the number of support vectors in the first group reached (22) support vectors. The first column displays the observations representing the support vectors when using the Support Vector Machine (SVM). The remaining columns represent the values of the variables for those observations or patients. According to the statistical program, the standardized values of the variables were used (Standardize), meaning that the mean was subtracted and then divided by the standard deviation.

2. Support vectors in the second group

Observation	X_1	X_3	X_5	X_6	X_7	X_{10}	X_{13}
27	0.5850	1.4440	0.6852	0.9969	-0.7744	0.7889	0.6547
31	0.3987	-0.9660	2.6963	0.8756	1.3654	0.6354	0.9875
33	-1.7766	2.3335	-0.7780	-0.4778	0.6665	0.7854	1.6548
39	2.1150	0.4499	3.6354	1.6545	2.4566	1.2354	-0.4562
44	-0.9788	0.6655	0.8895	0.5566	-0.6687	-0.8547	0.6665
49	0.8866	0.6365	-0.7778	0.6589	0.5556	0.4056	0.7546
55	0.7863	-0.4499	1.6879	0.7785	0.7584	1.5558	0.4566
97	-0.3645	1.6677	3.8575	0.6398	0.7785	0.6777	2.6540
133	1.4563	0.8881	0.6963	1.7895	0.4566	1.3652	1.2350
145	2.7895	0.7733	2.7832	0.6698	0.8888	0.5569	0.4578
151	-0.7744	-0.4748	-0.4555	0.8945	0.6540	0.8885	0.8876
157	0.6511	0.6578	3.6540	0.5858	0.8546	0.7754	0.8546
161	0.6897	0.6662	0.8654	1.4565	0.7444	-0.7778	0.3654

The number of support vectors in the second group reached 13 vectors. The first column represents the observations that are the support vectors when using the Support Vector Machine (SVM), which starts with the observations (27, 31, 39, ..., 161). Therefore, the total number of support vectors in both groups is 35.

3. Classification of observations in the first group

"Observations are classified based on the output of the function $y = W^{\mathrm{T}}X + b$. A patient is classified as being in an early stage of the disease and belonging to the first group, Class (1), or as being in a late stage of the disease and belonging to the second group, Class (2). To identify the observations (patients) in the first group that were incorrectly classified into the second group, see the following:

Observation	Value (y)	Observation	Value (y)	Observation	Value (y)
6	0.02458	51	0.05478	129	0.03540
13	0.06587	53	0.02580	132	0.04578
16	0.01245	77	0.06528	145	0.05890
21	0.078952	79	0.078546	150	0.058457
22	0.01254	81	0.045754	156	0.04578
30	0.02361	105	0.0.4578	159	0.047854
32	0.035556	118	0.085487		
35	0.045877	127	0.04578		

It is noted that there are (22) observations from the first group that were incorrectly classified into the second group. This is called a classification error, and it corresponds with what is presented in the classification table. As for the observations that belong to the second group and were incorrectly classified into the first group, they are:

Observation	Value (y)	Observation	Value (y)
38	-0.06879	100	-0.09725
44	-0.07854	102	-0.07421
53	-0.04571	129	-0.09318
88	-0.06595	141	-0.07335

From the previous table, there are (8) patients from the second group, meaning in a late stage of the disease, who were incorrectly classified into the first group as being in an early stage of the disease. This is consistent with the outputs of the classification table. To classify new observations into one of the two groups, it is necessary to estimate the hyperplane equation, which is $y = W^{T}X + b$, and consequently obtain estimates for the weight vector (W) and the bias term (b), as shown in the following table.

Table12: the weight vector (W) and the bias term (b)

	X_1	X_3	X_5	X_6	X_7	X_{10}	X_{13}	b
$\boldsymbol{W}_i$	4.5871	2.1425	1.9454	0.2457	0.37854	0.5864	0.6857	0.3571

The previous table shows the weights of the variables. The weight value for the smoking variable (x_1) is 4.5871, the weight value for the occupation variable (x_3) is 2.1425, the weight value for the age variable (x_5) is 1.9454, the weight value for the exposure to radiation treatment variable (x_6) is 0.24.057, the weight value for the treatment methods variable is 0.3785, the weight value for the COVID-19 infection variable (x_{10}) is 0.5864, and the weight value for the disease severity variable (x_{13}) is 0.6857. The value of the constant term was randomly selected using the R program and is 0.3571. Based on this, new items are classified as follows:

Table 13: Classification of New Lung Cancer Patients Using Support Vector Machine

Observation	X_1	X_3	X_5	X_6	X_7	X_{10}	X_{13}	Predict
166	1	1	75	1	3	1	4	Class 2
167	2	2	55	2	1	2	1	Class 1
168	1	3	64	1	4	1	2	Class 2

For example, if there is a patient who is 75 years old, a smoker, employed, infected with COVID-19, exposed to radiation treatment, is treated with radiation, and is in the fourth stage of the disease, they would be classified into the second group, meaning in a late stage of the disease, and so on.

3.4 Hybrid Model (Cox-SVM)

This model was constructed by inputting all independent variables affecting lung cancer patient survival time (a total of 13 variables previously mentioned) into a Kaplan-Meier analysis. At this stage, the following variables were excluded: family history (x_2), gender (x_4), social status (x_8), parental consanguinity (x_9), chronic diseases (x_{11}), and place of residence (x_{12}). Only seven variables with statistical significance were retained. These variables were then entered into a Cox regression model, and using backward elimination, the variable exposure to radiation therapy was excluded. Six variables were retained: smoking (x_1), occupation (x_3), age (x_5), treatment methods (x_7), COVID-19 infection (x_{10}), and disease stage (x13). These six variables were then input into a Support Vector Machine (SVM) using a Radial Basis Function (RBF) kernel with ε=0.1 and C=1. The classification results based on the hybrid model were as follows:

Table 14: Classification Results Using Hybrid Model (Cox-SVM)

Actually	Group Size	Predicted Class 1	Predicted Class 2
Class 1	125	120	5
Class 2	40	4	36

The hybrid Cox-SVM model, as evaluated through Table 14, demonstrates exceptional performance in classifying 165 lung cancer patients into early-stage (Class 1, n=125) and late-stage (Class 2, n=40) groups, achieving an overall accuracy of 94.5%. Key metrics highlight its robustness: sensitivity (90%)—calculated as TP / (TP + FN) = 36 / (36 + 4)—reflects its ability to correctly identify 90% of late-stage patients, critical for timely interventions; specificity (96%)—computed as TN / (TN + FP) = 120 / (120 + 5)—indicates a 96% success rate in identifying early-stage cases, minimizing unnecessary treatments; and precision (87.8%)—derived from TP / (TP + FP) = 36 / (36 + 5)—shows that 87.8% of predicted late-stage classifications are accurate. Additionally, the negative predictive value (NPV) of 96.8%—TN / (TN + FN) = 120 / (120 + 4)—underscores its reliability in ruling out advanced disease, while the F1-score (88.9%)—calculated as 2 × (Precision × Sensitivity) / (Precision + Sensitivity) = 2 × (0.878 × 0.90) / (0.878 + 0.90)—balances precision and recall effectively, affirming the model's strength despite the dataset's imbalance.

Compared to standalone models like Cox Regression (accuracy: 80.6%), Logistic Regression (67.88%), and SVM (81.8%), the hybrid model's superior performance stems from integrating Cox Regression's statistical rigor with SVM's ability to handle complex, non-linear relationships. The misclassification rate, computed as (FP + FN) / Total = (5 + 4) / 165 = 5.45%, is notably low, reinforcing its precision. These metrics collectively position the Cox-SVM model as a powerful tool for clinical decision-making, excelling in both identifying high-risk patients and confirming low-risk cases.

3.4.1 Classification of observations in the groups

In an analysis of patient data, observations were divided into two groups: Group 1, consisting of patients in an early stage of the disease with longer survival expectations, and Group 2, consisting of patients in a late stage of the disease with shorter survival expectations. Five misclassified observations were identified, where they belonged to Group 1 (early stage) but were classified into Group 2 (late stage). This result is consistent with what was found through classification table number 13, The observations were as follows:

Observation	Value (y)	Observation	Value (y)
13	0.07845	102	0.05231
19	0.04451	144	0.03670
87	0.01287		

As for the observations belonging to Group 2 that were incorrectly classified into Group 1, there were 4 such cases. This also aligns with classification table 17. The observations are as follows:

Observation	Value (y)	Observation	Value (y)
36	-0.07854	91	-0.06581
76	-0.02354	151	-0.05891

It is necessary to estimate the hyperplane equation, which is $y = W^{T}X + b$, and consequently obtain estimates for the weight vector (W) and the bias term (b), as shown in the following table.

Table 15: the weight vector (W) and the bias term (b)

	X_1	X_3	X_5	X_7	X_{10}	X_{13}	b
$\boldsymbol{W\ i}$	3.1570	1.0325	2.1135	2.7823	0.8790	1.97852	2.2157

The weights of Table 15 of the Cox-SVM hybrid model are valuable in understanding how individually they work towards classifying the lung cancer patients into early- and late-disease-stage groups. Smoking (X1) with a maximum weight of 3.1570 is discovered to be the most predictive variable, underpinning the established status it attains as one of the top risk factors responsible for lung cancer incidence and deaths. Following closely, Treatment Methods (X7) with a weight of 2.7823 determines the significant influence of treatment modalities on survival outcomes, in line with clinical experience that treatment effectiveness is cancer type- and stage-dependent. Age (X5) and Disease Stage (X13) with weights of 2.1135 and 1.97852 respectively reinforce the importance of patient age and disease progression, in line with the expectation that both rising age and advanced stage would be linked with poorer prognosis. These heavy weights together demonstrate the model's power to rank clinically significant factors within its classification.

By comparison, Occupation (X_3) and COVID-19 Infection (X_{10}), with weights of 1.0325 and 0.8790, respectively, have comparatively weaker contributions to the model. The modest weight for Occupation is consistent with a secondary role, perhaps through environmental exposures or socioeconomic status, but one whose effect is eclipsed by more proximal clinical factors. In the same manner, the lowest weight given to COVID-19 Infection suggests a smaller contribution to classification, either because its impact is overshadowed by other predictors such as age or disease stage, or there is not enough data to establish its complete impact. The bias term (b) 2.2157 positions the hyperplane for optimal class separation according to data distribution. This weighting plan illustrates the usefulness of the hybrid model in achieving a balance between the statistical rigor of Cox regression and SVM's ability to model complex, non-linear relationships, providing an efficient paradigm for survival analysis of lung cancer. Therefore, new observations are marked as follows:

Table 16: Classification of New Lung Cancer Patients Using the Cox-SVM Hybrid Model

Observation	X_1	X_3	X_5	X_7	X_{10}	X_{13}	Predict
166	1	2	80	4	1	3	Class 2
167	2	1	40	1	2	1	Class 1
168	1	3	67	3	1	2	Class 2

Table 16 demonstrates the classification of three new observations (166, 167, 168) using the Cox-SVM hybrid model, based on variables Smoking (X1), Occupation (X3), Age (X5), Treatment Methods (X7), COVID-19 Infection (X10), and Disease Stage (X13), employing the equation($y=w^T x_i+b$) with weights from Table 18 (b = 2.2157); patient 166 (80 years, smoker, stage III) is classified as Class 2 (late-stage) due to high-risk factors, patient 167 (40 years, non-smoker, stage I) as Class 1 (early-stage) for better survival prospects, and patient 168 (67 years, smoker, stage II) as Class 2 due to smoking and COVID-19 impact, highlighting the model's precision (90% sensitivity, 95.83% specificity) in distinguishing stages to enhance clinical decision-making.

3.5 Hybrid Model (logistic-SVM)

The hybrid model (Logistic-SVM) was constructed by incorporating all independent variables affecting the survival time of lung cancer patients, totaling (13) variables, using the Kaplan-Meier method. The variables of family genetic history (x2), gender (x_4), marital status (x_8), degree of kinship between parents (x9), chronic diseases (x_{11}), and place of residence (x_{12}) were excluded. Seven variables were retained: smoking (x_1), occupation (x_3), age (x_5), exposure to radiation therapy (x6), treatment methods (x_7), COVID-19 infection (x_{10}), and disease stage (x_{13}). These variables were then entered into a logistic regression model, which excluded exposure to radiation therapy (x_6) and COVID-19 infection (x_{10}), retaining the remaining five variables due to their statistical significance. These variables were subsequently fed into a Support Vector Machine (SVM) using the Radial Basis Function (RBF) kernel, with an error margin of e=0.1 and a margin size of c=1. The classification results based on the hybrid model are as follows:

Table 17: Classification Results Using Hybrid Model (logistic-SVM)

Actually	Group Size	Predicted Class 1	Predicted Class 2
Class 1	125	116	9
Class 2	40	6	34

The Logistic-SVM hybrid model, as presented in Table 17, demonstrates strong performance in classifying lung cancer patients into early-stage (Class 1) and late-stage (Class 2) groups, achieving an overall accuracy of 90.91%. This was calculated using the formula (TP + TN) / Total, where TP (true positives) = 34 (correctly classified Class 2 patients), TN (true negatives) = 116 (correctly classified Class 1 patients), and Total = 165, resulting in 150 / 165 ≈ 0.9091. Sensitivity, measuring the model's ability to detect Class 2, reached 85% via TP / (TP + FN) = 34 / (34 + 6), while specificity, assessing Class 1 accuracy, hit 92.8% through TN / (TN + FP) = 116 / (116 + 9). These metrics highlight the model's capacity to identify critical cases while minimizing errors, supported by a low misclassification rate of 9.09%, derived from (FP + FN) / Total = (9 + 6) / 165.

Additionally, the model boasts a high negative predictive value (95.08%), computed as TN / (TN + FN) = 116 / (116 + 6), indicating its reliability in ruling out late-stage disease. Precision for Class 2, reflecting the correctness of positive predictions, was 79.07% via TP / (TP + FP) = 34 / (34 + 9), while the F1-score for Class 2, balancing precision and sensitivity, reached 81.95% using 2 × (Precision × Sensitivity) / (Precision + Sensitivity) = 2 × (0.7907 × 0.85) / (0.7907 + 0.85). The macro-average F1-score (87.94%), calculated as the average of F1 for both classes (F1_Class1 = 0.9393 and F1_Class2 = 0.8195), confirms balanced performance.

The number of support vectors in the first group reached 18 vectors, while the number of support vectors in the second group was 11, resulting in a total of 29 support vectors for both groups combined.

3.5.1 Classification of observations in the groups

In patient data analysis, observations were categorized into two groups: Group 1, patients with an early stage of the disease and with greater survival expectations, and Group 2, patients with a late stage of the disease and with lesser survival expectations. Nine observation misclassifications were found, in which they were categorized under Group 1 (early stage) but classed as Group 2 (late stage). This outcome is the same as what was obtained via classification table number 20, The observations were as follows:

Observation	Value (y)	Observation	Value (y)
17	0.04578	131	0.02504
68	0.07856	139	0.07635
91	0.01239	147	0.08751
102	0.03879	159	0.08532
106	0.09634		

As for the observations belonging to Group 2 that were incorrectly classified into Group 1, there were 6 such cases. This also aligns with classification table 17. The observations are as follows:

Observation	Value (y)	Observation	Value (y)
43	-0.07635	107	-0.03772
70	-0.01996	145	-0.75201
95	-0.05642	161	-0.04361

It is necessary to estimate the hyperplane equation, which is $y = W^{T}X + b$, and consequently obtain estimates for the weight vector (W) and the bias term (b), as shown in the following table.

Table 18: the weight vector (W) and the bias term (b)

	X_1	X_3	X_5	X_7	X_{13}	b
$\boldsymbol{W\ i}$	3.5601	2.7614	0.8567	1.9624	1.2715	2.2927

Table 18 shows that the Logistic-SVM hybrid model ranks Smoking (X_1) and Occupation (X_3) as its top predictors with respective weights of 3.5601 and 2.7614 in contributing to its classification accuracy being high (90.91%). Treatment Methods (X_7), Disease Stage (X_{13}), and Age (X_5) are secondary factors, with a range of 0.8567 to 1.9624 for their respective weights, while the bias term (2.2927) moves the decision boundary. Its efficiency is indicated by the model's dependence on 29 support vectors, whereas the low weighting of Disease Stage and omission of variables such as X_6 and X_{10} indicate the potential for improvement. Statistically, this weighting scheme confirms the high performance of the model while indicating the possibility of additional improvement, i.e., re-establishing variable selection or kernel parameter optimization, to maximize its predictive value for lung cancer survival analysis. Thus, new observations are classified as follows:

Table 19: Classification of New Lung Cancer Patients Using the logistic-SVM Hybrid Model

Observation	X_1	X_3	X_5	X_7	X_{13}	Predict
166	1	3	80	3	3	Class 2
167	1	2	70	2	4	Class 2
168	2	1	50	1	1	Class 2
169	2	2	45	1	2	Class 2

The results show that all new observations were classified as Class 2, which may indicate a tendency of the model to predict late-stage disease, possibly due to the high weights assigned to

Smoking and Occupation or the selected variable values that tend to increase y_i. To enhance the model, reconsidering the exclusion of variables like X_6 (Radiation Therapy) and X_{10} (COVID-19 Infection) or adjusting the RBF kernel parameters could reduce potential over-sensitivity to certain variables. Overall, the model delivers strong performance (90.91% accuracy), but these findings underscore the need for further testing on diverse datasets to ensure balanced classification between the two classes.

3.6 Comparison Between Individual and Hybrid Models

Survival analysis in lung cancer patients involves measurement of time-to-event outcomes, disease staging classification, and prediction of prognosis, using both individual and hybrid statistical models. Individual models like Cox Regression, Logistic Regression, and Support Vector Machines (SVM) offer different strengths—statistical rigor, ease of use, and non-linear flexibility, respectively—but are at a disadvantage because they cannot cope with complex, heterogeneous data. Hybrid methods such as Cox-SVM and Logistic-SVM integrate these methods to enhance accuracy and resistance by merging the traditional survival analysis with machine learning capability. Comparison in this case relies on performance in respect to several measures to determine their efficacy in clinical use.

Table 20: Comparison Between Individual and Hybrid Models

Metric/Indicator	Cox Regression	Logistic Regression	Support Vector Machine (SVM)	Cox-SVM Hybrid	Logistic-SVM Hybrid
Overall Classification Accuracy	80.6%	67.88%	81.8%	94.5%	90.91%
Sensitivity (TruePositiveRate)	75%	65.57%	80%	90%	85%
Specificity (True Negative Rate)	85.88%	69.23%	82.4%	96%	92.8%
Positive Predictive Value (Precision)	N/A	55.56%	59.3%	87.8%	79.07%
Negative Predictive Value (NPV)	N/A	N/A	N/A	96.8%	95.08%
F1-Score (Harmonic Mean of Precision & Recall)	N/A	60.17%	68.1%	88.9%	81.95%
Misclassification Rate	19.39%	N/A	18.2%	5.45%	9.09%
R-squared (Variance Explained)	61.4%	54.18%	N/A	N/A	N/A

Table 20 shows the Cox-SVM hybrid model as the outright winner in lung cancer patient classification with an overall accuracy of 94.5%, sensitivity of 90%, specificity of 96%, and F1-score of 88.9%, with a misclassification rate of just 5.45%. This beats single models—Cox Regression (80.6% accurate, 61.4% R^2), Logistic Regression (67.88% accurate, 54.18% R^2), and SVM (81.8% accurate)—and the Logistic-SVM hybrid model (90.91% accurate, 85% sensitivity, 92.8% specificity) by emphasizing the statistical synergy of combining Cox's hazard-based understanding with SVM's ability to perform non-linear classification. Its high accuracy (87.8%) and negative predictive value (96.8%) emphasize its application in the diagnosis of advanced disease and ruling out advanced disease confidently, making it highly valuable in clinical practice.

Conversely, Table 20 exposes weaknesses in standalone models: Logistic Regression's sensitivity (65.57%) and precision (55.56%) falter due to its binary framework, while SVM's precision (59.3%) and misclassification rate (18.2%) indicate challenges with class imbalance (125 early-stage vs. 40 late-stage). Cox Regression, though explaining 61.4% of survival variance, lacks hybrid-level classification precision. The Logistic-SVM hybrid, with an F1-score of 81.95% and a 9.09% misclassification rate, performs admirably but trails Cox-SVM, possibly due to its reduced predictor set (five vs. six). These findings advocate for hybrid models in complex survival analysis, though cross-validation and broader dataset testing would enhance confidence in their generalizability.

4. Discussion

The comparative study of individual and hybrid survival models for lung cancer patients highlights the potency of revolutionizing traditional statistical approaches with machine learning models. The Cox-SVM hybrid model was the best, with a classification accuracy of 94.5%, sensitivity of 90%, and specificity of 96%, much better compared to isolated models such as Cox Regression (80.6%), Logistic Regression (67.88%), and SVM (81.8%). This model's superiority will be because of its capacity to take advantage of Cox Regression's hazard-based statistical paradigm on variable selection with SVM's versatility in modeling any type of very complex, not linear relationship capable of overcoming otherwise inherent constraints such as the proportional hazards assumption in the Cox models and linear constraints imposed on Logistic Regression. The model's high negative predictive value (96.8%) and low misclassification rate (5.45%) also emphasize its clinical usefulness in identifying early-stage patients correctly, minimizing unnecessary interventions, and marking late-stage cases for timely escalation of treatment.

The Logistic-SVM hybrid, though strong with a 90.91% accuracy, came in second behind Cox-SVM, most probably because it drew on fewer predictors (five vs. six) and did not include variables such as COVID-19 infection (X10) that were highly influential within the Cox-SVM model. This indicates that the selection of variables and model formation significantly affects performance, with Cox-SVM benefiting from a more extensive, statistically proven predictor base. Single-model approaches, although less complex, showed trade-offs: Cox Regression was very good at explaining variance (61.4% R^2) but poor in classification accuracy, whereas SVM performed poorly in precision (59.3%) due to class imbalance. These results are consistent with previous research, e.g., Xiao et al. (2022), which reported machine learning's advantage in discriminative capacity, and Cuthbert et al. (2022), which identified limited additional value in sophisticated methods unless combined thoughtfully, as done here.

Clinically, the Cox-SVM model's weighting on smoking (weight: 3.1570), treatment patterns (2.7823), and disease stage (1.97852) are evidence-based risk factors, which make it more interpretable and useful. However, the lower weighting of COVID-19 infection (0.8790) in Cox-SVM and exclusion in Logistic-SVM warrant further study, given that it can be a confounding variable for lung cancer prognosis, especially post-2020. More generally, these results validate hybrid approaches to survival analysis, offering the best possible balance of predictive ability and practical application in personalized medicine.

5. Limitations

The study involved 165 lung cancer patients from two centers in Egypt (2020–2024), with a sample comprising 125 early-stage and 40 late-stage cases. It utilized hybrid models, such as SVM with RBF kernel tuning, Kaplan-Meier and backward elimination methods for variable selection, reflecting a rigorous statistical approach. However, future studies are needed to strengthen the findings and expand their applicability.

6. Conclusion

This study demonstrates that hybrid models, and more precisely the Cox-SVM model, are a significant enhancement in survival analysis of lung cancer patients with greater classification accuracy (94.5%), sensitivity (90%), and specificity (96%) compared to single models like Cox Regression (80.6%), Logistic Regression (67.88%), and SVM (81.8%). By pairing the statistical strengths of traditional survival methods with the flexibility of machine learning, these hybrids effectively capture complex, non-linear relationships within survival data to enhance predictive accuracy and clinical decision-making. Both the Cox-SVM model's emphasis on smoking, treatment types, and disease stage adhering to traditional prognostic variables and its low misclassification rate (5.45%) reinforce its clinical viability and reliability.

Table 21: Improvement Percentages of Hybrid Models Compared to Individual Models

Metric/Individual Model	Cox Regression	Logistic Regression	SVM
Overall Classification Accuracy			
Cox-SVM(94.5%)	17.25%	39.23%	15.53%
Logistic-SVM(90.91%)	12.79%	33.95%	11.14%
Sensitivity (True Positive Rate(			
Cox-SVM(90%)	20%	37.36%	12.5%
Logistic-SVM(85%)	13.33%	29.66%	6.25%
Specificity (True Negative Rate(			
Cox-SVM(96%)	11.7%	38.53%	16.5%
Logistic-SVM(92.8%)	8.02%	34.06%	12.62%
Misclassification Rate			
Cox-SVM(5.45%)	-71.88%	Not Available	-70.05%
Logistic-SVM(9.09%)	-53.12%	Not Available	-50.05%

Table 21 highlights the substantial improvements offered by hybrid models over individual models in lung cancer survival analysis, with Cox-SVM demonstrating the highest gains—up to 39.23% in classification accuracy over Logistic Regression, 20.00% in sensitivity over Cox Regression, and a 71.88% reduction in misclassification rate compared to Cox Regression—underscoring its superior integration of statistical rigor and machine learning flexibility, while Logistic-SVM also shows notable enhancements, such as a 34.06% increase in specificity over Logistic Regression, though it lags slightly behind Cox-SVM due to its more limited predictor set.

7. Recommendations for Future Research

- Expand and Diversify the Sample: Increase the number of participants and diversify samples geographically and demographically to enhance the generalizability of hybrid models and reduce bias.
- Incorporate Variables and Optimize SVM: Include additional biological and clinical variables and optimize SVM parameters to improve predictive accuracy and computational efficiency.
- Develop Advanced Hybrid Models: Integrate deep learning with survival analysis to capture complex features and advance precision medicine.

References

Al-Essa, L. A., Soliman, A. A., Abd-Elmougod, G. A., & Alshanbari, H. M. (2023). A comparative study with applications for Gompertz models under competing risks and generalized *hybrid censoring schemes. Axioms*, 12(4), 322. https://doi.org/10.3390/axioms12040322

Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36(2), 3302–3308. https://doi.org/10.1016/j.eswa.2008.01.005

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92) (pp. 144–152). Association for Computing Machinery. https://doi.org/10.1145/130385.130401

Candès, E., Lei, L., & Ren, Z. (2023). Conformalized survival analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology, 85(1), 24–45. https://doi.org/10.1093/jrsssb/qkac013

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Cuthbert, A. R., Giles, L. C., Glonek, G., Kalisch Ellett, L. M., & Pratt, N. L. (2022). A comparison of survival models for prediction of eight-year revision risk following total knee and hip arthroplasty. *BMC Medical Research Methodology*, 22(1), 164. https://doi.org/10.1186/s12874-022-01647-6

Edwards, C. (2003). Assessing association: Logistic regression and logit analysis. Biometry, FRWS 6500, Fall 2003. [No DOI available; unpublished course material]

Garson, D. (2006). Logistic regression. Retrieved from http://www2.chass.ncsu.edu/garson/pa765/logistic.htm [No DOI available; online resource]

Germer, S., Rudolph, C., Labohm, L., Katalinic, A., Rath, N., Rausch, K., Holleczek, B., & Handels, H. (2024). Survival analysis for lung cancer patients: A comparison of Cox regression and machine learning models. *International Journal of Medical Informatics*, 191, 105607. https://doi.org/10.1016/j.ijmedinf.2024.105607

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. [No DOI *available for book; ISBN*: 978-0-470-58247-3]

Huang, Y.-M., et al. (2002). Support vector machines for classification and regression: A comparative study. *Journal of Machine Learning Research*, 3, 123–145. [Note: DOI not specified in original; retrieve from https://jmlr.org if available]

Kaplan, E., & Meier, P. (1958). Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Langley, P., & Carbonell, J. G. (1984). Approaches to machine learning. *Journal of the American Society for Information Science*, 35(5), 306–316. https://doi.org/10.1002/asi.4630350510

Lee, S. W. (2023). Kaplan-Meier and Cox proportional hazards regression in survival analysis: Statistical standard and guideline of Life Cycle Committee. Life Cycle, 3, e8. https://doi.org/10.54724/lc.2023.e8

Lu, Y., Aslani, S., Zhao, A., Shahin, A., Barber, D., Emberton, M., Alexander, D. C., & Jacob, J. (2023). A hybrid CNN-RNN approach for survival analysis in a lung cancer screening study. *Heliyon*, 9, e18695. https://doi.org/10.1016/j.heliyon.2023.e18695

Mandev, A. R., Muralimohan, P., Reddy, H., & Mathur, R. (2024). Hybrid survival analysis model for predicting automotive component failures. Automotive Technical Papers, 2024-01-5078. https://doi.org/10.4271/2024-01-5078

Salem, S. A., Abo-Kasem, O. E., & Khairy, A. A. (2024). Inference for generalized progressive hybrid type-II censored Weibull lifetimes under competing risk data. *Computational Journal of Mathematical and Statistical Sciences*, 3(1), 177–202. https://doi.org/10.21608/CJMSS.2024.256760.1035

Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., He, Y., & Zheng, Y. (2022). The application and comparison of machine learning models for the prediction of breast cancer prognosis: Retrospective cohort study. JMIR Medical Informatics, 10(2), e33440. https://doi.org/10.2196/33440

Xu, L., Cai, L., Zhu, Z., & Chen, G. (2023). Comparison of the Cox regression to machine learning in predicting the survival of anaplastic thyroid carcinoma. BMC Endocrine Disorders, 23(1), 129. https://doi.org/10.1186/s12902-023-01368-5

Yakubu, O. M., Mohammed, Y. A., & Imam, A. (2022). A mixture of gamma-gamma, and loglogistic-gamma distributions for the analysis of heterogeneous survival data. *International Journal of Mathematical Research*, 11(1), 1–9. https://doi.org/10.18488/24.v11i1.2924

المستخلص

تتناول الدراسة مقارنة أداء النماذج الهجينة لتحليل البقاء على قيد الحياة، والتي تجمع بين التقنيات الإحصائية التقليدية وتعليم الآلة، للتنبؤ بنتائج البقاء لدى مرضى سرطان الرئة. شملت الدراسة 165 مريضًا من معهد طنطا للأورام ومستشفى كفر الشيخ للصدر بين عامي 2020 و2024، حيث تمت مقارنة النماذج الفردية – انحدار كوكس، الانحدار اللوجستي، وآلات المتجهات الداعمة (SVM) – مع نموذجين هجينين هما كوكس–SVM ولوجستي–SVM. تم تحديد المؤشرات عالية المخاطر مثل التدخين، المهنة، العمر، أنواع العلاج، الإصابة بفيروس كورونا، ومرحلة المرض، وتم تصميمها باستخدام تحليل كابلان–مايير لاختيار السمات. أظهرت النتائج أن النموذج الهجين كوكس–SVM حقق أفضل أداء بين جميع النماذج بدقة تصنيف بلغت 94.5%، وحساسية 90%، ونوعية 96%، ومعدل خطأ في التصنيف 5.45%. تبعه النموذج الهجين لوجستي–SVM بدقة 90.91%، ثم النماذج الفردية (كوكس: 80.6%، لوجستي: 67.88%، SVM: 81.8%). يُعزى أداء النماذج الهجينة إلى دمج نموذج كوكس القائم على المخاطر مع قدرة SVM على التعامل مع العلاقات غير الخطية، مما يوفر دقة تنبؤية وأهمية سريرية أفضل. على الرغم من قيود حجم العينة الصغير واستبعاد بعض المتغيرات، تشير النتائج إلى أن الأساليب الهجينة في تحليل البقاء تمثل أداة قيمة في الطب الشخصي. يجب التحقق من هذه النماذج باستخدام بيانات أكبر وأكثر تنوعًا وعوامل تنبؤية أخرى في الأبحاث المستقبلية.

الكلمات المفتاحية: تحليل البقاء؛ النماذج الهجينة؛ كابلان–مايير، تعليم الآلة؛انحدار كوكس.