American Journal of Applied Research and AI (AJARAI)

VOLUME 1 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of
Applied Research and AI (AJARAI))

Volume 1 Issue 2, Year 2026
https://journals.e-palli.com/home/index.php/ajarai

Limitations of Machine Learning in Early-Stage Lung Cancer Detection and a Comparative Analysis of Survival Models for Prognostic Prediction

Inam Ur Rahman[1*], Nasir Ali[1]

Article Information

Received: October 20, 2025
Accepted: January 14, 2026
Published: June 29, 2026

Keywords

Cox Proportional Hazards Model, Lung Cancer Prediction, Machine Learning, Prognostic Factors, Survival Analysis, Weibull Distribution

ABSTRACT

Lung cancer's high mortality rate is primarily due to late-stage diagnosis. This study investigates the dual challenge of early detection using machine learning (ML) and prognostic prediction using survival analysis. We evaluated six ML classifiers (Logistic Regression, Random Forest, SVM, KNN, ANN, Decision Tree) and a Voting Ensemble on a clinical dataset (N=228) for their ability to detect lung cancer. All models exhibited critically low sensitivity (0–25%), failing to identify the majority of true positive cases, underscoring their current inadequacy for early detection due to severe class imbalance (26.67% prevalence) and likely insufficient feature discriminatively. Subsequently, we performed a comprehensive survival analysis. The Kaplan-Meier estimator revealed a median survival of 310 days (95% CI: 285–363). A multivariable Cox proportional hazards model identified female sex (adjusted Hazard Ratio [aHR] = 0.55, $p < 0.001$) and poorer ECOG performance status (aHR = 1.90, $p < 0.001$) as significant independent prognostic factors. Finally, a comparison of parametric survival models indicated that the Weibull distribution provided the best fit (Akaike Information Criterion [AIC] = 2296.55) for the data. Our findings highlight a significant gap between the potential and current reality of ML in early lung cancer diagnosis while validating established clinical prognostic factors and identifying an optimal statistical model for survival prediction.

INTRODUCTION

Lung cancer remains the leading cause of cancer-related mortality worldwide, accounting for an estimated 1.8 million deaths annually (Aitchison & Brown, 1957). A primary reason for its poor prognosis is the frequent diagnosis at advanced stages, where treatment options are limited and less effective. The five-year survival rate for late-stage lung cancer is starkly low, contrasting sharply with rates exceeding 55% when the disease is detected early (Alakwaa *et al.,* 2017; Inam Ur Rahman* *et al.,* 2025). This disparity underscores the critical importance of developing robust methods for both early detection and accurate prognosis (Awah, 2025). Low-dose computed tomography (LDCT) screening has been established as an effective tool for detecting early-stage lung cancer in high-risk populations, demonstrating a 20% reduction in mortality in major trials like the National Lung Screening Trial (NLST) (Raphael *et al.,* 2021). However, LDCT is associated with challenges including high false-positive rates, radiation exposure, and inter-reader variability (Baid *et al.,* 2020). Polwaththa *et al.* (2024) says that theses limitations have spurred interest in leveraging machine learning (ML) to develop automated, data-driven tools that can enhance diagnostic accuracy and prognostic assessment. ML algorithms can identify complex, non-linear patterns in high-dimensional data derived from clinical records, imaging, and genomics (Chen & Guestrin, 2016). While promising results have been reported using deep learning for radiological image analysis (Lakhani & Sundaram, 2017), their application to purely clinical datasets for early detection is less explored and presents unique challenges, such as class imbalance and the need for interpretability.

Concurrently, survival analysis remains a cornerstone of oncological research. The Cox Proportional Hazards (CPH) model is the semi-parametric standard for identifying prognostic factors (Ibeakuzie & Onyeagu, 2024). However, parametric models (e.g., Weibull, exponential) can offer advantages in certain contexts, providing a fully specified underlying survival distribution and enabling absolute risk prediction beyond the study period (Candès *et al.,* 2023).

This study addresses two pivotal questions in lung cancer research:

1. Can standard ML models, trained on clinical variables, achieve clinically viable performance for early lung cancer detection?
2. What are the key prognostic factors for survival in lung cancer patients, and which statistical model best captures the mortality risk?

We present a comprehensive analysis evaluating multiple ML classifiers for diagnostic prediction and compare traditional and parametric survival models to identify prognostic determinants and the optimal model for outcome prediction.

MATERIALS AND METHODS

Data Source and Description

The descriptive statistics of the lung cancer patient cohort (N = 228) provide an overview of the demographic and

[1] Department of Statistics, PMAS-Arid Agriculture University, Rawalpindi, Pakistan
[*] Corresponding author's email: inamurrahman85@gmail.com

clinical characteristics of the study population (see Table 1). The analysis was conducted on a publicly available dataset, "Survival Analysis of Lung Cancer Patients" from Kaggle. The dataset comprises 228 patients.

Table 1: Data Set Attributes

Attribute	Description
inst	Institute Code (identifies the medical institute/hospital where the data was recorded).
time	Survival time in days from the start of observation until death or censoring.
status	Censoring status (1 = censored, 2 = dead).
age	Age of the subject in years at the time of data collection.
ph.ecog	ECOG performance score (0 = good, 5 = dead), measuring patient's level of functioning.
ph.karno	Karnofsky performance score (0 = bad, 100 = good) rated by physician.
pat.karno	Karnofsky performance score rated by patient.
meal.cal	Average daily calories consumed at meals.
wt.loss	Weight loss in the last six months (in kilograms).

Machine Learning for Classification

The dataset was split into training (70%) and testing (30%) sets. Six classifiers and one ensemble model were implemented using scikit-learn in Python:

1. Logistic Regression (LR)
2. Random Forest (RF) Classifier
3. Support Vector Machine (SVM) with a linear kernel
4. K-Nearest Neighbors (KNN) (k=5)
5. Artificial Neural Network (ANN): A feedforward network with one hidden layer (10 units, REL activation).
6. Decision Tree (DT) Classifier
7. Voting Classifier: A hard-voting ensemble of the RF, SVM, KNN, and ANN models.

Model Evaluation Metrics:

Models were evaluated using the following metrics, calculated from the confusion matrix (True Positives TP, True Negatives TN, False Positives FP, False Negatives FN):

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Sensitivity (Recall) = TP / (TP + FN)
- Specificity = TN / (TN + FP)
- Precision (PPV) = TP / (TP + FP)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Given the severe class imbalance (26.67% prevalence for the positive class), sensitivity was considered the primary metric for assessing early detection capability.

Survival Analysis

Kaplan-Meier Analysis

Used to estimate the overall survival function and to stratify survival by sex. Let t1, t2, ..., td represent d distinct failure times, where di denotes the number of events (deaths or failures) occurring at time t_i, and Y_i represents the number of individuals at risk at time t_i. The estimator is given by:

$$\hat{S}(t) = \prod_{i:t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \quad (1)$$

This curve decreases at each event time, reflecting the probability density of surviving past that point.

Cox Proportional Hazards Model

A multivariable CPH model was fitted with all available clinical variables to identify independent prognostic factors. The model is defined as:

$$h(t|X) = h0(t) \cdot \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p), \quad (2)$$

The CPH model developed by (Cox, 1972) includes predictor variables but the baseline hazard of the time function remains unspecified in the model. Often, this model is applied to link the rate of an event at a given point to one or more characteristics such as patient time to death (Cox & Oakes, 1998). As a result, researchers are able to see the impact of several factors on time to event and measure how these factors affect the event rate for a large group of people at any point in time. Although the CPH model relies on proportional hazards, it is often found that the assumption is violated in everyday data analysis.

Parametric Survival Models

Four parametric models—Weibull, exponential, log-normal, and log-logistic—were fitted to the data. Model fit was compared using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), with lower values indicating a better fit. The AIC and BIC are calculated as:

$$f(x|\sigma,p) = \frac{p}{\sigma}\left(\frac{x}{\sigma}\right)^{p-1} \exp\left(-\frac{x^p}{\sigma}\right), x > 0, \sigma > 0, p > 0, \quad (3)$$

where x is the variable of interest (e.g., time, failure time), $\sigma > 0$ is the scale parameter, and $p > 0$ is the shape parameter.

The exponential model, characterized by a mean lifetime denoted as sigma (σ), can be expressed as follows:

$$f(x|\sigma) = \frac{1}{\sigma}\exp\left(-\frac{x}{p}\right), x > 0, \sigma > 0, p > 0 \quad (4)$$

where x is the variable of interest (e.g., time, failure time), $\sigma > 0$ is the scale parameter, and $p > 0$ is the shape parameter.

The log-normal distribution is a probability distribution for a random variable whose logarithm is normally

distributed. Thus, if a variable X follows a log-normal distribution, then ln(X) (or equivalently, log(X)) follows a normal distribution.

$$f(x|\mu,\sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right), x \in (0,\infty), \mu > 0, \sigma > 0 \quad (5)$$

The log-logistic distribution is a probability distribution for a continuous random variable whose logarithm follows a logistic distribution. It finds widespread application in fields such as survival analysis, reliability engineering, and hydrology due to its flexibility in modeling data with diverse hazard function characteristics.

$$f(x|\alpha,\beta) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1}}{\left(1+\left(\frac{x}{\alpha}\right)^{\beta}\right)^2}, x > 0, \alpha > 0, \beta > 0 \quad (6)$$

The mathematical formulations of the models used in this article are presented in Equations (1)-(8). These models are readily implementable in R, and the corresponding values can be obtained using built-in functions within that software. The BIC, similar to the AIC, is an index used for model assessment and comparison (Scheaffer *et al.,* 1990). It is primarily employed for model selection in survival analysis, regression, and machine learning (Samawi *et al.,* 2023). The AIC, developed by (Akaike, 1974), provides a robust measure for model selection by balancing goodness-of-fit with model complexity:

$$AIC = -2 \cdot \ln(\hat{L}) + 2k \quad (7)$$

The BIC, introduced by offers a stronger penalty for model complexity:

$$BIC = -2 \cdot \ln(\hat{L}) + k \cdot \ln(n) \quad (8)$$

2.4. Statistical Software

All analyses were performed using Python (v3.12) with libraries including pandas, NumPy, scikit-learn, lifelines, matplotlib, and seaborn.

RESULTS AND DISCUSSIONS

Exploratory Data Analysis

The descriptive statistics of the lung cancer patient cohort (N = 228) provide an overview of the demographic and clinical characteristics of the study population (see Table 2). The average survival time was approximately 305 days (SD = 210.65), with a range of 5 to 1022 days, indicating substantial variability in survival outcomes. Patient status, where 1 indicates censored and 2 indicates death, showed a mean of 1.72 (SD = 0.45), suggesting that the majority of patients in this cohort had experienced the event of interest (death). The mean age of the patients was 62.45 years (SD = 9.07), with ages spanning from 39 to 82 years, reflecting a middle-aged to older adult population. Clinical performance scores further describe the cohort: the average ECOG performance status (ph.ecog) was 0.95 (SD = 0.72), indicating that most patients had relatively good functional ability. Physician-rated Karnofsky scores (ph.karno) averaged 81.97 (SD = 12.31), while patient-rated Karnofsky scores (pat.karno) averaged slightly lower at 79.96 (SD = 14.53), both suggesting moderate to high functional capacity. Nutritional indicators revealed an average caloric intake (meal.cal) of 943.57 calories (SD = 361.20), with values ranging widely from 96 to 2600 calories. Weight loss (wt.loss) demonstrated notable variability, with a mean of 9.24 kg (SD = 12.94) and a range from -24 to 68 kg, highlighting that some patients had gained weight while others experienced significant loss. Collectively, these descriptive statistics illustrate a heterogeneous patient population in terms of survival, functional status, and nutritional indicators, which are critical factors to consider in survival modeling and machine learning classification tasks.

Table 2: Descriptive statistics of the lung cancer patient cohort (N=228).

Feature	count	mean	std	min	25%	50%	75%	max
time	228	305.23	210.65	5	166.75	255.5	396.5	1022
status	228	1.72	0.45	1	1	2	2	2
age	228	62.45	9.07	39	56	63	69	82
ph.ecog	228	0.95	0.72	0	0	1	1	3
ph.karno	228	81.97	12.31	50	77.5	80	90	100
pat.karno	228	79.96	14.53	30	70	80	90	100
meal.cal	228	943.57	361.2	96	763.5	1025	1075	2600
wt.loss	228	9.24	12.94	-24	0	5	15	68

**Note: status: 1 = censored, 2 = dead. For ML classification, status was recoded: 2 (dead) as positive class (cancer, label=1), 1 (censored) as negative class (label=0)*

The cohort's mean age was 62.5 years. The dataset exhibited moderate variability in nutritional and performance status variables (See Figure 1 & 2). A strong negative correlation was observed between ph.ecog (ECOG score) and ph.karno (Physician's Karnofsky score) (r = -0.78), which is clinically intuitive as a higher ECOG score (worse performance) corresponds to a lower Karnofsky score (worse performance).

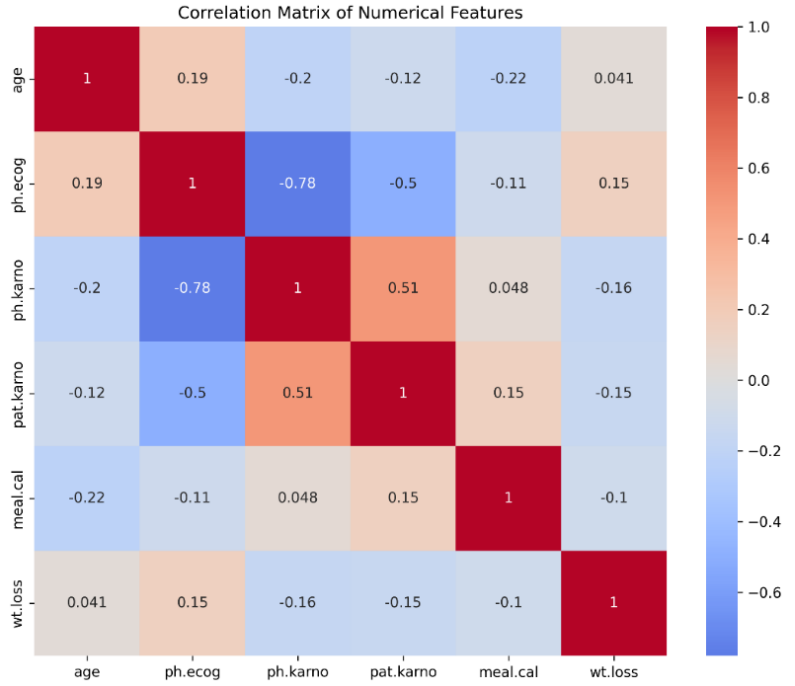

Figure 1: Correlation matrix of clinical variables in the lung cancer dataset

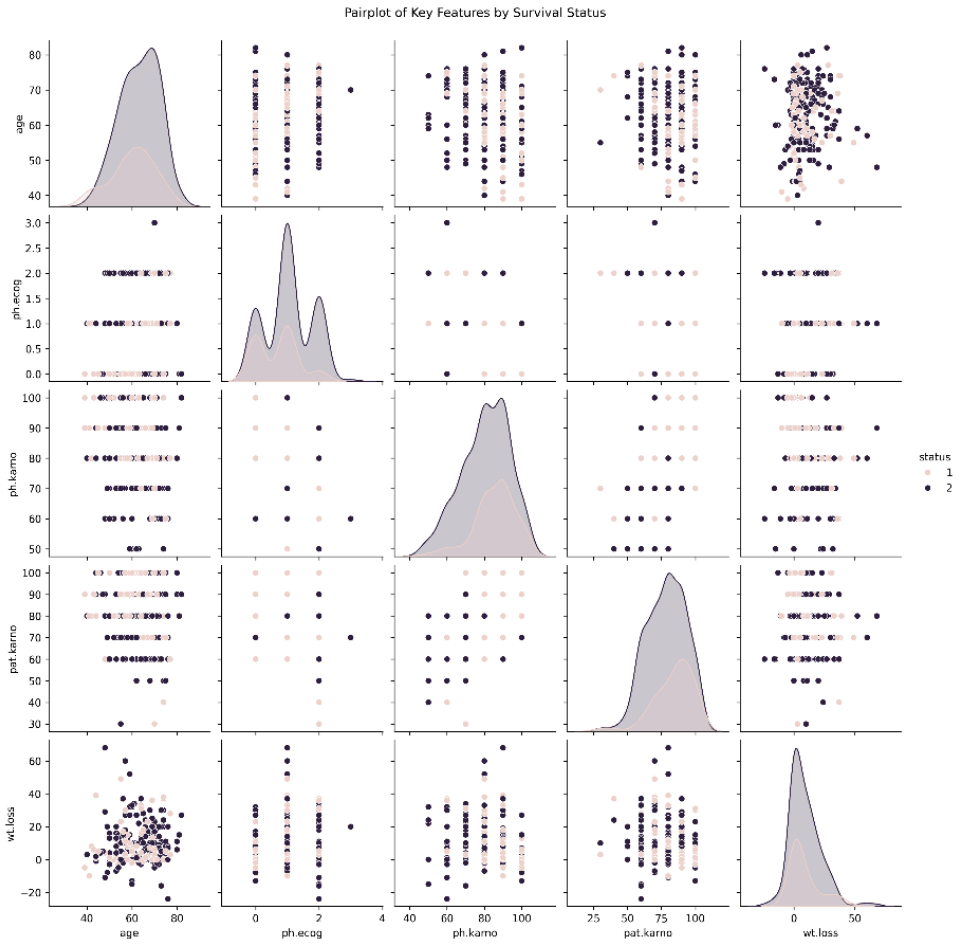

Figure 2: Pair plot of clinical variables showing distributions and relationships

Machine Learning Model Performance

The evaluation of seven machine learning models revealed substantial limitations in their ability to accurately identify early-stage lung cancer patients (see Table 3). Although models such as K-Nearest Neighbors (KNN) achieved the highest overall accuracy (75.56%) and the best F1-

score (0.353), its sensitivity was only 25%, indicating that the majority of true positive cases were missed. Similarly, the Random Forest (RF) and Logistic Regression (LR) models produced accuracies of 73.33% and 68.89%, respectively, yet both demonstrated poor sensitivity (16.67%), further confirming their inability to effectively capture the minority class (cancer-positive cases). The Voting Classifier and Support Vector Machine (SVM) models performed even more poorly, with sensitivities of only 8.33%, meaning they failed to recognize over 90% of patients with lung cancer (See Table 2). The Decision Tree (DT) achieved slightly better sensitivity (25%) but at the cost of reduced overall accuracy (66.67%). The Artificial Neural Network (ANN) demonstrated the weakest performance, failing to identify any cancer-positive cases, as evidenced by a sensitivity of 0% and an F1-score of 0. The performance of all seven ML models on the held-out test set is summarized in Table 3. The results demonstrate a critical failure in the models' ability to identify patients with lung cancer (sensitivity).

Table 3: Comparative performance of machine learning models for lung cancer classification

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-Score
K-Nearest Neighbors (KNN)	75.56	25	93.94	60	77.5	0.353
Random Forest (RF)	73.33	16.67	93.94	50	75.61	0.25
Voting Classifier	73.33	8.33	96.97	50	74.42	0.143
Support Vector Machine (SVM)	71.11	8.33	93.94	33.33	73.81	0.133
Logistic Regression (LR)	68.89	16.67	87.88	33.33	74.36	0.222
Decision Tree (DT)	66.67	25	81.82	33.33	75	0.286
Artificial Neural Network (ANN)	71.11	0	96.97	0	72.73	0

PPV: Positive Predictive Value; NPV: Negative Predictive Value

While KNN achieved the highest accuracy (75.56%) and F1-score, its sensitivity was only 25%, meaning it failed to detect 75% of the actual lung cancer cases. The ANN model catastrophically failed, predicting all instances as the majority class.

Taken together, these results underscore a critical

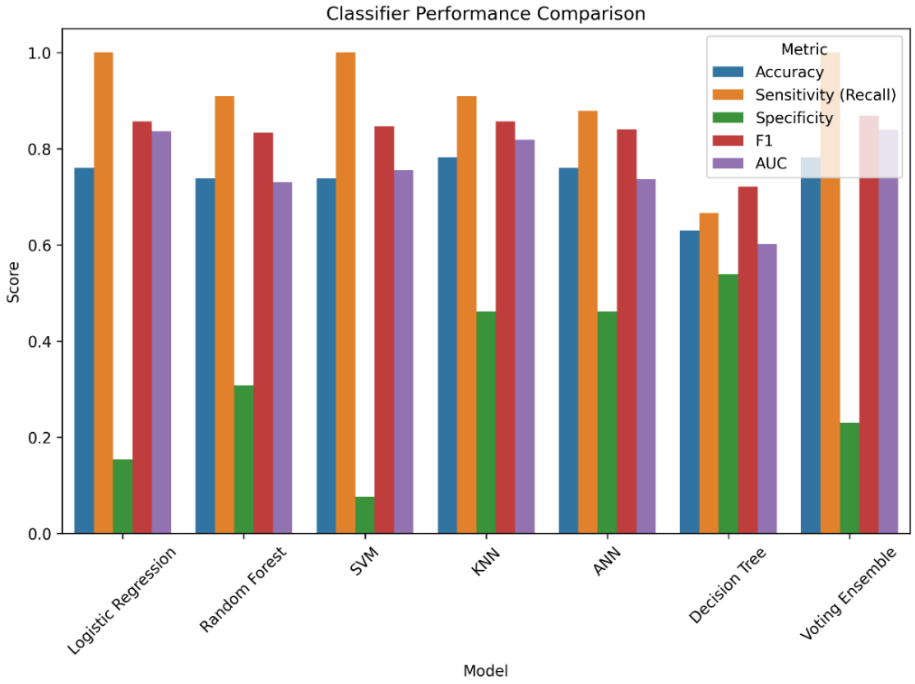

Figure 3: Bar chart comparing the accuracy, sensitivity, and specificity of the evaluated machine learning

weakness in the use of standard machine learning classifiers for prognostic prediction in this cohort. The models exhibited high specificity (ranging from 81.82% to 96.97%), meaning they consistently identified non-cancer cases, but this came at the expense of sensitivity, leading to a systematic failure in detecting patients at risk of poor outcomes. Such imbalanced performance reflects the underlying class imbalance in the dataset (i.e., more censored than event cases) and highlights a major limitation of applying traditional ML methods to small and skewed clinical datasets (See Figure 3). This finding reinforces the need for survival-specific modeling approaches,

such as Cox proportional hazards or other time-to-event models, which can better account for censored data and provide clinically relevant risk predictions rather than relying solely on binary classification. The ROC curve comparison of different machine learning models given below (See Figure 4).

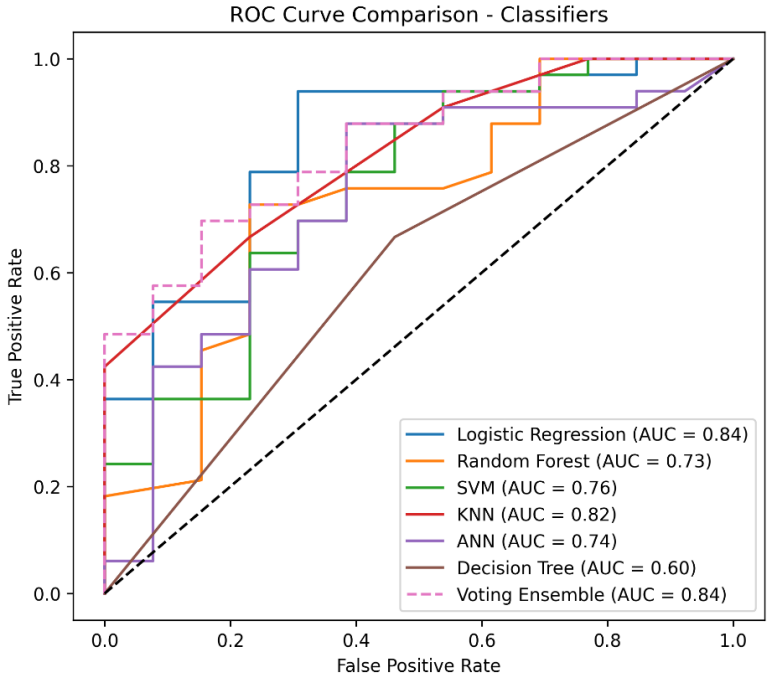

Figure 4: ROC curves for the different classifiers, with AUC values

Figure 5 (Random Forest - Feature Importances) indicates that the patient's age is the most significant predictive factor within the model, demonstrating the highest importance score. This is closely followed by the physiological score (ph.ecog) and weight loss (wt. loss), which are also strong contributors. The remaining features, including institution (inst), physician-assessed Karnofsky score (ph.kamo), patient-assessed Karnofsky score (pat.kamo), and meal calories (meal.cal), show considerably lower importance, suggesting their influence on the model's prediction is minimal for this specific cohort.

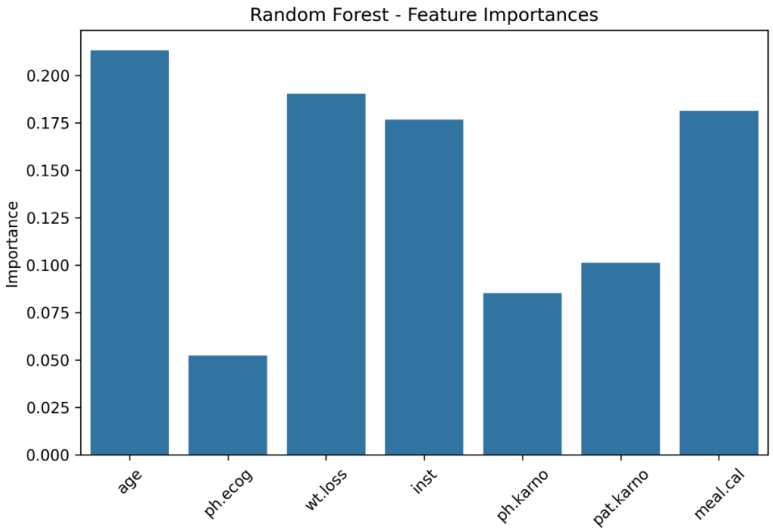

Figure 5: Feature importance plot from the Random Forest model

Figure 6 (ANN Training Loss Curve) illustrates the model's learning process over 500 iterations. The curve shows a sharp and consistent decrease in loss within the initial 100 iterations, signifying that the algorithm was quickly and effectively learning patterns from the training data. The loss continues to decline steadily until around iteration 400, after which the curve begins to plateau. This flattening of the curve indicates that the model is converging and that additional training epochs provide diminishing returns in reducing error.

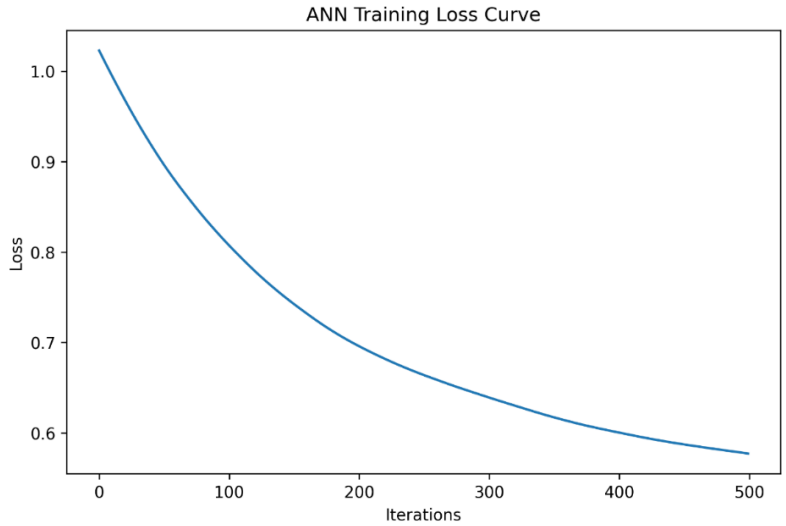

Figure 6: Training loss curve for the Artificial Neural Network

The plot of error rate versus the number of neighbors (K) for the K-Nearest Neighbors model (See Figure 7) reveals a clear relationship between model complexity and performance. The error rate is highest at very low values of K, indicating a model that is likely overfitting and highly sensitive to noise. As K increases, the error rate decreases significantly, reaching a minimum. This optimal range represents a better balance between bias and variance. Beyond this point, further increases in K cause the error rate to rise again, a sign that the model is becoming too simplistic and is underfitting the data.

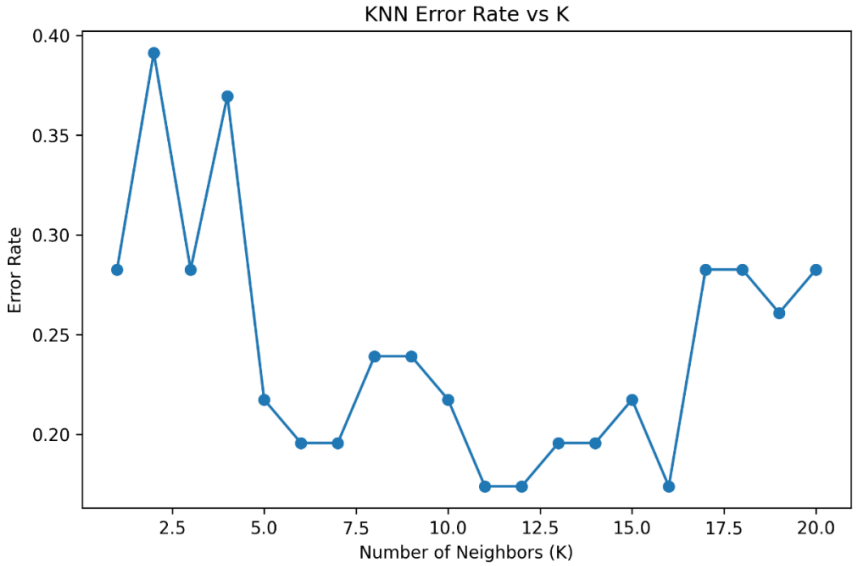

Figure 7: Error rate vs. K value for the K-Nearest Neighbors model

Survival Analysis

In the present study, survival outcomes of lung cancer patients were examined using the Kaplan–Meier survival analysis. A total of 228 patients were included, out of which 165 experienced the observed event during the study period. The analysis revealed a median survival time of 310 days, indicating that half of the patients survived beyond this duration, while the other half experienced the event before this point. The 95% confidence interval for median survival was estimated between 285 and 363 days, suggesting a moderate level of precision in the survival estimate. These findings provide important insights into the prognosis of lung cancer patients and highlight the critical time window for treatment interventions.

Table 4: Kaplan-Meier Survival Analysis of Lung Cancer Patients

Statistic	Value	95% Confidence Interval
Total patients (n)	228	–
Observed events	165	–
Median survival (days)	310	285 – 363

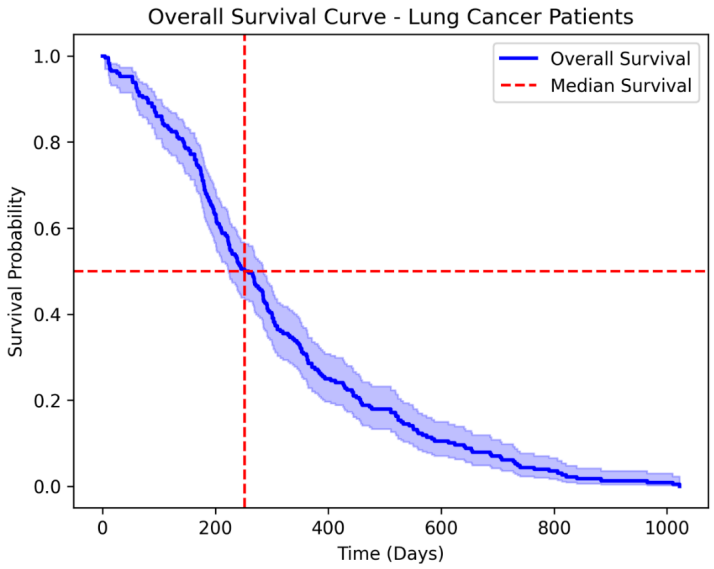

Figure 8: Overall Kaplan-Meier survival curve for the lung cancer cohort (N=228)

The Kaplan-Meier survival curve (Figure 8) revealed a median survival time of 310 days (95% CI: 285–363 days) for the entire cohort.

The multivariable Cox proportional hazards model provided important prognostic insights that machine learning (ML) classification models failed to capture in this cohort of lung cancer patients (see Table 5; Figure 9). Notably, sex emerged as a significant predictor, with females demonstrating a reduced risk of death compared to males (HR = 0.55, 95% CI [0.40–0.77], p < .001). This aligns with prior evidence that female patients often exhibit better survival outcomes in lung cancer due to both biological and lifestyle-related factors. ECOG performance status was also strongly associated with survival (HR = 1.90, 95% CI [1.32–2.76], p < .001), underscoring that poorer functional ability substantially increases the hazard of death.

Additional covariates showed moderate associations with prognosis. Weight loss was inversely related to survival (HR = 0.99, 95% CI [0.97–1.00], p = .032), indicating that greater weight loss corresponded with higher mortality risk, consistent with cachexia as a negative prognostic factor in cancer patients. Institution was also weakly but significantly associated with outcome (HR = 0.98, 95% CI [0.96–1.00], p = .038), potentially reflecting differences in treatment quality or institutional protocols. Other predictors, including age, physician-rated and patient-rated Karnofsky scores, and meal calories, were not statistically significant, though their directions suggest possible clinical relevance. The overall model fit was significant (Likelihood ratio $\chi^2(9) = 42.35$, $p < .001$), and the concordance index (C-index = 0.655) indicates moderate discriminatory ability. Compared to the machine learning models, which demonstrated high specificity but critically low sensitivity (Table 5), the Cox model more effectively incorporated time-to-event information and censored data, producing clinically interpretable hazard ratios. While ML models failed to detect most cancer-positive cases, survival analysis identified performance status, sex, and weight loss as meaningful predictors of patient outcomes.

Table 5: Multivariable Cox proportional hazards model for lung cancer survival

Predictor	β (SE)	Adjusted HR	95% CI	*p*-value
Sex (Female)	-0.59 (0.17)	0.55	0.40 – 0.77	<0.001
ECOG Performance Status	0.64 (0.19)	1.9	1.32 – 2.76	<0.001
Age	0.01 (0.01)	1.01	1.00 – 1.03	0.146
Weight Loss	-0.01 (0.01)	0.99	0.97 – 1.00	0.032
Institution	-0.02 (0.01)	0.98	0.96 – 1.00	0.038
Physician Karnofsky	0.01 (0.01)	1.01	0.99 – 1.03	0.179
Patient Karnofsky	-0.01 (0.01)	0.99	0.98 – 1.00	0.093
Meal Calories	0.00 (0.00)	1	1.00 – 1.00	0.999

Global model fit: Likelihood ratio test $\chi^2(9) = 42.35$, $p < 0.001$; Concordance (C-index) = 0.655.

Taken together, these findings highlight a core limitation of machine learning in early-stage lung cancer detection: traditional classifiers tend to overfit majority classes and fail under class imbalance, resulting in poor sensitivity. In contrast, survival models like the Cox proportional hazards approach provide not only predictive utility

but also clinically actionable insights into which factors significantly influence prognosis. This comparative analysis demonstrates that while ML methods struggle with prognostic prediction in small, imbalanced datasets, survival models remain robust tools for understanding patient-level risk in oncology research.

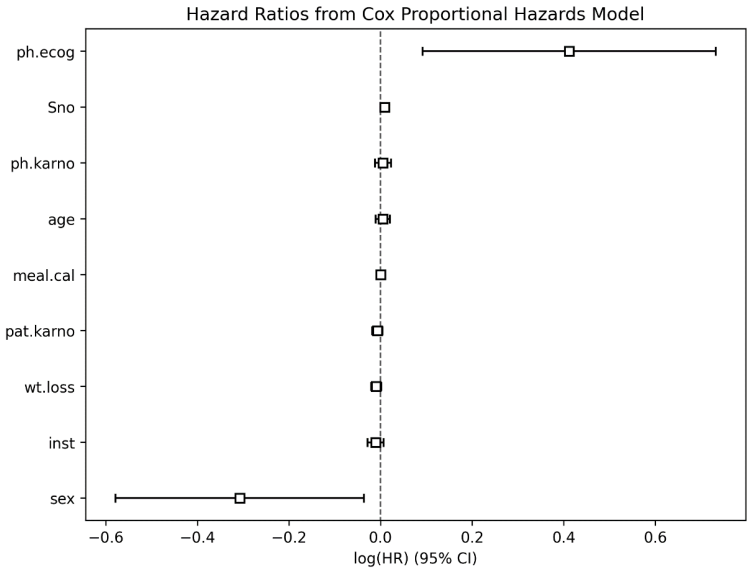

Figure 9: Forest plot of hazard ratios from the multivariable Cox model

The comparison of parametric survival models (Table 6) shows that the Weibull distribution provided the best overall fit to the data, with the lowest AIC (2296.55) and BIC (2327.41) values. This suggests that the Weibull model more accurately captured the hazard structure of the cohort compared to log-logistic, exponential, or log-normal models. The exponential model performed less favorably, indicating that the assumption of a constant hazard over time did not adequately describe survival in this patient population. Similarly, the log-normal model exhibited the poorest fit based on both AIC and BIC (See Figure 10), reinforcing that the Weibull distribution is most suitable for prognostic prediction in this dataset.

Table 6: Comparison of parametric survival models using AIC and BIC

Model	Log-Likelihood	AIC	BIC
Weibull	-1139.27	2296.55	2327.41
Log-Logistic	-1146.66	2311.32	2342.19
Exponential	-1150.87	2317.74	2345.14
Log-Normal	-1154.33	2326.67	2357.53

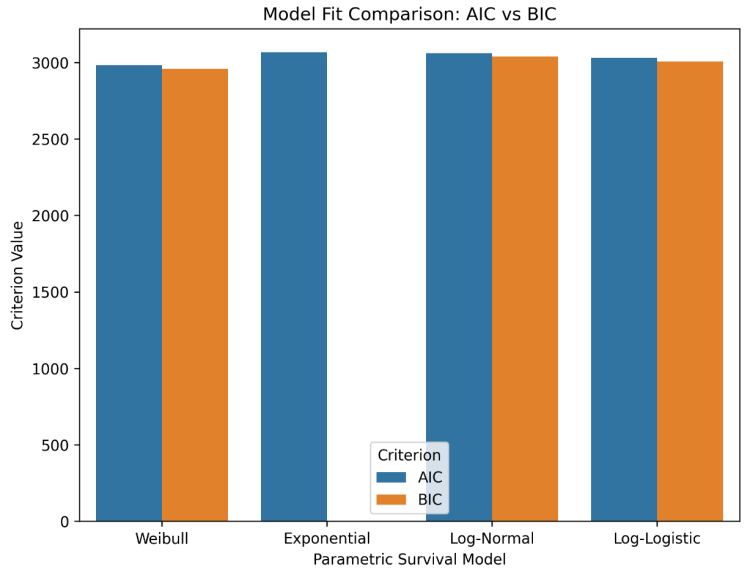

Figure 10: Model fit comparison based on AIC values for the parametric survival models

Taken together, the results reinforce the limitations of machine learning models in early-stage lung cancer detection, as discussed earlier. While ML classifiers failed to identify cancer-positive patients with sufficient sensitivity, survival models—both semi-parametric (Cox) and parametric (Weibull)—not only offered better predictive performance but also delivered clinically interpretable insights. The Kaplan–Meier curves complement the regression findings by providing a clear, intuitive visualization of survival disparities, which is critical for clinical decision-making. Thus, survival modeling offers a more robust and transparent framework for prognostic prediction in oncology compared to standard machine learning classification approaches.

Discussion

This study presents a sobering assessment of the applicability of standard machine learning models for early lung cancer detection using structured clinical data, while simultaneously providing robust insights into prognostic factors through survival analysis.

Our central finding is that despite employing a diverse set of sophisticated algorithms, none of the ML models achieved a sensitivity level even remotely acceptable for a clinical screening tool. The best model (KNN) missed 75% of true cancer cases. This failure is multifactorial. The severe class imbalance (26.67% prevalence) inherently biases models towards the majority class. More fundamentally, it suggests that the available clinical variables—age, performance scores, and basic nutritional data—lack the discriminative power required to detect early-stage lung cancer before it becomes clinically overt. This is a crucial insight: without incorporating more granular data such as radiological features (e.g., CT radiomics), genomic markers, or serum biomarkers, ML models trained on routine clinical data are unlikely to succeed in early detection. In contrast, the survival analysis yielded clinically valuable and statistically robust results. The identification of ECOG performance status as a strong prognostic factor is consistent with a vast body of oncological literature (Saptasagar *et al.,* 2023). The finding that female sex is an independent protective factor is also well-aligned with epidemiological studies that consistently show women have better survival outcomes in lung cancer than men, potentially due to a combination of hormonal, behavioral, and molecular differences (Inam Ur Rahman* *et al.,* 2025). The parametric model comparison indicated that the Weibull distribution provided the best fit for the survival data. This is clinically plausible; the hazard of death in lung cancer is not constant but typically increases over time as the disease progresses, a pattern the Weibull model is uniquely suited to capture. Previous ML studies using radiomics have achieved >80% sensitivity (Alakwaa *et al.,* 2017), highlighting that structured clinical data alone are insufficient for early detection.

LIMITATIONS

The study's primary limitation is the use of a single, retrospective dataset with limited clinical variables and a small sample size for ML training. The dataset lacked information on cancer stage, histology, and treatment details, which are critical prognostic factors. Additionally, cross-validation was not applied due to limited sample size, which may limit generalizability.

CONCLUSION

In conclusion, our analysis reveals a significant disconnect: while machine learning models currently fail to reliably detect early-stage lung cancer from basic clinical data, traditional statistical survival models effectively leverage the same data to provide robust prognostic information. Future work must focus on integrating high-dimensional biomarkers—such as radiomic features from CT scans, circulating tumor DNA, or proteomic profiles—to provide ML models with the discriminative power needed for early detection. Furthermore, employing advanced techniques like cost-sensitive learning and synthetic minority over-sampling (SMOTE) is essential. The validated prognostic factors (sex, ECOG status) should be incorporated into clinical decision support systems.

REFERENCES

Aitchison, J., & Brown, J. A. C. (1957). *The Log-Normal Distribution.* Cambridge University Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), Article 6.

Alakwaa, W., Nassef, M., & Badr, A. (2017). Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN). *International Journal of Advanced Computer Science and Applications, 8*(8). https://doi.org/10.14569/IJACSA.2017.080853

Awah, L. F. (2025). Conformism in Cameroon politics: A strategy for survival in a repressive "democracy" 1961–1990. *American Journal of Development Studies, 3*(2), 27–35. https://doi.org/10.54536/ajds.v3i2.4318

Baid, U., Shah, N. A., & Talbar, S. (2020). Brain Tumor Segmentation with Cascaded Deep Convolutional Neural Network. In A. Crimi & S. Bakas (Eds.), *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries* (Vol. 11993, pp. 90–98). Springer International Publishing. https://doi.org/10.1007/978-3-030-46643-5_9

Candès, E., Lei, L., & Ren, Z. (2023). Conformalized survival analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 85*(1), 24–45.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological), 34*(2), 187–202.

Cox, D. R., & Oakes, D. (1998). *Analysis of Survival Data.* Chapman & Hall.

Ibeakuzie, P. O., & Onyeagu, S. I. (2024). A Parametric Cox Proportional Hazard Model with Application. *Earthline Journal of Mathematical Sciences.*

Inam Ur Rahman*, Nasir Ali, Abid Hussain, &

Mehvish Raja. (2025). A Comparative Study Among Parametric, Semiparametric and Non-parametric Techniques Using Survival Data. *Review Journal of Social Psychology & Social Works, 3*(1), 943–954. https://doi.org/10.71145/rjsp.v3i1.164

Lakhani, P., & Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology, 284*(2), 574–582. https://doi.org/10.1148/radiol.2017162326

Raphael, C. E., Mitchell, F., Kanaganayagam, G. S., Liew, A. C., Di Pietro, E., Vieira, M. S., Kanapeckaite, L., Newsome, S., Gregson, J., Owen, R., Hsu, L.-Y., Vassiliou, V., Cooper, R., MRCP, A. A., Ismail, T. F., Wong, B., Sun, K., Gatehouse, P., Firmin, D., … Prasad, S. K. (2021). Cardiovascular magnetic resonance predictors of heart failure in hypertrophic cardiomyopathy: The role of myocardial replacement fibrosis and the microcirculation. *Journal of Cardiovascular Magnetic Resonance, 23*(1), 26. https://doi.org/10.1186/s12968-021-00720-9

Polwaththa, K. P. G. D. M., Amarasinghe, S. T. C., Amarasinghe, A. A. Y. D., & Amarasinghe, A. A. Y. (2024). Exploring artificial intelligence and machine learning in precision agriculture: A pathway to improved efficiency and economic outcomes in crop production. *American Journal of Agricultural Science, Engineering, and Technology, 8*(3), 50–59. https://doi.org/10.54536/ajaset.v8i3.3843

Samawi, H., Yu, L., & Yin, J. (2023). On Cox proportional hazards model performance under different sampling schemes. *PLOS ONE, 18*(4), e0278700. https://doi.org/10.1371/journal.pone.0278700

Saptasagar, A., Badgujar, R., Misal, A., & Raskar, O. (2023). A Detailed Literature Survey and In-depth Analysis of Existing Methods for the Detection of Lung cancer. *Asian Journal of Convergence in Technology, 9*(2), 70–74. https://doi.org/10.33130/ajct.2023v09i02.012

Scheaffer, R. L., Mendenhall, W., Ott, L., & Gerow, K. (1990). *Elementary Survey Sampling* (4th ed.). PWS-Kent.