Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 27 No. 3 (2024)

Nonlinear Analysis of Machine Learning Applicability for Survival Analysis of Lung Cancer Patients

Rupa Fadnavis [1],Shilpa M.Dhopte [2],Shweta Sondawale [3],Swati Kale [4],Rupali Shyam Saha[5],Jagdish. D. Kene[6]

[1]Yeshwantrao Chavan college of Engineering, Department of Computer Science and Engineering, Nagpur, Maharashtra, India. rafadnavis@ycce.edu

[2]MIT School of computing, MIT Art, design and Technology University, Department of Computer Science and Engineering, Pune, Maharashtra, India. shilpa.dhopte@gmail.com

[3]Sinhagad academy of Engineering, Department of Computer Science and Engineering, Pune, Maharashtra, India.

shweta.sondawale123@gmail.com

[4]Yeshwantrao Chavan college of Engineering, Department of information Technology, Nagpur, Maharashtra, India.

swati79kale@gmail.com

[5]Department of Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra, India. saharupali01@gmail.com

[6]Department of Electronics and Communication Engineering, Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India. kenejd@rknec.edu

Article History:

Received: *16-02-2024*

Revised: *26-04-2024*

Accepted: *18-05-2024*

Abstract

This study about looks into how nonlinear machine learning (ML) can be utilized to foresee the survival of lung cancer patients in arrange to create expectations more exact and make strides understanding comes about. Lung cancer is still one of the foremost deadly types of cancer within the world, so better approaches ought to be found to precisely analyze and arrange medicines. Whereas conventional measurements strategies have made a difference us get it survival examination, they aren't continuously great at catching the complicated, nonlinear connections that exist in organic information. The think about looks at how to combine diverse machine learning models, like Irregular Woodlands, Slope Boosting Machines, and Neural Systems, to see at nonlinear designs and connections in DNA and clinical datasets of lung cancer cases. A part of consideration is paid to methods for choosing highlights and utilizing progressed arrangement strategies to bargain with expansive sums of information and lower clamor. We utilize strict cross-validation strategies and comparison investigation to check how well ML models work compared to conventional factual strategies. To see how well a demonstrate can anticipate how long a understanding will live, measurements just like the Concordance File, Brier Score, and Log-Rank Test are utilized. The comes about appear that machine learning models, particularly those that are great at recognizing nonlinearities, do much way better than standard strategies, making gauges that are more exact and dependable. The comes about appear that machine learning has the ability to alter the way survival analysis is worn out cancer

by giving personalized data around forecast and making it easier to select centered treatments. Within the future, analysts will center on combining information from numerous omics and making machine learning models that can be caught on. This will make these progressed examination devices indeed more valuable in clinical settings and boost believe in them. This work includes to the developing sum of information that machine learning is an vital device for moving forward cancer inquire about and quiet care.

Keywords: Nonlinear Survival Analysis, Lung Cancer Prognosis, Machine Learning in Oncology, Predictive Modeling, Clinical and Genomic Data, Personalized Therapy Decisions.

1. Introduction

As per research, about 25% of cancer fatalities are caused by lung cancer making it one of the leading causes of mortality in humans. Over the past few years, millions of new lung cancer cases, and lung. cancer deaths have been observed. In many medical studies, the endpoint evaluated is the time to the “event of particular interest” which can be can be death also. This period is called as time of survival or time of event or time of response. Time of survival can be applied to survival from complete remission to recurrence or progression, and from diagnosis to death. In analysis of cancer data, the response time between treatment and recurrence of cancer or time of recurrence free survival is measured. Here the key features are: what is the event and start and end of observation period. For example, one may be curious to find recurrence of the time between a confirmed response and the recurrence of cancer for the first time. If the event happens to everyone, there are many analytical methods available. However, the real time of the event is generally unknown as at the endpoint, some people did not have an event of interest. also, survival data may not be normally distributed always, it can be distorted and usually contains many early events and a relatively small number of slow events. Due to these data properties, it requires a special method called survival analysis, that uses statistical approach to compute the time required to occur a particular event. So, survival analysis is related to history of event and probability. And very useful for studying various situations. This paper discusses about application of survival analysis for lung cancer data. Survival analysis in cancer research focuses on certain following features:

- Clinical features affecting a patient's survival, like difference between groups of people with high blood sugar and low blood sugar or may be based on gender.
- Survival time probability of a person may be computed for year, month, days, etc.
- Survival rate may vary from patient to patient, if they receive different types of treatment.

1.1. Basic terms of Survival Analysis

Survival time is the time it takes to actively investigate whether a person is alive. Events can be: death, recurrence: a deterioration of patient’s health after temporary improvement, gradual development into a process of appearance or a more advanced state, may be health improvement.

Data Censoring: Survival analysis cornerstone the happening of events like death, but it may not be observed because of variety of reasons. The survival time of subset of the study group is unknown. These observations are named as censored observations. Censoring can be done in following ways:

1. The patient did not witness death or recurrence of any event during the study period.

2. Patients are no longer observed.

3. If the patient relocates to other place(city), then aftercare by the hospital staff may not be maintaining the integrity of the specifications possible.

This survival time will be censored and can miscalculate the true but unknown time to the event. Censored survival time can be of three types: In Right censoring, patient's survival process as a time-line, when the event is beyond the end of the follow-up period presuming it happened, is considered. Censoring also consider the situations like presence of a state or condition is observed but where it began may not be known. For example, the time required for recurrence of cancer after surgical resection of a primary tumor. The patients who had a recurrence of cancer, when examined certain months after surgery would have a Left censored survival time as the actual recurrence time was less than specified months after surgical resection. When data is available for some intervals only, that is individuals come in and out of observation is called Interval censoring. Generally, the survival data has right censored observations.

Survival and Hazard functions

Types of Probabilities calculated for analysis of survival data are:

- The survival probability
- The hazard probability

Probability that a patient survival time from detection of disease to a particular time t in future is called Survival Probability and is denoted by S(t). For example, S(100) = 0.7 indicates the patient's survival probability is 0.7 after 100 days has passed since the diagnosis of disease. Data will be censored, if the person stays alive at the end of an experiment.

Probability that a patient undergoing treatment at a time t died, (that is event of interest) that time is called "Hazard probability" shown as h(t), For example,

if h (210) = 0.8, indicates 80 percent chances of that patient being dead at time t=210 days. As survivor function focuses on not experiencing an event, and hazard probability focuses on the experiencing an event, high chances of survival and low hazard probability is efficient. So, main objectives of Survival analysis include:

- Consider survival data and estimate and interpret survival or hazard functions and relate them.
- To find and analyze the relation between time to live and independent variables.

2. Related Work

The paper [1] describes use of survival analysis to predict period of how long a COVID19 patients will stay in hospital. In their study, they have used survival analysis techniques to create predictive

models that could predict patient's stay period by selecting patient discharge time as a relevant event. This time it was very important because it allows hospital management to prepare for hospital congestion. Authors concluded that the gradient boosting survival model found to be more efficient for predicting patient survival. This research study helped health authorities make more informed decisions in the event of an outbreak. In paper [2], authors explained the two-year period of cell degradation in lung cancer patients by a new methodology presented by oncologists to examine the skin. around the growth volume. The author has determined the best strategy for predicting the 2-year endurance of malignant growth using machine learning methods. In this article, advanced predictions were observed at 6 pixels outside the cancer volume and about 5 mm outside the first GTV, while achieving 71.18% accuracy using a support vector machine.

Authors in Paper [3] describes survival analysis for tuberculosis patients. They investigated risk factors which impact the survival period and resistance to multiple drugs of tuberculosis patients. Various statistical techniques were used to measure differences in survival time between patients. Authors used logistic regression analysis and Cox regression model. The results provide early intervention for tuberculosis, increased health awareness socially, and better control over factors which affect the survival. Authors of paper [4] discussed about deep learning methods used for survival in medical domain. They have computed cost to process censored survival data using their cost functions. They applied a flexible two-tier approach for predicting survival risk. Each subject's time of survival was first translated into a set of pseudo survival probabilities which are conditional and then "deep neural network" model was applied to it. Authors have thereby transformed survival analysis problem to regression problems by using these pseudo values, and simplified the construction of neural networks. In paper [5], authors have applied a parametric model on breast cancer patients' data to evaluate predictive attributes that may affect the survival of these patients. The log rank test and forward approach Weibull models were used for univariate as well as multivariate analysis. Both methods i.e Univariate and Multivariate analysis identified lymph node status, histological, vascular invasion, and grade to be statistically significant attributes. The survival rate was found to be 0.98 yearly.

The authors study provides a detailed classification of contemporary survival analysis methods used in various real-world application domains [6]. They focused on one of the major issues of survival analysis called censorship. over statistical methods developed to overcome censorship problem, many machine learning algorithms were also applied to process survival data. In paper [7], the significance of machine learning for survival analysis has been emphasized by author. The study's main goal was to present some conventional and machine learning methods for survival analysis that have been used in the literature. Survival analysis has been an important part of medical study for a long time, especially when trying to figure out how cancer patients will do. Lung cancer is one of the most deadly types of cancer, so many studies have used different statistical and machine learning methods to try to improve how well they can predict life. Traditional Statistical Methods: Cox proportional hazards regression has been the main tool for survival analysis for a long time. The semi-parametric Cox model, which was first used in 1972, has been used a lot because it can handle restricted data. Researchers have used Cox models to find important factors that affect how well someone with lung cancer will do in the future, such as age, stage, and type of treatment [8]. The Cox model is widely used, but it assumes that the relationship between variables and the hazard

function is linear. This means that it might not be able to fully capture the complicated, nonlinear interactions in the data.

The development of machine learning has led to the creation of more advanced tools that can find complex trends in surviving data. Random Survival Forests (RSF), a variation on the Random Forest algorithm, are being used more and more in this situation. Researchers have shown that RSF can handle large amounts of genetic data and make more accurate guesses about life than older methods [9]. Also, the conditional inference approach is a strong option to the Cox model because it fixes problems with variable selection bias [10].
In survival analysis, neural networks and deep learning models have recently shown promise. The DeepSurv model, which is a Cox proportional hazards model built on deep learning, uses neural networks to show how complex factors interact with each other. Studies on lung cancer patients showed that DeepSurv was more accurate at predicting the future than standard Cox models [11]. Also, deep learning methods used to predict survival in breast cancer can be used to predict survival in lung cancer because the two types of data are similar in how complicated they are [12]. A lot of new study has been done on integrating multimodal data, like putting together clinical, genetic, and imaging data. One study looked at how multi-omics data could be used to predict life from lung cancer. It showed that mixing different types of data can make models work better. This study combined RSF and deep learning to combine clinical and genetic data, which led to more accurate estimates of life [13].

Feature selection and dimensionality reduction are very important when working with the very large amounts of data that are common in cancer research. To shrink the feature space while keeping important data, methods like Lasso regression, Principal Component Analysis (PCA), and Autoencoders have been used. For example, using Lasso-Cox regression to find important factors for prognosis in lung cancer has made it much more accurate to predict life [14]. Several studies have compared how well different machine learning models work with old-fashioned statistical methods. One study looked at lung cancer datasets and compared the Cox model, RSF, and a few deep learning models. It found that machine learning models are usually better at making predictions [15]. This study showed how important it is for models to be easy to understand, especially in clinical settings where it's important to know how each prediction affects the whole. Even with the progress, there are still some problems to solve. One big problem is that complicated machine learning models are hard to understand. Models that can be understood are still being worked on. Examples include attention-based neural networks and AI methods that can be explained. Putting together different kinds of data sources is also hard because it needs advanced preparation and normalization methods [16]. In the future, researchers should work on making models easier to understand and adding more types of data, such as radiomics and proteomics. It is also very important to create strong evaluation tools to make sure that models can be used with a variety of groups and settings [17]. The switch from old-fashioned statistical methods to machine learning methods has made survival analysis in lung cancer studies a lot more advanced. Machine learning models, especially those that can handle complex relationships, are better at making predictions than human models. In the future, a lot of work will go into combining different types of data and making models that can be understood. This will lead to better individual prognoses and treatment plans for lung cancer patients [18].

Table 1: Summary of literature review

Method	Approach	Finding	Limitation	Advantages
Survival Analysis	Predictive models using survival analysis techniques for COVID-19 patient discharge prediction	Gradient boosting survival model was found to be most efficient	Focused on a single event (patient discharge)	Helps in hospital management and preparing for congestion
Machine Learning	New methodology to predict 2-year survival in lung cancer patients using SVM	Achieved 71.18% accuracy with SVM in predicting 2-year survival	Limited to a specific prediction window (2 years)	Provides strategy for better prediction of lung cancer patient survival
Logistic Regression, Cox Regression	Analyzing risk factors for tuberculosis patients' survival and drug resistance	Identified significant risk factors impacting survival time	Focus on tuberculosis; limited generalizability	Helps in early intervention and better control over tuberculosis-related factors
Deep Learning	Two-tier approach using pseudo survival probabilities and deep neural networks	Transformed survival analysis into regression problems using pseudo values	Complexity in transforming survival problems	Simplifies neural network construction for survival analysis
Parametric Models	Log-rank test and Weibull models for breast cancer patient survival analysis	Identified significant attributes impacting survival with a 0.98 yearly survival rate	Limited to univariate and multivariate analysis	Detailed classification of predictive attributes for breast cancer survival
Various Statistical and ML Techniques	Comparative analysis of survival analysis methods addressing censorship	Machine learning algorithms showed promise in handling censored data	General challenges in censorship remain	Comprehensive overview of contemporary survival analysis methods
Conventional and ML Methods	Presenting conventional and machine learning methods for survival analysis	Emphasized the importance of ML in survival analysis	Need for better interpretability of complex ML models	Demonstrated the superiority of ML models in predictive performance
Cox	Traditional	Identified age,	Assumes linear	Widely used due

Proportional Hazards Model	statistical method for identifying prognostic factors in lung cancer	stage, and treatment type as significant prognostic factors	relationships between variables and hazard function	to its semi-parametric nature and ability to handle censored data
Random Survival Forests (RSF)	Handling high-dimensional genomic data for survival predictions	RSF provided more accurate survival predictions compared to traditional methods	Variable selection bias	Effective in uncovering complex, nonlinear patterns in survival data
Conditional Inference Framework	Addressing issues of variable selection bias in survival analysis	Robust alternative to Cox model	Limited exploration in diverse datasets	Corrects variable selection bias, enhancing model robustness
DeepSurv Model	Deep learning-based Cox model for survival analysis	Outperformed traditional Cox models in predictive accuracy	Requires substantial computational resources	Captures complex interactions between covariates using neural networks
Deep Learning	Applying deep learning to survival prediction in breast cancer	Achieved improved survival prediction accuracy	Focus on breast cancer; applicability to lung cancer needs further validation	Similar data complexity allows application to lung cancer survival analysis
Hybrid ML Approach	Combining RSF and deep learning for integrating clinical and genomic data	Improved model performance with multimodal data integration	Complexity in data integration and preprocessing	Enhanced accuracy by leveraging diverse data types
Lasso-Cox Regression	Feature selection technique for high-dimensional data	Improved survival prediction accuracy for lung cancer using key prognostic biomarkers	High-dimensional data processing challenges	Effective in reducing feature space while retaining essential information
Comparative Study	Evaluation of Cox, RSF, and deep learning models on lung cancer datasets	ML models generally offered superior predictive performance	Need for improved model interpretability	Highlights the importance of machine learning in enhancing survival

				predictions
Explainable AI Techniques	Development of interpretable models for survival analysis	Ongoing efforts to make complex ML models more interpretable	Integration of heterogeneous data sources	Facilitates understanding and trust in AI models for clinical applications
Robust Validation Frameworks	Ensuring generalizability of survival analysis models	Emphasis on creating robust validation frameworks	Generalizability across diverse populations and settings	Critical for ensuring reliability and applicability of survival analysis models
Machine Learning and Data Integration	Advancing survival analysis with machine learning models and multimodal data	Superior predictive performance with machine learning models capturing nonlinear relationships	Complexity in combining and preprocessing various data types	Enhanced personalized prognosis and treatment planning for lung cancer patients

3. Materials and Methods

Statistical methods that can be applied for survival analysis include Kaplan-Meier, Log-Rank Test, etc. Major drawback with these methods is that it considers one variable at a time. Also, they only perform operations on categorical variables. As this method is usually not used for numeric data such as age and weight, a machine learning approach can be applied. Survival analysis machine learning algorithms described in the literature include Bayesian method, support vector machines (SVM), neural networks, survival trees, and advanced techniques like ensemble methods , and transfer learning methods. In general, medical research considers several factors to diagnose a person's health and survival. Data can be grouped by age to find out which age group is most likely to survive and by gender to determine which gender is most likely to survive.

3.1 Data requirements in Survival Analysis

Survival analysis does not require an exact start and end point of data. The whole observations always start from zero. The total period is relative. All subjects (i.e patients or machines) are assumed to have common starting point for time t (t = 0), and correspond to the survival probability, that is, the increase in the event of interest (death, churn, etc) is100%. There may be a situation where the large amount of data being used in survival analysis, in such cases, following sampling techniques may be applied: i) simple random sampling ii) stratified sampling.

simple random sampling can be used to select specific number of subjects from each group. First, the total number of subjects determined, then the required total number is divided between each group, and the number of subjects are randomly selected from each group. In stratified sampling, the entire population is separated into groups (called as strata) according to certain traits.

3.2 Machine Learning Approach to Survival Analysis

Machine learning is being used in survival analysis due to various reasons like:

1. Feature selection: machine learning (ML) techniques can automatically identify related features that have a significant impact on survival outcomes

2. Accuracy in Prediction: ML methods like neural networks (NN), random forest (RF), SVM obtain higher prediction Accuracy over traditional survival analysis techniques and it helps in developing accurate risk prediction.

3. Capability of Handling censored data: survival analysis often include censored data, where the required event has not happened for some subjects till the end of the study. Machine learning can handle such censored data effectively and include it in modeling process

4. Handling complex relationship: machine learning algorithms can capture patterns and complex relationships in survival data over traditional statistical methods. This is particularly useful for large and heterogeneous data

5. Ensemble Methods: ensemble methods in machine learning like boosting and bagging can improve the generalization of survival models by combining multiple base learners leading to more reliable predictions

6. Model Flexibility: as machine learning models are flexible, they can adapt to survival data with competing risks, time dependent covariates and non-proportional hazards.

Various machine learning techniques that are commonly used in survival analysis and prediction include:

1. Cox proportional hazards model:

Although it is a statistical method, it is often used for analysis of survival time. It estimates hazard function and assess the impact of covariates on survival time while assuming proportional hazards. Cox Proportional Hazard method is also used to investigate impact of different attributes in a particular dataset on the event of interest. Cox proportional hazards regression analysis can be applied on numeric and categorical variables both. Hazard ratio (HR) is computed part exp(bi). HR value more than 1 indicates that the hazard ration increases as the value of i^{th} covariate increases and so survival time decreases. HR value less than 1 indicates there is reduction in the hazard.HR value= 1 indicates there is no effect.

1. $Define\ the\ hazard\ function$:
   $The\ hazard\ function\ h(t|X) represents\ the\ instantaneous\ risk\ of\ an\ event\ occurring\ at\ time\ t\ given\ the\ covariates\ X.$
   $h(t|X) = h0(t) * \exp(\beta 1 * X1 + \beta 2 * X2 + \ldots + \beta p * Xp)$
2. $Baseline\ hazard\ function$:
   $The\ baseline\ hazard\ function\ h0(t) is\ the$

$hazard\ function\ when\ all\ covariates\ are\ zero.$

$h0(t) = h(t|X = 0)$

3. $Exponentiation\ of\ linear\ predictor$:

$The\ linear\ predictor\ is\ the\ sum\ of\ the\ product\ of$
$coefficients\ \beta i\ and\ covariates\ Xi.$

$\eta = \beta 1 * X1 + \beta 2 * X2 + \ldots + \beta p * Xp$

4. $Hazard\ ratio$:

$The\ hazard\ ratio\ compares\ the\ hazard\ of\ two\ individuals\ with\ different\ covariates\ X\ an$

$$\frac{h(t|X)}{h(t|X')} = \exp\left(\beta 1 * (X1 - X1') + \beta 2 * (X2 - X2') + \ldots + \beta p * (Xp - Xp')\right)$$

5. $Log - likelihood\ function$:

$The\ log - likelihood\ function\ L\ for\ the\ Cox\ model\ is\ given\ by$:

$L(\beta) = \Sigma[\delta i * (\beta^T * Xi - \log \Sigma[\exp(\beta^T * Xj)])]$

$where\ \delta i\ is\ an\ event\ indicator, and\ R(ti)\ is\ the\ risk\ set\ at\ time\ ti.$

6. $Partial\ likelihood$:

$The\ partial\ likelihood\ for\ the\ Cox\ model\ focuses\ on\ the\ order\ of\ event\ times\ rather\ than$

$$L(\beta) = \Pi\left[\frac{\exp(\beta^T * Xi)}{\Sigma[\exp(\beta^T * Xj)]}\right]^{\delta i}$$

7. $Gradient\ of\ log - likelihood$:

$The\ gradient\ (first\ derivative)\ of\ the\ log - likelihood\ function\ with\ respect\ to\ \beta\ is$:

$\partial L/\partial \beta = \Sigma[\delta i * (Xi - \Sigma[Xj * \exp(\beta\hat{}T * Xj) / \Sigma[\exp(\beta\hat{}T * Xj)])]$

8. $Hessian\ matrix$:

$The\ Hessian\ matrix\ (second\ derivative)\ of\ the\ log$
$- likelihood\ function\ with\ respect\ to\ \beta\ is$:

$H(\beta) = -\Sigma[\delta i * (\Sigma[Xj * Xj\hat{}T * \exp(\beta\hat{}T * Xj) / \Sigma[\exp(\beta\hat{}T * Xj)] - (\Sigma[Xj$
$* \exp(\beta\hat{}T * Xj) / \Sigma[\exp(\beta\hat{}T * Xj)])\hat{}2)]$

9. $Newton - Raphson\ update$:

$The\ coefficients\ \beta\ are\ updated\ iteratively\ using\ the\ Newton - Raphson\ method$:

$$\beta(k+1) = \beta(k) - H\left(\beta(k)\right)^{-1} * \frac{\partial L}{\partial \beta(k)}$$

10. $Convergence\ criterion$:

$The\ iterative\ process\ continues\ until\ convergence, typically\ when\ the\ change\ in\ log$
$- likelihood\ or\ the\ change\ in\ \beta\ is\ below\ a\ specified\ threshold.$

$\left||\beta(k+1) - \beta(k)|\right| < \varepsilon$

2. Random forests(RF):

An ensemble learning technique that builds many decision trees and combine their predictions. They are effectively used for survival analysis as it can handle complex interactions and nonlinear relationship present in data

1. Random Forest Decision Function:

The decision function f(X) for a random forest is the average of the decision functions of the individual trees f_m(X):

$$f(X) = \left(\frac{1}{M}\right) * \Sigma \left[f_{m(X)}\right], for\ m\ =\ 1\ to\ M$$

2. Tree-based Decision Function:

The decision function of an individual tree f_m(X) can be represented using indicator functions I for the leaf nodes:

$$f_{m(X)} = \Sigma \left[c_{mj} * I\left(X\ \in\ R_{mj}\right)\right], for\ j\ =\ 1\ to\ J_m$$

where J_m is the number of leaf nodes in tree m, c_mj is the prediction for leaf node j of tree m, and R_mj is the region associated with leaf node j of tree m.

3. Out-of-Bag (OOB) Error Estimation:

The OOB error estimate for a random forest is calculated by averaging the prediction error over all out-of-bag samples:

$$OOB\ Error\ =\ \left(\frac{1}{N}\right) * \Sigma \left[I\left(y_i \neq\ f_{(-i)(X_i)}\right)\right], for\ i\ =\ 1\ to\ N$$

where f_(-i)(X_i) is the prediction for the i-th observation using only the trees that did not include X_i in their bootstrap sample.

4. Gini Impurity for Node Splitting:

The Gini impurity I_G for a node t is given by:

$$I_{G(t)} = \Sigma \left[p_k * (1\ -\ p_k)\right], for\ k\ =\ 1\ to\ K$$
$$= 1\ -\ \Sigma \left[p_k^2\right]$$

where p_k is the proportion of observations of class k in node t and K is the number of classes.

5. Integration of Feature Importance:

The feature importance for a feature X_j can be represented as an integral over all possible splits involving X_j:

$$Importance\left(X_j\right) = \int \left[\Sigma\ \Sigma \left[\Delta I\left(t, X_j, \theta\right)\right] d\theta\right], for\ m\ =\ 1\ to\ M\ and\ t\ in\ T_m$$

where $\Delta I(t, X\_j, \theta)$ is the decrease in impurity at node t for tree m due to a split on feature X_j at threshold θ, and T_m is the set of all nodes in tree m.

3. Support vector machine(SVM): SVM are used for both classification and regression tasks including survival analysis . They work well for non linearly related predictors and survival outcomes as well as high dimensional data.

1. Objective Function:

The objective of SVM is to find the hyperplane that maximizes the margin between two classes. This can be formulated as:

$$\min_{w,b} \left(\frac{1}{2}\right) * \left||w|\right|^2$$

$$subject\ to:$$

$$y_i * (w \cdot x_i + b) \geq 1, for\ all\ i$$

where w is the weight vector, b is the bias, x_i are the training samples, and y_i are the corresponding labels.

2. Lagrangian Formulation:

The constrained optimization problem can be transformed into its Lagrangian dual form:

$$L(w, b, alpha) = \left(\frac{1}{2}\right) * \left||w|\right|^2 - \Sigma \left[alpha_i * \left(y_i * (w \cdot x_i + b) - 1\right)\right], for\ i = 1\ to\ N$$

where alpha_i ≥ 0 are the Lagrange multipliers.

3. Dual Problem:

The dual problem is derived by taking the derivatives of the Lagrangian with respect to w and b and setting them to zero. The dual form of the SVM optimization problem is:

$$\max_{alpha} \Sigma[alpha_i] - \left(\frac{1}{2}\right) * \Sigma\Sigma\left[alpha_i * alpha_j * y_i * y_j * \left(x_i \cdot x_j\right)\right]$$

$$for\ i = 1\ to\ N\ and\ j = 1\ to\ N$$

subject to:

$$\Sigma\ [alpha_i * y_i] = 0, alpha_i \geq 0, for\ all\ i$$

4. Decision Function:

Once the optimal Lagrange multipliers alpha* are found, the decision function for classifying a new sample x is given by:

$$f(x) = sign\ [\Sigma\ [alpha_i ** y_i * (x_i \cdot x)] + b], for\ i = 1\ to\ N$$

where b can be computed using the support vectors.

4. Artificial neural networks (ANN): ANN specifically deep learning models are increasingly applied to survival analysis of high dimensional as well as large datasets and can capture complex patterns in that data.

The forward propagation step computes the output of the neural network by passing the input through multiple layers. For a neural network with L layers, the output a[l] of layer l is given by:

$$z[l] = W[l] * a[l-1] + b[l]$$

$$a[l] = \sigma(z[l])$$

The loss function for survival analysis, such as the partial likelihood of the Cox proportional hazards model, can be integrated with neural networks. The loss function L is given by:

$$L(\beta) = -\Sigma\left[\delta_i * \left(\beta^T * h(X_i) - \log\Sigma\left[\exp\left(\beta^T * h(X_j)\right)\right]\right)\right], for\ i = 1\ to\ N\ and\ j\ in\ R(t_i)$$

The backpropagation step updates the weights and biases of the neural network to minimize the loss function. The gradients of the loss with respect to the weights W[l] and biases b[l] are computed as:

$$\frac{\partial L}{\partial W[l]} = \delta[l] * (a[l-1])^T$$

$$\frac{\partial L}{\partial b[l]} = \delta[l]$$

- where δ[l] is the error term for layer l, computed as:

$$\delta[l] = (W[l+1])^T * \delta[l+1] * \sigma'(z[l])$$

The weights and biases are updated using gradient descent or an optimization algorithm like Adam. The update rule for the weights W[l] and biases b[l] is:

$$W[l] \leftarrow W[l] - \eta * \frac{\partial L}{\partial W[l]}$$

$$b[l] \leftarrow b[l] - \eta * \frac{\partial L}{\partial b[l]}$$

5. Gradient boosting machines (GBM): Like decision trees, GBM is also an ensemble learning technique that generates a sequence of slow learners to create a predictive model. It is robust and performs well in survival analysis.

1. $Gradient\ Boosting\ Framework$:
$The\ goal\ of\ gradient\ boosting\ is\ to\ minimize\ a$
$loss\ function\ L(y, F(x))\ by\ iteratively\ adding\ weak\ learners\ h\_m(x)\ to\ the\ model.The\ genet$
$F_{M(x)} = F_{0(x)} + \Sigma\left[\gamma_m * h_{m(x)}\right], for\ m = 1\ to\ M$
$where\ F\_0(x)\ is\ the\ initial\ model, h\_m(x)\ is\ the\ m$
$-\ th\ weak\ learner, and\ \gamma\_m\ is\ the\ step\ size\ or\ learning\ rate.$
2. $Loss\ Function\ Gradient$:
$At\ each\ iteration, the\ model\ fits\ a\ weak\ learner\ h_{m(x)}$
$to\ the\ negative\ gradient\ of\ the\ loss\ function.The\ negative\ gradient$
$\backslash tilde\{y\}\_i\hat{}\{(m)\}\ at\ iteration\ m\ for\ sample\ i\ is\ given\ by$:
$\tilde{}\{y\}\_i\hat{}\{(m)\} = -(\partial L(y\_i, F(x\_i)) / \partial F(x\_i))|\_\{F(x) = F\_\{m-1\}(x)\}$
3. $Weak\ Learner\ Fit$:
$The\ weak\ learner\ h_{m(x)is} fit\ to\ the\ negative\ gradients$
$\{y\}_i^{\widetilde{\{(m)\}by}}\ minimizing\ the\ sum\ of\ squared\ residuals$:
$h_{m(x)} = \arg\min_h \Sigma\left[\left(\{y\}_i^{\{(m)\}\sim} - h(x_i)\right)^2\right], for\ i = 1\ to\ N$

4. *Update Rule for the Model*:
The model is updated by adding the scaled weak learner to the current model.
The scaling factor γ_m is found by solving:
$\gamma_m = \arg\min_{\gamma} \Sigma\left[L\left(y_i, F_{\{m-1\}(x_i)} + \gamma * h_{m(x_i)}\right)\right], for\ i = 1\ to\ N$
Then, the updated model is:
$F_{m(x)} = F_{\{m-1\}(x)} + \gamma_m * h_{m(x)}$
5. *Integral Form for Continuous Case*:
In a continuous case, the update can be represented as an integral over the function space:
$F(x) = F_{0(x)} + \int [\gamma(t) * h(t,x)dt], from\ 0\ to\ T$
where $\gamma(t)$is a continuous step size function, $h(t,x)$ is a continuously evolving weak learner, and T is the total time.

4. Results and Discussion

Following section of the paper discusses comparison of results of two different methods applied for survival analysis of lung cancer patients.

Dataset Description: dataset shows survival of lung cancer patients in advanced stage from the North Central Cancer Treatment Group (NCCTG). It has data of 228 patients. Performance values indicate how well a patient can perform routine activities in normal manner. various attributes used but details of variables participated in survival analysis is as follows :

1. Survival time (measured in days) as **time**
2. Censoring status as **Status**, where 1 value indicates censored and 2 as dead
3. Age (in years) as **Age**
4. **Sex**: 1 and 2 for male and female respectively .
5. **ph.ecog:** ECOG performance score given by the physician

ph.ecog value- 0 indicates asymptomatic,1- indicates symptotic but completely ambulatory ,2- indicates in bed< 50% of the day , 3-indicates in bed >50%of the day but not bedridden and 4- indicates – bedridden

6. Karnofsky performance score (on a scale of 0 bad to 100 -good) indicated by ph**.karno**
7. **pat.karno** : Karnofsky performance score
8. Loss of weight measured (in pounds) in previous 6 months as **wt.loss**
9. Amount of Calories absorbed at meals as **meal.cal**

Pre-processing on dataset:

i) This step includes removing missing values under one or more attributed from dataset. Rows having missing values are deleted, to avoid error in analysis. After removing such rows , dataset reduced to 167 observations having 47 right censored observations.

ii) Then data is organized as per status. An additional attribute data is considered. if status=1, then patient survived is indicated by dead=0 and if status=2, then patient is dead is indicated by dead=1.

iii) Also, dataset is sorted on time attribute.

Method1: Kaplan Meier (KM)test for survival analysis:

KM method is widely used statistical test which is nonparametric for estimating the survival function. Nonparametric meaning it makes no assumption about the underlying distribution of survival times and directly estimates the survival function from the observed data. KM is particularly used in survival analysis, with the goal of estimating the time until a particular event occurs . KM appropriately handles censored data in estimating the survival function. Probability of a survival of subject after certain time t is given by S(t). The KM calculates this survival function as the product of conditional probabilities of surviving up to each observed time point.

Following steps are applied for estimation using KM test:

- Sorted the observed survival times in ascending order
- Considered events occurred at each point of time and subjects at risk and calculated Kaplan Meier estimator for each observed time point,
- To get overall survival function estimate, multiply conditional survival probabilities
- When Kaplan Meier test applied on above dataset, survival function computed as in Figure1 :

timeline	KM_estimate
0.0	1.000000
5.0	0.994012
11.0	0.988024
12.0	0.982036
13.0	0.976048
...	...
814.0	0.061903
821.0	0.061903
840.0	0.061903
965.0	0.061903
1022.0	0.061903

Figure 1. KM estimate

The estimated survival probability over time is shown in Figure2 :

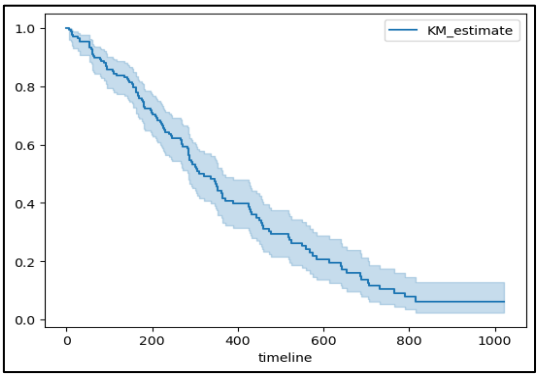

Figure2. Estimated survival probability curve using KM

Here time is indicated on x-axis, and the estimated survival probability on y-axis. This curve illustrates how the survival probability changes over time, typically decreasing as time progresses (assuming the event of interest is death).

Method2: Cox regression method for Survival analysis and prediction.

Objective of this implementation is to analyze the survival probability for patients and to know which factor affects survival. Result summary when cox regression model was fitted on dataset is given in Table 1:

Table 1. Cox regression result summary

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	log2(p)
age	0.01	1.01	0.01	-0.01	0.03	0.99	1.03	0.0	0.93	0.35	1.51
sex	-0.55	0.57	0.20	-0.95	-0.16	0.39	0.85	0.0	-2.75	0.01	7.37
ph.ecog	0.74	2.09	0.22	0.30	1.18	1.35	3.26	0.0	3.29	0.005	9.95
ph.karno	0.02	1.02	0.01	0.00	0.03	1.00	1.05	0.0	2.00	0.05	4.45
pat.karno	-0.01	0.99	0.01	-0.03	0.00	0.97	1.00	0.0	-1.49	0.14	2.87
meal.cal	0.00	1.00	0.00	-0.01	0.01	1.00	1.00	0.0	0.11	0.91	0.13
wt.loss	-0.01	0.99	0.01	-0.03	0.00	0.97	1.00	0.0	-1.83	0.07	3.89

Likelihood ratio test = 28.165 on 7 degrees of freedom, pvalue =0.00021. As p-value obtained here found to be significant, (less than 0.05),on the basis of p- value obtained for ph.ecog and sex, data can be grouped using them. The p-value for gender is 0.01 and the HR (hazard ratio) is 0.57, signifies a prominent association between the patient's gender and decreased death risk. If other covariates kept constant, the risk is reduced by 0.58, or 42%, for females (gender = 2).

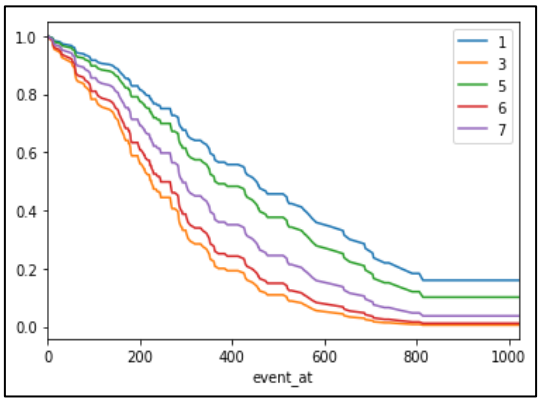

Figure 3. Survival Probability Chart

This indicates that women are more likely to survive. Also the higher value of ph.ecog, indicates lower the survival rate, provided other covariates remain constant. People with high ph.ecog values have an increased risk of death, shown in figure 4. Therefore, physicians try to decrease ph.ecog value by providing the right medicine. This is only a 1% increase in the elderly group, as the age HR is 1.01. So different age groups made no significant difference The following survival probability charts (Figure.3) show that Person1 has the highest chance of survival, Person3 has the lowest chance of survival, and has a high ph.ecog value.

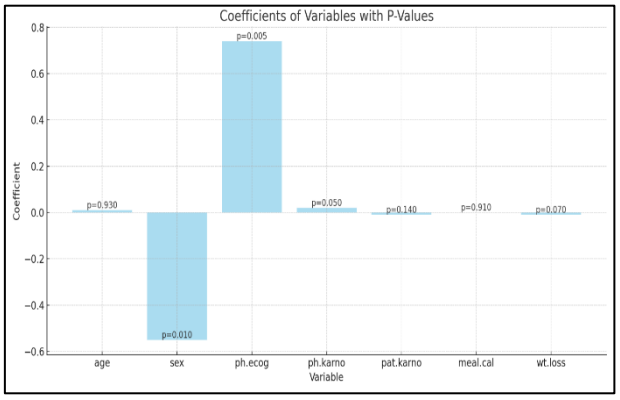

Figure 4: compare the coefficients and highlight their p-values.

5. Conclusion

This paper discusses the application of two survival analysis tests : i)traditional statistical test- Kaplan Meier test and ii) machine learning technique- Cox regression method to Lung cancer data. Kaplan Meier test have provided the estimated survival probabilities for different point times. The limitation with Kaplan Meier method is that it may not perform well with small sample sizes or when large amount of censoring is observed in data. also, it assumes independence of censoring, that is censoring is not related to the likelihood of experiencing the event. The aim of the Cox proportional

hazard method here was to find impact of various attributes in given dataset on the event of interest. In this Lung cancer dataset gender, pc.ecog values are major factors which impact survival time. Here survival probability chart provided the survival probability of different patients.

References

[1] Mohammadreza Nemati, Jamal Ansary, Nazafarin Nemati, Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data", Patterns1,100074, https://doi.org/10.1016/j.patter.2020.100074 (2020)

[2] Alanna Vial, David Stirling, Matthew Field, Montserrat Ros, Christian Ritz, Martin Carolan, Lois Holloway, and Alexis A. Miller, A comparative study of machine learning techniques for the improved prediction of NSCLC survival analysis, 978-1-5386-8494-8/18/$31.00 ©(2018) IEEE

[3] Carter, B. B., Zhang, Y., Zou, H., Zhang, C., Zhang, X., Sheng, R., Qi, Y., Kou,C., & Li, Y. Survival analysis of patients with tuberculosis and risk factors for multidrug-resistant tuberculosis in Monrovia, Liberia. PLOS ONE. https://doi.org/10.1371/journal.pone.0249474

[4] Zhao, L., & Feng, D. Deep neural networks for survival analysis using pseudo values. IEEE Journal of Biomedical and Health Informatics, 24(11), 3308–3314. https://doi.org/10.1109/JBHI.2020.2980204

[5] Baghestani, A. R., Moghaddam, S. S., Alavi Majd, H., Akbari, M. E., Nafissi, N., Gohari, K. Survival analysis of patients with breast cancer using Weibull parametric model. Asian Pacific Journal of Cancer Prevention, vol.16. (2015).

[6] Wang, P., Li, Y., & Reddy, C. K.. Machine learning for survival analysis: A survey. ACM Computing Surveys, 1(1), Article 1. (March 2017)

[7] Fadnavis, R. A. Application of machine learning for survival analysis-a review. IOSR Journal of Engineering (IOSRJEN), 9(5), 56-60. (2019).

[8] Ajani, S. N. ., Khobragade, P. ., Dhone, M. ., Ganguly, B. ., Shelke, N. ., & Parati, N. . (2023). Advancements in Computing: Emerging Trends in Computational Science with Next-Generation Computing. International Journal of Intelligent Systems and Applications in Engineering, 12(7s), 546–559

[9] Changhee Lee, William R. Zame, Jinsung Yoon, Mihaela van der Schaar ,DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks", , Association for the Advancement of Artificial Intelligence www.aaai.org Copyright (2018)

[10] Stephane Fotso, Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework, arXiv:1801.05512v1 [stat.ML] (17 Jan 2018)

[11] Alaa, A. M., and van der Schaar, M. 2017." Deep multi-task gaussian processes for survival analysis with competing risks "In Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)

[12] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger, DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network, arXiv:1606.00931v3 [stat.ML] (9 Aug 2017)

[13] Y. Li, V. Rakesh, and C. K. Reddy." Project success prediction in crowdfunding environments. " In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM '16, pages 247-256, (2016).

[14] Y. Li, B. Vinzamuri, and C. K. Reddy "Regularized weighted linear regression for high-dimensional censored data". In Proceedings of SIAM International Conference on Data Mining, pages 45-53, (2016).

[15] Y. Li, K. S. Xu, and C. K. Reddy." Regularized parametric regression for high-dimensional survival analysis". In Proceedings of SIAM International Conference on Data Mining, pages 765-773, (2016).

[16] Yan Li,Jie Wang ,Jieping Ye ,Chandan K. Reddy, "A Multi-Task Learning Formulation for Survival Analysis", KDD '16, San Francisco, CA, USA c 2016 ACM. ISBN 978-1-4503-42322/16/08. . . $15.00 DOI: http://dx.doi.org/10.1145/2939672.2939857 (August 13-17,2016)

[17] Sattar Ameri, Mahtab J Fard, Ratna B Chinnam, and Chandan K Reddy. "Survival analysis-based framework for early prediction of student dropouts." In Proceedings of ACM International Conference on Conference on Information and Knowledge Management. ACM, 903–912. (2016)

[18] Yifei Chen, Zhenyu Jia, Dan Mercolaand Xiaohui Xie" Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index", Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2013, Article ID 873595, 8pages http://dx.doi.org/10.1155/2013