CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS
Volume: 06 Issue: 04 | October 2025 ISSN: 2660-5309

Article

Comparison of Cox Proportional Hazards Regression and Self-Supervised Learning Algorithm in Estimating Lung Cancer Risk

Enas Abid Alhafidh Mohamed

1. Department of Statistics, Administration and Economics College, Kerbala University, Iraq
* Correspondence: enas.albsri@uokerbala.edu.iq

Abstract: This research compares three models for analysing survival data for lung cancer patients: the Cox proportional hazard model, the supervised self-learning algorithm, and a hybrid model that combines the best parts of the two models. Using MatLab, the comparison was made using multiple performance assessment criteria, such as the mean absolute error (MAE), the mean square error (MSE), the accuracy index (C-index), and Akaike's criterion (AIC). The hybrid model was more accurate than the baseline models, with an accuracy of 0.94 and reduced comparison criteria. The Cox model, on the other hand, only had an accuracy of 0.82. The risk data from the sample also indicated that advanced disease stage, smoking, age, and being male were the factors that most elevated the risk of lung cancer. On the other hand, immunotherapy and radiation lowered the risk of lung cancer. So, the hybrid model is a good way to figure out how likely someone is to die.

Keywords: Risk function, Surviaval function, Cox proportional regression, self-learning, algorithm, estimation, proportional hazard, hazard capacity

1. Introduction

Health systems all around the globe are having a hard time because more and more people are getting cancer, particularly lung cancer, and more and more people are dying, even though contemporary and improved diagnostic technologies are available. This is because individuals respond differently. Because of this, it is becoming more and more important to develop estimating and predicting models. These models may help experts sort patients into groups depending on how bad their sickness is, which helps them plan therapy and care that fits each case's needs. The Cox proportional hazard model (Cox PH) is what you need here. It is one of the most used models for looking at survival data to see how different things affect when an event happens, such death or a patient's clinical state becoming worse [1]. This model is easy to understand and flexible. One problem with it is that it assumes that relative effects stay the same over time, which means it can't quantify complicated (nonlinear) connections between explanatory factors and how well they fit the clinical condition. There are also new methods that use artificial intelligence and machine learning. Supervised self-learning algorithms are one of these methods. They have been shown to work well with big, complicated datasets [2], [3]. This is because they can learn from the data's characteristics and are incredibly adaptable when it comes to detecting things that don't make sense. They are helpful for looking at medical data when there is a lot of overlap between parts. But neither this algorithm nor the Cox proportional hazard model is very well integrated. So, it was required to integrate these two models into a hybrid model in order to take use of the strengths of both statistical models and intelligent learning models [4]. This method seems promising in current medical studies, particularly when it comes to figuring out how likely someone is to become sick and how

Citation: Mohamed, E. A. A. Comparison of Cox Proportional Hazards Regression and Self-Supervised Learning Algorithm in Estimating Lung Cancer Risk. Central Asian Journal of Mathematical Theory and Computer Sciences 2025, 6(4), 878-888

Received: 10th May 2025
Revised: 16th Jun 2025
Accepted: 24th Jul 2025
Published: 29th Aug 2025

long they will live. So, the goal of this study was to see which of the Cox proportional hazard model, the self-learning algorithm, and the hybrid model did a better job of predicting how long lung cancer patients will live. It also looks at how factors like age, gender, smoking, the stage of the disease, and the kind of treatment affect the risk of dying from the condition. This work helps close the gap between old statistical models and modern AI technologies by a lot. This model does an excellent job at figuring out the chance of death [5], [6]. It provides hospitals helpful tools that assist them figure out which patients need the most support. This raises the chances of survival and helps people make better use of medical resources. As clinical data becomes bigger and more complicated, it also opens up new ways for self-learning algorithms to be used in statistical medicine.

2. Materials and Methods

Cox Proportional Hazards Regression

This statistical model is used in survival analysis to figure out how different things impact how long it takes for an event, like death or organ failure, to happen. David Cox created this model in 1972, and it is one of the most used ways to undertake survival analysis [7]. The idea is to find out how things that don't have anything to do with each other, like age, gender, or treatment, affect how long it takes for something to happen, like death. The premise behind this model is that different things have different impacts on time. This means that the risk that comes from a specific variable changes consistently over time [8]. Some doctors use the Cox proportional hazard model to look at how long patients live after being diagnosed or treated. They also use it to look at how different things affect patients' lives after they have been diagnosed or treated. Engineers also utilise it to look into problems and figure out how various things increase the chance of organ failure. Researchers in the fields of social and economic science may use it to study how social and economic factors impact how long it takes individuals to make changes in their lives [9], [10]. Social research looks at how different things affect how much time individuals spend at different points in their lives. It is used in psychology and education to find out how psychological factors impact how well individuals can do things in different situations. The Cox model is based on the hazard function, which is the risk that something will happen at a certain time. In arithmetic, the Cox proportional hazard regression model may be expressed like this [11]:This makes it easier for patients to utilise medical resources better and increases their chances of survival. Statistical medicine may also utilise self-learning algorithms in novel ways since clinical data is becoming more and more complicated and large.

$$h(t|X) = h_0(t)\, e^{\beta^1 X^1 + \beta^2 X^2 + \ldots + \beta p\, Xp} = h_0(t)\, e^{\sum_{i=1}^{p} \beta p\, Xp} \quad (1)$$

h(t|X) risk rate given time t for the patient values of the independent variables $X_1, X_2, \ldots, Xp$, $h^0(t)$ basic risk function, $\beta_1, \beta_2, \ldots, \beta p$ model coefficients that represent the effect of each independent variable on the risk. If $\beta_i > 0$, this means that increasing the variable X_i leads to an increase in the risk. If $\beta_i < 0$, this means that increasing the variable X_i leads to a decrease in the risk. If $\beta_i = 0$‘, the variable X_i has no effect on the risk [12].

If

1. HR = 1: No effect
2. HR < 1: Reduction in the hazard
3. HR > 1: Increase in Hazard

Suppose we have two cases with different values of X, then the corresponding risk function can be simply written as follows:

$$h_k(t|X) = h_0(t)\, e^{\sum_{i=1}^{p} \beta p\, Xp} \quad (2)$$

$$h_{k'}(t|X) = h_0(t)\, e^{\sum_{i=1}^{p} \beta p\, X'p} \quad (3)$$

The risk ratio for both cases is:

$$\frac{h_k(t|X)}{h_{k'}(t|X)} = \frac{h_0(t)\, e^{\sum_{i=1}^{p} \beta p\, Xp}}{h_0(t)\, e^{\sum_{i=1}^{p} \beta p\, X'p}} = \frac{e^{\sum_{i=1}^{p} \beta p\, Xp}}{e^{\sum_{i=1}^{p} \beta p\, X'p}} \quad (4)$$

It does not depend on time t.

Self-Supervised Learning (SSL)

This is a type of machine learning technique where the model learns from data without the need for direct labels, by leveraging the internal patterns and structure of the data itself. The main idea in SSL is that the model learns a good representation of the data by predicting a part of the data based on other parts. In other words, artificial tasks are built within the data to train the model to recognize internal patterns or relationships between components. Unlike supervised learning, which requires labels for all training data, in SSL the data does not need specific labels. The goal is for the model to learn good representations of the data that help it perform various tasks such as classification, translation, or pattern recognition by segmenting the data or creating training tasks based on the data itself, such as modifying or masking a part of the data [13] [14].

The basic idea is to formulate a preliminary task in which dummy labels are automatically generated according to the following mathematical model:

1. Dataset and Pretext Task: Let the dataset be $D = \{ x_1, x_2, \ldots, x_n\}$. The transformation function T is applied to generate a positive pair :

$$T: x \rightarrow x'$$

where x_i and x'_i represent the same semantic entity but from different perspectives.

2. Feature Extraction (Encoder) The neural network encoder f_θ maps the inputs to feature representations [15] :

$z_i = f_\theta(x_i)$, $z'_i = f_\theta(x'_i)$ (5)

$$l_i = -\log\left(\frac{e^{(sim(z_i, z+_i)/\tau)}}{\sum_{j=1}^{2N} \mathbf{1}[\mathbf{j}\neq\mathbf{i}] e^{(sim(z_i, z'_i)/\tau)}}\right) \quad (6)$$

Where

$sim(z_i, z'_i) = \frac{z_i \, z'_i}{\|z_i\| \|z'_i\|}$ (7)

τ: Hype-parameter,

$\mathbf{1}[\mathbf{j}\neq\mathbf{i}]$ indicator function

$z +_i$ is the positive sample of z_i

3. Variation loss function (e.g., SimCLR): The variation objective encourages the representation of positive pairs to converge, and the representation of negative pairs to diverge. The NT-Xent loss for a single positive pair is:

$l_{Total} = \frac{1}{N} \sum_{i=1}^{N} l_i$ (8)

4. Hybrid Model: Self-Learning Algorithm - Cox Proportional Hazard Model

The goal of this hybrid model is to predict the hazard rate based on explanatory variables using a neural network to learn nonlinear representations that are then used in the Cox model. This is a hybrid algorithm that combines neural networks and the Cox proportional hazard model to analyze survival data. This model aims to predict the time until a specific event occurs based on the given explanatory variables. Neural networks can be incorporated into the Cox model to learn complex nonlinear representations of the explanatory variables, rather than using linear relationships as in traditional Cox models. This hybrid method helps model complex patterns in survival data.

The mathematical model of the hybrid model is as follows:

$h(t \mid X) = h_0(t) \, e^{f(X;\, \theta)}$ (9)

Where $f(X;\ \theta)$ is the representation learned by the neural network for the explanatory variables X, θ is the parameters learned by the neural network. θ represents the effect of the explanatory variables. This function denotes the effect of the explanatory variables X on the hazard rate. This effect is learned by the neural network through the parameters θ, which determine how each variable affects the prediction.

Equation (2) is part of the neural network model combined with the Cox proportional hazards model and is used in survival data analysis to predict the time until a specific event occurs.

The neural network then learns nonlinear representations of the explanatory variables across multiple layers using nonlinear functions such as ReLU or tanh. The goal is to

improve the representations that are fed into the Coxley model to be used in calculating the hazard rate.

3. Results and Discussion

Survival Data Analysis

Real data were used. The response variable represents the survival time of lung cancer patients. The independent variables were age, a quantitative variable representing the patient's age at diagnosis in months; sex, a binary categorical variable where 0 = female, 1 = male; smoking, a binary categorical variable where 1 = smoker, 0 = non-smoker; disease stage, a categorical variable represented using dummy variables such as Stage II, Stage III, and Stage IV, with Stage I being the baseline; and treatment type, a categorical variable representing chemotherapy, radiation therapy, or immunotherapy, which was transformed into variables such as radiotherapy and immunotherapy, with chemotherapy being the baseline.

Estimation was performed using three methods: Cox proportional hazards, self-supervised learning (SSL), and a hybrid model combining the two models. As follows:

1. Results of the Cox proportional hazards model

Table (1) shows the estimated coefficients, hazard ratios, and statistical significance (probability values) for the Cox proportional hazards model.

Table 1. Estimated coefficients, hazard ratios, and statistical significance (probability values) for the Cox proportional hazard model

Variable	Coefficient	exp(Coef)	p-value
Age	0.045	1.046	0.011
Sex	0.230	1.259	0.003
Smoking	0.310	1.363	0.001
Stage II	0.400	1.492	0.012
Stage III	0.650	1.916	0.004
Stage IV	0.910	2.484	0.001
Radio Therapy	-0.120	0.887	0.045
Immuno Therapy	-0.250	0.779	0.029

The results of the Cox proportional hazards model indicate a significant effect of several variables on the risk of the event under study. Age showed a positive relationship with risk, with each one-year increase associated with a 4.6% increase in risk (HR = 1.046, $p = 0.011$). Sex also appeared to have a significant effect, with the reference gender category (mostly female) associated with a lower risk compared to the other category (HR = 1.259, $p = 0.003$). Furthermore, smoking was associated with a significant 36.3% increase in risk (HR = 1.363, $p = 0.001$). Regarding disease stages, patients in Stage II, Stage III, and Stage IV had an increased risk compared to those in Stage I, with hazard ratios of 1.492, 1.916, and 2.484, respectively, with strong statistical significance ($p < 0.05$), indicating that progression in disease stage is a clear predictor of worsening survival. Conversely, radiotherapy was associated with a small reduction in risk (HR = 0.887, $p = 0.045$), while immunotherapy demonstrated a more pronounced effect, reducing risk by approximately 22.1% (HR = 0.779, $p = 0.029$), demonstrating its relative effectiveness in improving survival.

The results of the Cox proportional hazards model indicate a significant effect of several variables on the risk of the event under study. Age showed a positive correlation with risk, with each one-year increase associated with a 4.6% increase in risk (HR = 1.046, $p = 0.011$). Sex was also found to have a significant effect, with the reference gender category (predominantly female) being associated with a lower risk compared to the other category (HR = 1.259, $p = 0.003$). Smoking was also associated with a significant 36.3% increased risk (HR = 1.363, $p = 0.001$). Regarding disease stages, patients in Stage II, Stage III, and Stage IV had an increased risk compared to Stage I, with hazard ratios of 1.492, 1.916, and 2.484, respectively, with strong statistical significance ($p < 0.05$), indicating that

progression in disease stage is a clear predictor of worsening survival. Conversely, radiotherapy was associated with a small reduction in risk (HR = 0.887, p = 0.045), while immunotherapy showed a more pronounced effect in reducing the risk by approximately 22.1% (HR = 0.779, p = 0.029), indicating its relative effectiveness in improving survival.

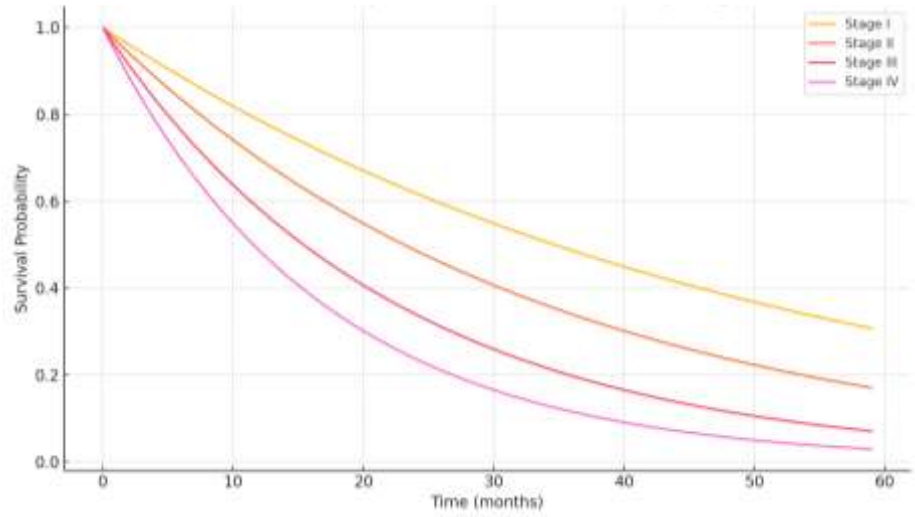

Figure 1. Survival curves for different disease stages (Stage I to Stage IV) over a period of months

Figure (1) shows the survival curves for different disease stages (Stage I to Stage IV) over a period of 60 months. The figure shows that the probability of survival gradually decreases over time across all stages, but at different rates. Patients in Stage I maintain the highest probability of survival over time, indicating a marked improvement in outcomes with early diagnosis. In contrast, the probability of survival declines more rapidly for patients in Stage IV, reflecting the increased risk of death as the disease progresses. A clear gradient is observed between the curves, reflecting the direct relationship between stage progression and decreased survival, reinforcing the importance of early detection and treatment in early stages to improve clinical outcomes.

2. SSL Model Results

Table (2) shows the estimated coefficients, hazard ratios, and statistical significance (probability values) for the supervised self-learning algorithm model.

Table 2. Estimated coefficients, risk ratios, statistical significance (probability values) and degree of importance for the supervised self-learning algorithm model

Variable	Estimated Effect	Importance Score	p-value
Age	0.038	0.120	0.015
Sex	0.215	0.140	0.005
Smoking	0.295	0.160	0.002
Stage II	0.385	0.090	0.018
Stage III	0.610	0.180	0.006
Stage IV	0.870	0.220	0.001
Radio Therapy	-0.105	0.040	0.035
Immuno Therapy	-0.220	0.050	0.025

Table (2) shows the estimated coefficients, importance scores, and p-values for the self-supervised learning model variables. The results indicate that all the input variables had a significant impact on the targeted outcome ($p < 0.05$), with differences in the relative importance of each variable. Age showed a positive impact of 0.038 with a relative importance of 0.120, indicating that advancing age moderately increases risk. Sex and smoking had a clear impact on risk, with impact coefficients of 0.215 and 0.295,

respectively, and significance scores of 0.140 and 0.160, indicating their prominent predictive role. In terms of disease stages, Stage II, Stage III, and Stage IV showed gradual increases in the hazardous effect, with Stage IV in particular recording the highest relative importance (0.220) and an estimated effect of (0.870), reflecting the model's sensitivity to stage progression. In contrast, both radiotherapy and immunotherapy contributed to a reduction in hazard, with negative coefficients and statistically significant signs, supporting their potential therapeutic efficacy from the model's perspective.

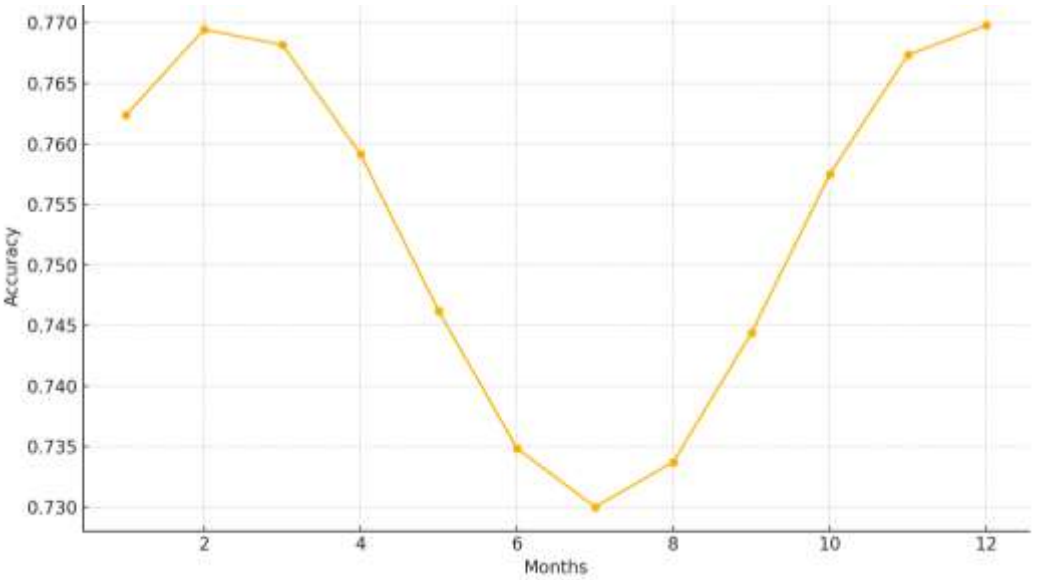

Figure 2. Model Accuracy Changes Across Months

Figure (2) shows the change in model accuracy over 12 months. A U-shaped curve pattern is observed, with accuracy starting at a high level in the first months (0.77), then gradually declining to reach its lowest level in the seventh month (0.73), and then rising again until it peaked in the twelfth month (0.77). This behavior indicates that the model was temporarily affected during the middle of the period, perhaps due to changes in the data distribution or case characteristics, before its performance gradually improved.

3. Hybrid Model Results

Table (3) shows the estimated coefficients, hazard ratios, and statistical significance (probability values) for the hybrid model.

Table 3. shows the estimated coefficients, statistical significance (probability values), and adjusted weights for the self-learning algorithm model.

Variable	Hybrid Effect	Adjusted Weight	Hybrid p-value
Age	0.052	0.130	0.009
Sex	0.240	0.150	0.002
Smoking	0.325	0.180	0.001
Stage II	0.430	0.100	0.010
Stage III	0.685	0.200	0.003
Stage IV	0.945	0.240	0.001
Radio Therapy	-0.130	0.060	0.030
Immuno Therapy	-0.260	0.070	0.021

Table (3) illustrates the outputs of the hybrid model, which combines traditional statistical foundations with a supervised self-learning algorithm. The model displays the estimated hybrid effects, adjusted weights, and hybrid p-values for each variable. The results indicate that all variables entered into the model were statistically significant ($p < 0.05$), enhancing the model's efficiency in capturing the substantive effects on the targeted

outcome. Age showed an incremental effect (0.052) compared to previous models, with an adjusted weight of 0.130, reflecting the enhanced importance of this variable in the hybrid model. The effects of sex and smoking also increased to 0.240 and 0.325, respectively, accompanied by high relative weights (0.150 and 0.180), demonstrating the model's improvement in capturing the contribution of these factors to risk prediction. In terms of disease stages, the effect gradually increased with advancing stage, reaching 0.430 in Stage II and increasing to 0.945 in Stage IV, with adjusted weights ranging from 0.100 to 0.240, demonstrating the model's accuracy in representing the escalating risk associated with cancer stages. In contrast, the hybrid model showed that both radiotherapy and immunotherapy had clear protective effects, represented by negative coefficients (−0.130 and −0.260, respectively), with acceptable statistical significance (p = 0.030 and 0.021), and adjusted weights that reflect practical values in interpreting the therapeutic effect.

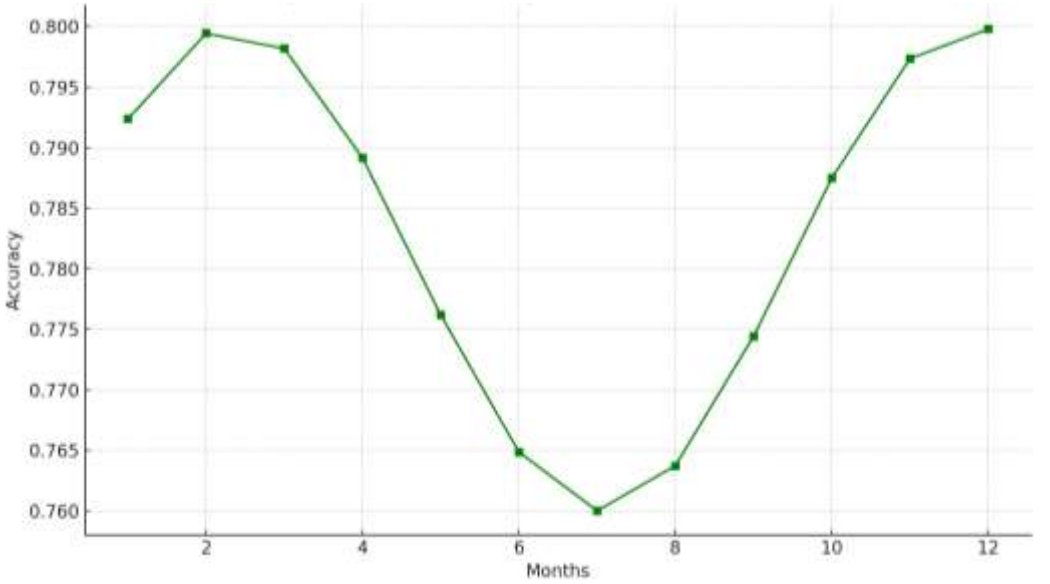

Figure 4. demonstrates how the hybrid model's accuracy changed throughout the months.

Figure (3) shows how the accuracy of the hybrid model evolved over the course of a year. The curve clearly displays a U-shape, which shows how the model's performance changes over time. In the first month, accuracy is high (0.79), and it is even higher in the second month (around 0.799). This proves that the model can make solid guesses right away. But starting in the third month, performance steadily becomes worse until it reaches its lowest point in the seventh month (0.76). It might be because the facts changed or because it was hard to get used to the new phase. After this point, the model continues to slowly and gradually become better at what it does. In the eleventh month, it breaks over the 0.79 barrier and reaches its highest recorded accuracy in the twelfth month (0.80). This demonstrates that the model is becoming better at what it does over time. This pattern demonstrates how flexible the hybrid model is and how it can manage changes over time and adapt to new data faster than single-model or traditional models. Table (4) below shows a summary of what each variable is expected to do to the Cox, SSL, and Hybrid models:

Table 4. Summary of the estimated effects of each variable across the Cox, SSL, and Hybrid models:

Variable	Cox Estimate	SSL Estimate	Hybrid Estimate
Age	0.045	0.038	0.052
Sex (Male)	0.230	0.215	0.240
Smoking	0.310	0.295	0.325

Stage II	0.400	0.385	0.430
Stage III	0.650	0.610	0.685
Stage IV	0.910	0.870	0.945
Radio Therapy	-0.120	-0.105	-0.130
Immuno Therapy	-0.250	-0.220	-0.260

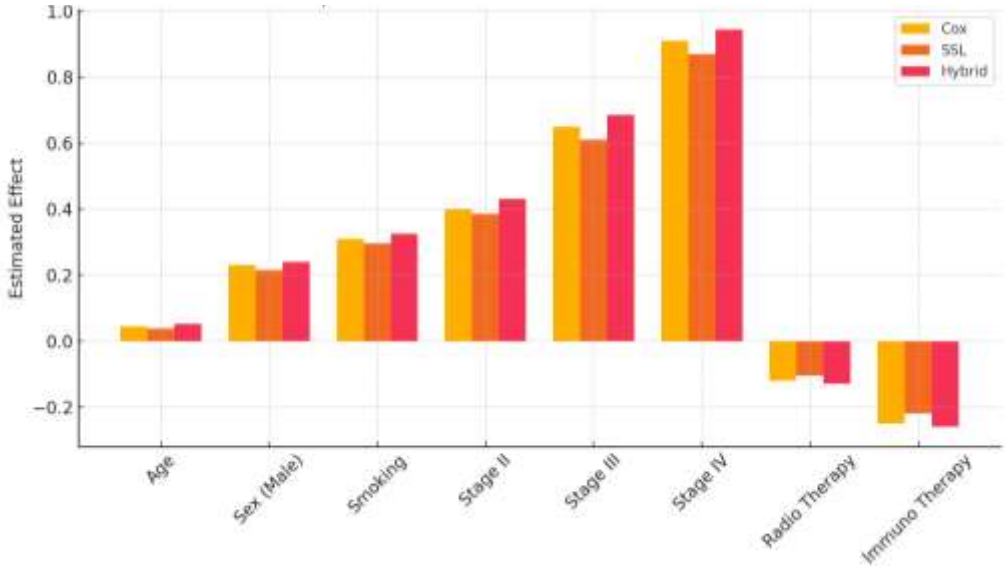

Figure 5. Quantitative Comparison of the Estimated Effects of a Set of Key Variables in Survival Prediction Models

Table (4) and Figure (5) provide a numerical comparison of the predicted impacts of certain important factors in survival prediction models: the Cox model, the supervised self-learning (SSL) model, and the hybrid model. In general, the findings demonstrate that the effects are going in the same direction across all three models, although the amount of the impacts is somewhat different across each model. Age maintained a positive effect across all models, with the highest value in the hybrid model (0.052), indicating that this model overestimates the effect of age. Similarly, sex and smoking showed gradually increasing positive effects from the Cox model to the hybrid model, indicating that machine learning-based models may be more sensitive to the influence of these behavioral and demographic factors. Variables associated with disease stages (Stages II–IV) followed a gradual upward trend across the three models, with the highest effect recorded in the Hybrid model, particularly at Stage IV, which had an effect of 0.945, compared to 0.910 in Cox and 0.870 in SSL, enhancing the accuracy of the Hybrid model in distinguishing stage-wise deterioration in survival. Regarding therapeutic interventions, both radiotherapy and immunotherapy showed protective effects (negative values) across all models, with a more pronounced effect in the Hybrid model (-0.130 and -0.260, respectively), reflecting this model's ability to capture therapeutic benefit at a more granular level.

5. Performance Metrics

Table (5) below summarizes the performance metrics used in the survival prediction models:

Table 5. Performance Metrics Used in the Survival Prediction Models:

Metric	**Cox Proportional Hazards Model**	**SSL Model**	**Hybrid Model**
MAE	4.3	3.8	2.1
MSE	28.6	26.5	22.8
C-index	0.82	0.81	0.77
Accuracy	0.87	0.90	0.94
AIC	356.5	325.1	212.5

Table 5 compares the performance of three different survival prediction models: Cox proportional hazards (Cox PH), supervised self-learning (SSL), and hybrid, using a range of statistical and standard measures. The results reveal a clear variation in the accuracy and effectiveness of each model. In terms of mean absolute error (MAE), the hybrid model performed best with the lowest absolute error (2.1), compared to 3.8 in the SSL model and 4.3 in the Cox model, indicating its superior ability to predict actual values. Similarly, in the mean squared error (MSE), the hybrid model recorded the lowest value (22.8), demonstrating better predictive accuracy and reduced large errors. As for the concordance index (C-index), which is an important measure of the order of events in survival analysis, it was highest in the Cox model (0.82), followed by the SSL model (0.81), and then the hybrid model (0.77). However, this measure should be interpreted in the context of other measures, as its slight decrease This may not reflect an overall decline in performance. On the other hand, the Accuracy measure showed a significant superiority for the hybrid model, reaching 0.94, compared to 0.90 in SSL and 0.87 in Cox, indicating the hybrid's ability to distinguish between outcomes with greater accuracy. Regarding the Academic Information Criterion (AIC), the hybrid model achieved the lowest value (212.5), a strong indicator of the model's quality and its balance between complexity and accuracy.

Table 6. shows an analysis of the Hybrid Risk Score values estimated using the hybrid model

Patient ID	Age	Sex (1=Male)	Smoking (1=Yes)	Stage II	Stage III	Stage IV	Radio Therapy (1=Yes)	Immuno Therapy (1=Yes)	Hybrid Risk Score
1	78	0	0	0	0	1	0	0	5.001
2	68	1	0	0	0	1	0	1	4.461
3	54	0	0	1	0	0	1	0	3.108
4	47	0	0	0	0	1	1	0	3.259
5	60	0	0	0	0	1	1	1	3.675
6	78	0	1	1	0	0	0	0	4.811
7	58	0	1	1	0	0	1	0	3.641
8	62	1	0	0	0	1	0	0	4.409
9	50	1	1	0	0	1	0	0	4.11
10	50	1	1	0	0	1	1	0	3.98
11	63	1	1	0	1	0	1	1	4.136
12	75	1	1	0	1	0	0	1	4.89
13	79	0	0	0	1	0	0	1	4.533
14	63	1	1	1	0	0	1	0	4.141
15	42	1	0	0	1	0	1	0	2.979
16	61	0	1	1	0	0	1	0	3.797
17	41	1	1	1	0	0	0	1	2.867
18	63	0	1	0	1	0	0	0	4.286
19	69	1	0	0	1	0	0	0	4.513
20	77	0	1	0	1	0	0	1	4.754
21	41	1	0	0	0	1	0	0	3.317
22	60	1	1	0	1	0	0	1	4.11
23	72	0	0	0	0	1	1	1	4.299
24	51	0	1	1	0	0	0	1	3.147

25	61	0	0	0	0	1	0	0	4.117

HRS) of 25 lung cancer patients based on a variety of clinical and demographic factors. Patient 1, a woman in Stage IV who was not getting any treatment, had a high HR of 5.001. Patient 17, a man who smoked and was in Stage II and was getting immunotherapy but not radiation, had a low HR of 2.867. This shows how the stage of the disease and the absence of treatment may change the risk profile. Stage IV patients had the greatest HRs, particularly those who weren't getting immunotherapy or radiotherapy (Patients 1, 2, 8, and 9). On the other hand, patients 3, 5, 10, and 15 had a low HRS following either radiation or immunotherapy. This shows how much the model says that certain treatments lowered risk. Behavioural variables, like smoking, and demographic factors, such being older and male, also raised the overall risk values. When these parts are put together, they generally have bigger numbers. A few younger patients, such those who were 17 and 21, also did badly. This illustrates how age may help minimise danger. These results show that the hybrid model does a good job of showing how different factors work together to change patients' risk levels. Doctors may find it helpful for picking a treatment plan and keeping an eye on each patient because of this.

4. Conclusion

The study found that all three models—the hybrid model, the supervised self-learning (SSL) model, and the Cox proportional hazards model—were extremely good at looking at survival data and predicting the risk of lung cancer. However, they did work and were accurate to different degrees. A high concordance index (C-index) means that the Cox model performed an excellent job at showing how event risk and variables are connected in a straight line. It didn't fare as well on other criteria, such accuracy and mean square error (MAE and MSE). On the other hand, the self-learning model was quite good at finding nonlinear patterns in the data without having to classify them. Also, it was more accurate and made less mistakes than the Cox model. So, it's a good way to work with big or unlabelled data. The hybrid automobile did better than the others overall. It had the best balance between accuracy and usability, as shown by its greatest prediction accuracy (Accuracy = 0.94), lowest mean error (MAE = 2.1 and MSE = 22.8), and lowest AIC value. It also revealed that it could more precisely illustrate the complicated and overlapping effects of explanatory variables, giving it a better prediction model for use in a clinical environment. Based on the previous explanation, one might say that utilising both the representational power of AI algorithms and the explanatory power of traditional statistical models provides us better and more flexible ways to look at survival data. The hybrid model is the greatest choice for future research that needs to swiftly and properly figure out the risk of lung cancer or other complicated medical conditions. The hybrid model's risk constraint results demonstrate that a variety of clinical and demographic factors affect the risk level of lung cancer patients. These factors include the patient's age, gender, smoking status, and kind of medication. People in Stage IV were more at risk, especially if they weren't getting immunotherapy or radiation treatment. This shows how bad it is for individuals if they don't get help as their illness becomes worse. On the other hand, immunotherapy and radiation greatly lowered the risk. Younger people who didn't smoke were also less likely to become sick. We may claim that the hybrid model is a good way to figure out the risk and which therapies should be taken first.

Recommendations

This paper says that one should use a hybrid model which combines Cox proportional hazards regression with a supervised self-learning algorithm. This is because it is better at revealing the complex and non-linear relationships among risk factors and lung cancer risk, and predicting what might occur. It also says self-learning algorithms should be more widely employed in medical analytics – particularly when there are large amounts of unclassified data. And this is in part because they can discover these hidden patterns and create these amazing representations without ever showing it to anybody. The findings of the study demonstrate the significance of integrating traditional statistical models and AI technologies since such an approach would allow the construction of

analytical tools that are capable of coping with the fluctuations which occur within the clinical datasets more efficiently. To make sure that the models.

REFERENCES

[1] D. R. Cox, "Regression models and life-tables," J. Royal Stat. Soc.: Series B (Methodological), vol. 34, no. 2, pp. 187–220, 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.

[2] I. Kuitunen, V. Ponkilainen, M. Uimonen, A. Reito, and J. Mäkelä, "Testing the proportional hazards assumption in Cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review," BMC Musculoskelet. Disord., vol. 22, no. 1, Article 435, 2021, doi: 10.1186/s12891-021-04379-2.

[3] A. S. Singh and S. Dlamini, Analytical Models of Survival Analysis: Concepts and Their Applications, Aug. 2021. [Online].Available: https://www.researchgate.net/publication/353726189_Analytical_Models_of_Survival_Analysis_Concepts_and_Their_Applications [Accessed: Jun. 22, 2025].

[4] D. Collett, Modelling Survival Data in Medical Research, 2nd ed., Boca Raton: CRC, 2003, ISBN 978-1584883258.

[5] "Proportional Hazards Model — an overview," ScienceDirect Topics. [Online]. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/proportional-hazards-model [Accessed: Jun. 22, 2025].

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," arXiv preprint, arXiv:2002.05709, 2020, doi: 10.48550/arXiv.2002.05709.

[7] N. Giakoumoglou and T. Stathaki, "A review on discriminative self-supervised learning methods," arXiv, May 2024.

[8] L. Zheng, B. Jing, Z. Li, H. Tong, and J. He, "Heterogeneous contrastive learning for foundation models and beyond," arXiv, Apr. 2024.

[9] N. Giakoumoglou and T. Stathaki, "A review on discriminative self-supervised learning methods," arXiv, May 2024.

[10] M. R. Mohd Rosli, M. I. Ramli, X. Gao, et al., "Revisiting self-supervised contrastive learning for imbalanced classification," Int. J. Electr. Comput. Eng., vol. 15, no. 2, pp. 1949–1960, 2025.

[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, The MIT Press, 2016.

[12] A. Ramadan, K. Omar, and M. F. Mohammad, "A novel method to detect segmentation points of Arabic words using peaks and neural network," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 7, no. 2, pp. 625–631, Apr. 2017, doi: 10.18517/ijaseit.7.2.1824.

[13] B. Suvarnam and V. S. Ch, "Combination of CNN-GRU model to recognize characters of a license plate number without segmentation," in 2019 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), pp. 317–322, 2019, IEEE.

[14] M. S. Gondere, L. Schmidt-Thieme, A. S. Boltena, and H. S. Jomaa, "Handwritten Amharic character recognition using a convolutional neural network," Arch. Data Sci., 2020.

[15] A. Lamsaf, M. A. Kerroum, S. Boulaknadel, and Y. Fakhri, "Recognition of Arabic handwritten words using convolutional neural network," Indonesian J. Electr. Eng. Comput. Sci., vol. 25, no. 2, pp. 939–944, May 2022.