Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 4 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i04.52889

PAPER

Lung Cancer Survival Period Prediction: Exploring Machine Learning Approaches

Rooshan Ghous, Seyed Ebrahim Hosseini(✉), Shahbaz Pervez Chattha

Technology Innovation Research Group, Whitecliffe, New Zealand

seyedh@whitecliffe.ac.nz

ABSTRACT

Lung cancer imposes the highest disease burden among all cancers and has the highest expected mortality rate, with 1.8 million deaths annually. It also has the lowest five-year survival rate, averaging at 20% among all diagnosed cancers. Machine learning (ML) offers a novel approach that has been utilized in healthcare for early detection, treatment planning, and survival time estimation. In this study, we applied various supervised, ensemble, and unsupervised ML algorithms to surveillance, epidemiology, and end results (SEER) lung cancer data to predict disease-specific survival (DSS) at 0.5-year, one-year, three-year, and five-year intervals. Our results show that ML models were effective in predicting short-term survival outcomes, but their ability to predict three-year and five-year survival was suboptimal. The limited performance of the models to predict survival outcomes may be attributed to the class imbalance that inherently exists in lung cancer patients. It may also be an indication of limited capacity of the selected features to predict long-term survival. Among the models tested, logistic regression (LR) and XGBoost were most robust algorithms to predict survival outcomes using given features. K-nearest neighbour (K-NN) and deep neural network (DNN) showed relatively weak performance as compared to other models in survival prediction. Additionally, the study found that household income, a socioeconomic factor, was the most significant predictor of survival across all time intervals. These findings highlight the potential of ML in survival prediction, particularly in the short term for lung cancer. The study also emphasizes the importance of addressing socioeconomic disparities as part of public health strategies to improve lung cancer outcomes.

KEYWORDS

artificial intelligence (AI), machine learning (ML), lung cancer, SEER data, logistic regression (LR), decision tree (DT), random forest (RF), XG Boost, support vector machines (SVM), K-nearest neighbour (K-NN), deep neural networks (DNN), socioeconomic status

1 INTRODUCTION

Lung cancer is responsible for approximately 350 deaths per day, making it the leading cause of cancer-related mortality in both men and women [1]. The annual

All authors contributed equally to this work.

Ghous, R., Hosseini, S.E., Chattha, S.P. (2025). Lung Cancer Survival Period Prediction: Exploring Machine Learning Approaches. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(4), pp. 45–60. https://doi.org/10.3991/ijoe.v21i04.52889

Article submitted 2024-10-14. Revision uploaded 2025-01-10. Final acceptance 2025-01-10.

death toll from lung cancer exceeds that of colon, breast, and prostate cancers combined [2]. It remains one of the most common cancers with a significantly poorer survival rate compared to other malignancies diagnosed at the same stage. To estimate the survival time of patients diagnosed with lung cancer, staging systems are widely utilized by clinicians for treatment selection and informed decision-making by the patients. However, while these systems provide valuable insights into population-level trends, they often lack the precision necessary to predict individual survival outcomes [3], particularly for high-risk cases. Consequently, developing robust survival prediction models is critical for improving patient outcomes.

The "survival rate" refers to the percentage of a population still alive at specific intervals after receiving a diagnosis [2]. For lung cancer, survival rates are typically assessed for time frames 0.5-year, one-year, and five-years. Traditional statistical methods have historically relied on patient data to estimate survival periods with a limited ability to handle multi-dimensional data. However, advancements in artificial intelligence (AI), and machine learning (ML) in particular, has opened new frontiers in cancer data modelling. These technologies have demonstrated significant potential in areas such as early detection, treatment selection, drug response prediction, and survival analysis due to their ability to handle complex high dimensional data [4]. An ML model with high predictive accuracy for survival rates can help clinicians avoid unnecessary interventions, optimize treatment plans, and allocate resources more effectively. For instance, patients with a higher anticipated survival time may be prioritized for aggressive treatments, while those with lower survival projections could be directed toward quality-of-life-focused care.

This study utilizes ML techniques to predict disease specific survival (DSS) outcomes in lung cancer patients using data from the U.S. Surveillance, Epidemiology, and End Results (SEER) database. The key objectives of this study are as follows:

1. To develop a robust ML model for predicting lung cancer survival and compare the performance of various algorithms, with the goal of identifying the most effective model.
2. To investigate temporal trends and identify significant risk factors associated with lung cancer survival.

2 LITERATURE REVIEW

For over four decades, patient data has been utilized through traditional statistical models to understand disease processes and facilitate healthcare delivery; however, these methods are limited in their ability to handle complex multidimensional data. Recently, AI has revolutionized the landscape of data inference by uncovering complex relationships inherent in cancer data [5]. ML, a subset of AI, is now being employed for early detection, treatment selection, and survival prediction by identifying patterns within patient data [6].

Studies have demonstrated that ML achieves superior accuracy compared to clinical staging systems and traditional statistical models [7]. This advantage stems from its ability to handle large datasets, process non-linear relationships, and decipher multi-variable interactions. These capabilities have enabled clinicians to make informed decisions for treatment selection and triage, particularly for critically ill in-hospital patients [8].

Among supervised ML algorithms, logistic regression (LR) is widely used for binary classification problems and has been extensively applied in survival prediction. It is favoured for its simplicity, as it does not require hyperparameter tuning while maintaining high accuracy [9]. Naive Bayes (NB), another supervised learning

technique, has also shown success in cancer survival prediction and has outperformed Bayesian networks (BN) [10]. Moreover, NB has been effectively used as a screening tool for early lung cancer detection [11]. Gradient boosting machines (GBM), while limited in their capacity to handle highly complex data, have demonstrated superior performance among various supervised ML techniques in predicting lung cancer prognosis [12].

In a comparative study, Patra evaluated multiple algorithms, including LR, random forest (RF), K-nearest neighbour (K-NN), artificial neural networks (ANN), and support vector classifiers (SVC), using lung cancer data. It was found that the radial basis function network (RBFN) outperformed all algorithms [13]. Further studies compared supervised learning models, such as LR, RF, and GBM, with deep learning models (DLMs), including ANN, recurrent neural networks (RNN), and convolutional neural networks (CNN). The DLMs demonstrated superior predictive capabilities, both in classification and regression tasks, particularly on datasets such as SEER [14][15]. Additionally, DLMs have shown superior results in cancer survival prognosis when evaluated with time-dependent metrics [16].

Ensemble learning techniques, such as decision trees (DT), have outperformed LR, K-NN, NB, and RF classifiers in predicting five-year survival rates [17]. K-NN, while not always the most accurate, has the advantage of interpretability, making it accessible for patient communication [9]. Similarly, RF has demonstrated higher accuracy in predicting cancer mortality risk factors compared to DT, neural networks, LR, and support vector machines (SVM) [18]. Performance discrepancies among models are often observed when applied to different datasets. For instance, LR achieved the highest accuracy on UCI data, while SVM outperformed other models on lung cancer data from Data World [18].

Similarly, model performance can vary across different survival periods. A study using SEER data found that GBM provided the highest accuracy for predicting survival between 0.5 to two-years, whereas RF was the most robust for survival predictions of less than 0.5-years and more than two-years [12]. The development of robust survival prediction models requires the inclusion of relevant risk factors. For lung cancer, key predictors include smoking history, gender [1], race [3], [19], tumour size, stage at diagnosis [20], treatment received [1], and the presence of metastasis. This study aims to predict survival in critical time windows and investigate how significant risk factors change over time [21].

3 METHODOLOGY

Data for this study was extracted from the SEER database, a comprehensive resource of cancer statistics in the United States. Features were selected based on domain knowledge and a thorough review of relevant literature to ensure their relevance to lung cancer survival outcomes. The SEER dataset, being highly structured and well-categorized, required minimal pre-processing, which facilitated a straightforward data preparation phase.

Machine learning algorithms were chosen based on their established effectiveness in cancer survival studies, as highlighted in the literature. A diverse range of modelling approaches was employed, including supervised learning methods, ensemble learning techniques, and deep learning models, enabling a robust comparison of algorithmic performance. The specific algorithms used included LR, DT, RF, XGBoost, SVM, K-NN, and deep neural networks (DNN).

The dataset was divided into training and testing subsets; the models were trained on the training subset. Parameter tuning was performed to optimize model

performance and ensure its generalizability. Model validation was conducted on the test dataset using standard evaluation metrics including accuracy, precision, recall, F-1 score, and AUC. The results were systematically analysed to evaluate algorithmic performance and derive insights into the significance of selected risk factors in lung cancer survival prediction. A brief overview of the methodology is provided in Figure 1.

Feature ranking techniques were employed to assess the role and importance of risk factors in survival analysis. This step provided a deeper understanding of how individual features contributed to the predictive models and the overall survival outcomes.

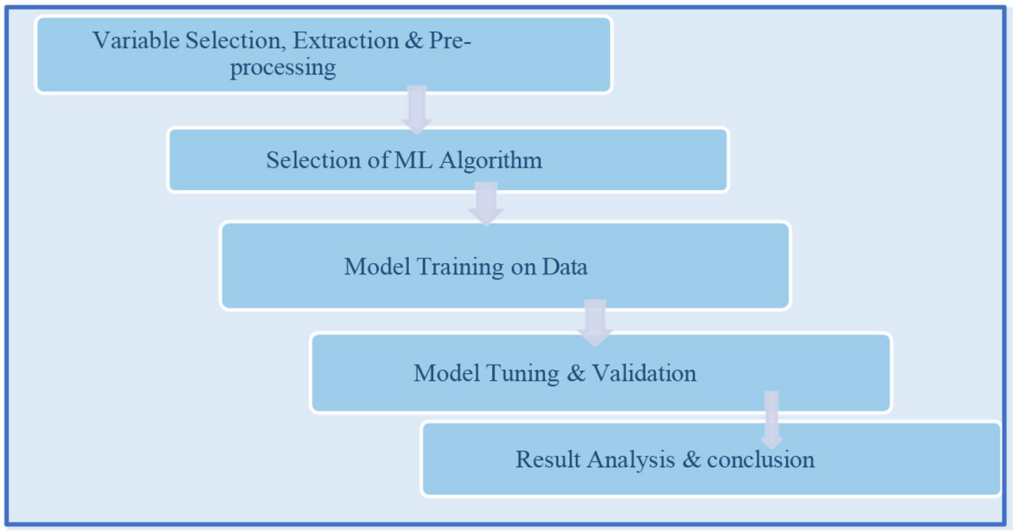

Fig. 1. Development of survival prediction model

3.1 Data selection

Data was retrieved from SEER data, which is one of the largest most comprehensive US-based cancer data repository. It provides extensive coverage (47.0% US population), standardized data, with large number of variables that can form generalizable predictions. The patient factors selected as independent variables were patient demographics, primary tumour site, tumour morphology, evidence of metastasis, stage at diagnosis, first course of treatment, and follow-up status given in Table 1.

Table 1. Feature selection for lung cancer survival prediction

SEER Data Field	Description
Age	Indicates patient's age at diagnosis
Race/Origin	Indicates demographic variable categorizing patients by race and ethnicity
Gender	Indicates patient's gender
Seqnum	Indicates the sequence number of the primary tumour for an individual patient
StageSummary	Indicates the summary stage at diagnosis (localized, regional, distant)
Site	Indicates anatomical site of the primary lung tumour
HouseholdIncome	Indicates the median household income in the patient's geographical area

(Continued)

Table 1. Feature selection for lung cancer survival prediction *(Continued)*

SEER Data Field	Description
Urbun-Rural	Indicates the patient's residential status as urban or rural, influencing access to healthcare
RxsysSxSeq	Indicates the sequence of systemic and surgical treatments
SxRdSq	Indicates sequence of surgery and radiation treatments
SxPrimarysite	Indicates whether surgery was performed on the primary tumour site
CxRecode	Indicates chemotherapy administration
Radrecode	Indicates whether radiation therapy was administered
SxLN	Indicates whether lymph node surgery was performed
SxDistant	Indicates whether surgery was performed on distant metastatic sites
noSxreason	Indicates the reason for not performing surgery, such as patient refusal or clinical ineligibility
Rxdelay	Indicates time from diagnosis to treatment in days
Metslung	Indicates the presence of lung metastases
Metsbrain	Indicates the presence of brain metastases
Metsbone	Indicates the presence of bone metastases
Metsliver	Indicates the presence of liver metastases

Inclusion criteria:

1. All cases diagnosed with malignant lung cancer during the period 2015–2017 reported in SEER Research data with Delay-Adjustment from 22 US Registries.
2. Primary tumour site was lung and bronchus.

Exclusion criteria:

1. Benign lesions and non-invasive tumours were excluded.
2. Patients lost to follow-up and those who died due to reasons other than lung cancer.

3.2 Data pre-processing

Feature selection was performed based on availability of field knowledge and literature. Variables were extracted from the SEER database, with the outcome variable defined as the survival period in months, representing the time a patient survives post lung cancer diagnosis until the end of the follow-up period or the event of death, whichever occurs first. Four survival brackets were selected: 0.5-year (0.5Yr), one-year (1Yr), three-years (3Yr), and five-years (5Yr).

The survival status was engineered as a new variable by creating cut-off values in survival months, resulting in three variables corresponding to the 0.5Yr, 1Yr, and 5Yr survival brackets. A binary classification approach was used to predict the number of survival months where 0 is assigned to non-survival and one to survival class. The patient cohort of three years will be followed up for three time bracket to study their survival characteristics.

Survival period estimation:

$$\text{0.5-year Survival} = \begin{cases} 0 \ if \ patient \ survives \leq 6 \ months \ post \ diagnosis \\ 1 \ if \ patient \ survives > 6 \ months \ post \ diagnosis \end{cases}$$

$$\text{1-year Survival} = \begin{cases} 0 \ if \ patient \ survives \leq 12 \ months \ post \ diagnosis \\ 1 \ if \ patient \ survives > 12 \ months \ post \ diagnosis \end{cases}$$

$$\text{3-year Survival} = \begin{cases} 0 \ if \ patient \ survives \leq 36 \ months \ post \ diagnosis \\ 1 \ if \ patient \ survives > 36 \ months \ post \ diagnosis \end{cases}$$

$$\text{5-year Survival} = \begin{cases} 0 \ if \ patient \ survives \leq 60 \ months \ post \ diagnosis \\ 1 \ if \ patient \ survives > 60 \ months \ post \ diagnosis \end{cases}$$

All object data types were converted to the 'category' type to prepare the dataset for ML. Additionally, variables were encoded for use in the study. Variables providing overlapping information were removed. For instance, 'Year of diagnosis,' 'Year of death recode,' and 'Survival months' extracted to provide information on the 'Survival Status' of patient, were dropped once the variable was engineered using their information. Variables such as 'Site recode' and 'Behaviour recode,' initially extracted to identify malignant cases of lung cancer, were also removed as they were no longer needed.

Imputation techniques in health data are not recommended if data size is sufficient for dropping missing data. Missing data for lymph node surgery 'RX Summ-Scope Reg LN Sur (2003+)' was significantly large therefore category was dropped. All rows with missing data in the four metastasis categories were eliminated. Rows marked as 'blank' for critical variables, such as survival months, were deleted. Fields with 'unknown' outcomes for metastasis (mets) variables were dropped as they provided no actionable information.

A suitable cohort window was selected to ensure that all variable metrics adhered to standardized collection criteria. Numerical variables were standardized using a standard scaler to ensure uniformity in data. Data containing whitespaces was cleaned to make it machine-readable. Time from diagnosis to treatment was binned into predefined intervals based on a literature review: 8, 60, 120, 180, 240, 360, 480 days, and greater than 480 days [22]. One-hot encoding was applied to convert categorical variables into dummy variables for ML models. The final dataset included a total of 136587 cases diagnosed with invasive lung cancer, the number of cases selected over the three-year period is given in Figure 2.

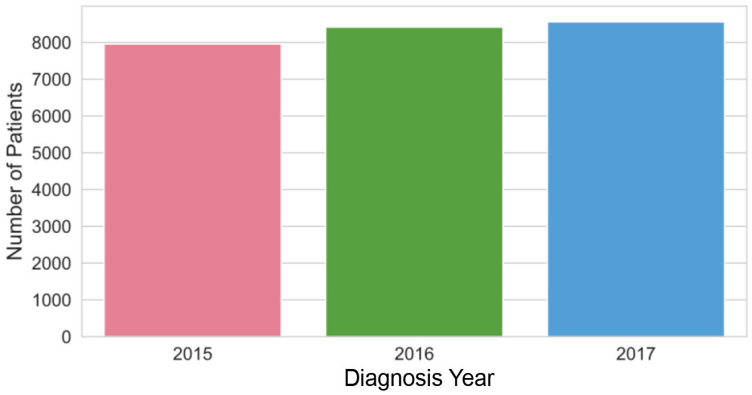

Fig. 2. Number of cases selected from three-year period

Exploratory data analysis (EDA) showed the expected finding that 0.5-year the number of patients who survived was highest as compared to non-survivors, as seen in Figure 3, at one year time the class size between survival vs. non- survival was almost the same as seen in Figure 4. As time progresses after diagnosis, the number of survivors decreases compared to non-survivors, as seen over a three-year period shown in Figure 5, and even further over a five-year period as shown in Figure 6.

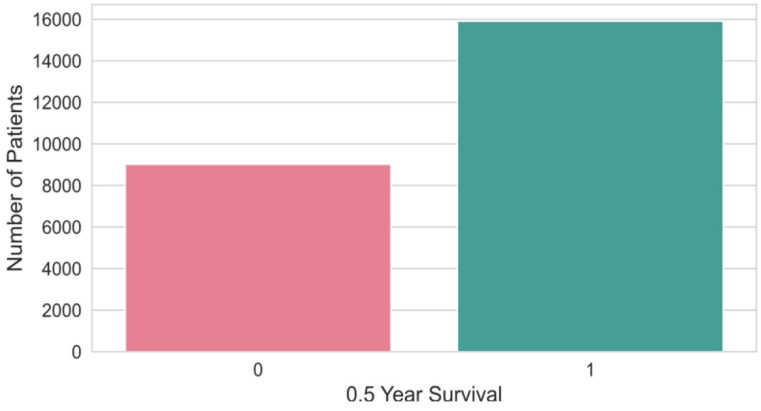

Fig. 3. Comparison of survival vs. non-survival class at 0.5-year

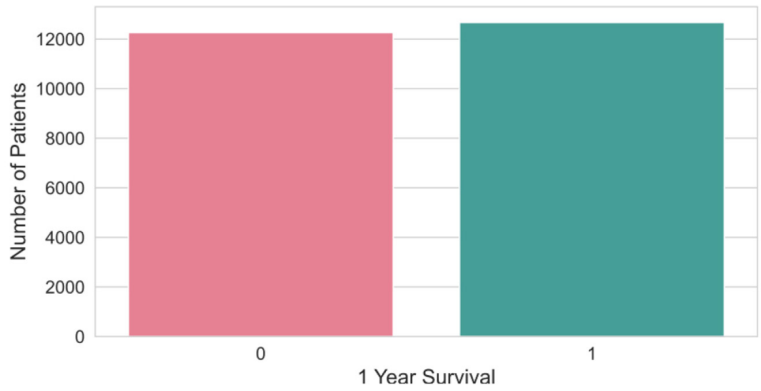

Fig. 4. Comparison of survival vs. non-survival class at one-year

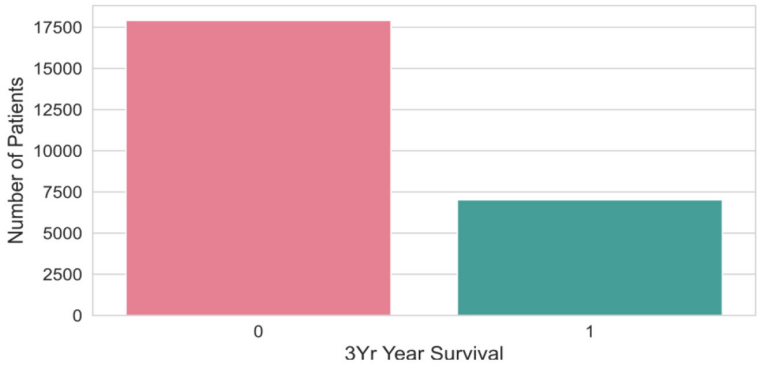

Fig. 5. Comparison of survival vs. non-survival class at three-year

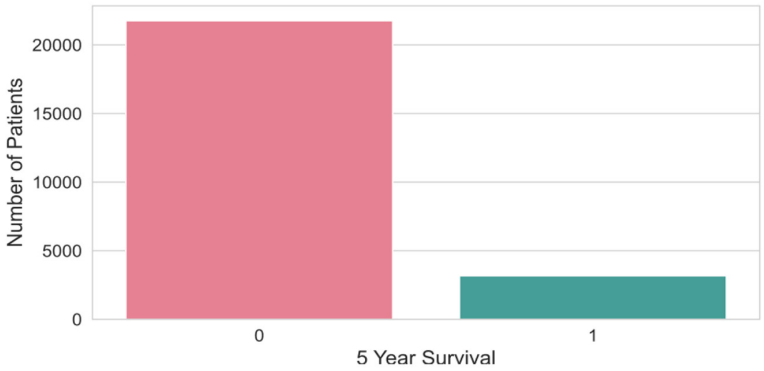

Fig. 6. Comparison of survival vs. non-survival class at five-years

3.3 Model training

A variety of ML algorithms were employed in this study, including LR, DT, RF, XGBoost, SVM, K-NN, and DNN to predict lung cancer survival outcomes, reflecting their established relevance in survival analysis from the literature. LR has been successfully used in cancer survival modelling due to its interpretability and ability to estimate survival probabilities based on risk factors [23]. DT are frequently applied in survival studies as they provide interpretable decision rules for identifying prognostic factors [24]. RF has been shown to perform well in survival prediction by capturing complex feature interactions and providing robust variable importance measures [25]. Similarly, XGBoost, a gradient boosting framework, has demonstrated superior accuracy in survival analysis tasks, particularly in handling high-dimensional and imbalanced datasets, and has been successfully applied in survival prediction [26]. SVM have been effectively used in survival analysis for datasets with non-linear relationships, often coupled with appropriate kernel functions [29]. K-NN, while less common, has shown utility in lung cancer prognosis by leveraging proximity-based survival estimates [27]. Lastly, DNN have emerged as powerful tools for modelling survival outcomes and have been successfully applied in lung cancer survival studies [15]. These algorithms collectively provide complementary strengths, enabling robust predictions of lung cancer survival outcomes.

In this study, we employed hyperparameter tuning to optimize the performance of the ML models used. The hyperparameter tuning process was performed using cross-validation to identify the optimal model configurations, thereby improving the generalization ability of the models in predicting long-term survival outcomes. Given the significant class imbalance observed in the dataset in the long-term survival period, we applied both the synthetic minority over-sampling technique (SMOTE) and random over-sampling to address this issue. SMOTE was utilized to generate synthetic samples of the minority class, enhancing its representation, while random over-sampling was applied to duplicate minority class instances. These techniques were combined to ensure a more balanced dataset, allowing the models to better capture the patterns associated with both survivors and non-survivors.

4 RESULTS

The results presented in Table 2 indicate that XGBoost was the most effective model for predicting 0.5-year disease-specific survival (DSS) in lung cancer, achieving a well-balanced classification between survivors and non-survivors. LR and

SVM show high accuracy, recall, and F1 scores, making them reliable for identifying survival cases, but their precision could be improved. In contrast, simpler methods such as K-NN demonstrated the lowest performance relative to more advanced algorithms. The suboptimal performance of the DNN could be the result of the challenges associated with applying deep learning to this particular dataset, showing relatively low accuracy and precision as compared to other models.

The results of model performance on 1-year data show similar results with LR, DT, and XGBoost showing best performance as seen in Table 3. All three models show high accuracy, AUC, and balanced recall. LR slightly outperformed XGBoost in recall and F1-score. SVM showed a high recall but has lower precision, leading to more false positives. LR or XGBoost are the preferred models for practical use due to their balanced performance across accuracy, precision, recall, and AUC, hence their ability to correctly classify both survivor and non-survivors. DT can be used as an alternative if interpretability is required for decision-making. RF is slightly less reliable albeit the high AUC of 0.72 due to its low recall and F-1 values as compared to other models. Both K-NN and DNN showed poor performance due to their weaker ability to identify positive cases, making them less suitable for survival prediction.

For the 3-year survival prediction, the models exhibited a mixed performance, influenced significantly by the class imbalance between survivors and non-survivors as seen in Table 4. K-NN was able to correctly predict a high number of non-survivors (thus achieving high accuracy), it failed to identify most of the survivors, leading to poor recall and a low F-1 score. LR, DT, XGBoost, and DNN all showed relatively consistent performance. These models had accuracy values ranging from 62.18% to 67.10%, which reflect their ability to distinguish between survivors and non-survivors, but there were challenges with imbalanced class distribution. The DT and RF models had better precision (42.69% and 43.56%, respectively) than some other models but their recall was lower (83.02% for DT and 58.87% for RF), suggesting that while they were relatively good at detecting non-survivors, they missed some survivors, which contributed to their moderate F-1 scores (56.38% for DT and 50.07% for RF). XGBoost and DNN performed similarly in terms of overall metrics, both had an AUC score of 0.74, the highest among the models. These models also achieved relatively balanced results in terms of recall (82.92% for XGBoost and 83.16% for DNN), but their precision (41.66% for XGBoost and 42.33% for DNN) remained lower, reflecting the challenge of maintaining a balance between detecting survivors while avoiding false positives. Both models had F-1 scores around 55%, indicating they were reasonably good at predicting the minority class of survivors without compromising too much on precision.

The results for the five-year survival prediction, as given in Table 5, show that the models struggled with the significant class imbalance, leading to generally low performance in terms of precision and F-1 scores. DT and SVM exhibited high recall (91.95% and 89.88%, respectively), but their low precision (18.30% and 19.26%) and F-1 scores indicate that they heavily misclassified non-survivors as survivors, resulting in a poor overall balance between precision and recall. LR, RF, XGBoost, and DNN all achieved moderate accuracy levels (ranging from 48.58% to 65.30%). Despite this, the precision values across these models were low (ranging from 18.75% to 21.55%), and the recall values were moderate, indicating that while these models did capture some survivors, they still struggled with false positives.

Overall, the five-year survival prediction models demonstrated suboptimal performance, likely due to the severe class imbalance, which skewed the models' ability to correctly predict the minority class (survivors). Even with the application of techniques such as SMOTE analysis for the five-year analysis, the overall balance between precision and recall remained a challenge, with models performing well on recall but poorly on precision. The DNN had the best AUC, but still showed limited

performance when accounting for both precision and the F-1 score, making it unsuitable for use in survival prediction.

Table 2. Model performance on 0.5-year survival

Model Name	Accuracy	Precision	Recall	F-1	AUC
Logistic Regression	69.36%	70.60%	89.06%	78.76%	0.73
Decision Tree	69.56%	73.85%	80.95%	77.24%	0.70
Random Forest	69.20%	73.85%	80.07%	76.83%	0.70
XGBoost	70.97%	74.67%	82.46%	78.37%	0.73
SVM	68.24%	69.59%	89.19%	78.18%	0.70
K-NN	63.07%	67.11%	82.59%	74.05%	0.60
DNN	65.76%	66.59%	92.96%	77.60%	0.69

Table 3. Model performance on one-year survival

Model Name	Accuracy	Precision	Recall	F-1	AUC
Logistic Regression	70.67%	68.07%	79.51%	73.35%	0.75
Decision Tree	71.07%	69.91%	75.51%	72.60%	0.73
Random Forest	67.66%	67.94%	68.71%	68.32%	0.73
XGBoost	70.04%	68.46%	75.98%	72.02%	0.75
SVM	69.28%	66.23%	80.54%	72.69%	0.72
K-NN	57.81%	57.34%	65.92%	61.33%	0.60
DNN	65.67%	67.70%	61.89%	64.67%	0.72

Table 4. Model performance on three-year survival

Model Name	Accuracy	Precision	Recall	F-1	AUC
Logistic Regression	62.18%	41.30%	82.97%	55.15%	0.73
Decision Tree	64.02%	42.69%	83.02%	56.38%	0.71
Random Forest	67.10%	43.56%	58.87%	50.07%	0.70
XGBoost	62.67%	41.66%	82.92%	55.46%	0.74
SVM	59.62%	39.98%	88.07%	55.0%	0.70
K-NN	67.29%	38.35%	27.58%	32.08%	0.59
DNN	63.53%	42.33%	83.16%	56.10%	0.74

Table 5. Model performance on five-year survival

Model Name	Accuracy	Precision	Recall	F-1 Score	AUC
Logistic Regression	64.32%	20.91%	68.44	32.03%	0.73
Decision Tree	48.58%	18.30%	91.95%	30.52%	0.71
Random Forest	63.84%	21.41%	72.80%	33.09%	0.73
XGBoost	65.30%	21.51%	68.88%	32.78%	0.73
SVM	52.47%	19.26%	89.88%	31.72%	0.72
K-NN	85.31%	18.75%	5.87%	8.95%	0.56
DNN	64.99%	21.55%	70.08%	32.97%	0.74

The feature ranking of risk factors showed household income consistently as the most important predictor across all three-time brackets, indicating the critical role of socioeconomic factors in lung cancer survival as seen in Table 6. Features related to metastasis, such as metastasis to the liver, bone, brain, and lung, appeared significant in the short-term survival of 0.5-year period, reflecting their relevance in predicting immediate survival outcomes.

Demographic factors including age and race or origin showed comparatively minimal importance in short-term survival. Age and urban-rural classification gained more significance with time, rising in feature ranking in one-year period, highlighting demographic and geographical influences over this survival period. Gender on the other hand consistently retained its value as middle-value risk factor. Clinical features, such as treatment delay and stage summary, also show increased importance in the long term compared to the short-term bracket.

In the long-term survival, age and urban-rural classification increase in importance, reflecting their cumulative influence on long-term survival. Metastasis-related features are critical for short-term survival but drop slightly in significance over longer periods.

Table 6. Temporal variation between risk factors

Feature Ranking for 0.5-Year Survival		Feature Ranking for 1-Year Survival		Feature Ranking for 5-Year Survival	
Feature	**Importance**	**Feature**	**Importance**	**Feature**	**Importance**
HouseholdIncome	0.903	HouseholdIncome	0.157	HouseholdIncome	0.157
Metsliver	0.021	Age	0.132	Age	0.131
Metsbone	0.020	Urbun-Rural	0.066	Urbun-Rural	0.068
Metsbrain	0.010	Site	0.064	Site	0.063
Race/Origin	0.010	Metsbone	0.057	Race/Origin	0.057
Metslung	0.009	Race/Origin	0.057	Metsbone	0.057
Seqnum	0.008	SxPrimarysite	0.055	SxPrimarysite	0.054
Age	0.004	Metsliver	0.052	Metsliver	0.051
Urbun-Rural	0.002	Rxdelay	0.049	Rxdelay	0.048
SxPrimarysite	0.002	Seqnum	0.046	Seqnum	0.046
SxLN	0.002	StageSummary	0.044	StageSummary	0.045
Site	0.002	SxLN	0.038	SxLN	0.038
Rxdelay	0.002	Metsbrain	0.035	Metsbrain	0.035
StageSummary	0.002	Gender	0.034	Gender	0.035
RxsysSxSeq	0.001	Metslung	0.032	Metslung	0.031
Gender	0.001	RxsysSxSeq	0.019	RxsysSxSeq	0.019
noSxreason	0.001	Radrecode	0.017	Radrecode	0.016
SxRdSq	0.001	SxRdSq	0.015	SxRdSq	0.015
Radrecode	0.001	noSxreason	0.013	noSxreason	0.013
SxDistant	0.001	CxRecode	0.010	CxRecode	0.010
CxRecode	0.000	SxDistant	0.010	SxDistant	0.010

5 DISCUSSION

Currently several prediction models have been developed; however, they are weak in statistical analysis due to weak data pre-processing and parameter selection [28]. Although there is a rapid increase in the use of ML for time-to-event analysis in healthcare, more work is still needed to refine the models. The current research has been limited in robust parameters and vigorous data pre-processing methods that are warranted [6]. Li et al., [4] found that optimization of ML models can be improved via more standardization and validation and can be promising in generating superior predictions of survival rate.

We aimed to develop a DSS prediction model by selecting relevant risk factors, a large patient cohort, and optimal parameter tuning.

In the present study, LR, XGBoost, and DT outperformed other methods in survival analysis. These findings are consistent with previous research, which has demonstrated the successful application of LR and ensemble learning techniques in cancer classification tasks [29, 30, 23]. However, as the survival time extended to the three- and five-year periods, the ability of the ML models to accurately predict survival outcomes diminished, as has been previously shown in predictive analysis for lung cancer survival [25].

Although neural networks are often expected to outperform other models due to their ability to capture complex non-linear relationships in cancer data [14], our study did not identify neural networks as the superior algorithm. This outcome may be attributed to the inherent strengths of XGBoost in handling tabular data and the superior ability of LR to handle bivariate analysis, the typical format of survival datasets. Both algorithms excel at capturing feature interactions and modelling non-linear relationships without requiring extensive feature engineering or hyperparameter optimization. In contrast, neural networks often demand careful tuning of their architecture and hyperparameters to achieve optimal performance. The observed superior performance of XGBoost is consistent with findings reported in existing literature [31].

Socioeconomic factors emerged as critical determinants in lung cancer survival, highlighting their significant influence on access to care, early diagnosis, advanced treatments, and therapy adherence. Temporal trends show that socioeconomic disparities impact outcomes at all stages but are particularly dominant in the short term. Addressing these disparities through targeted policies and education programs could significantly improve survival rates.

Metastatic burden (e.g., liver, bone, brain, lung metastases) is critical in the short term but diminishes in importance over time, as patients with advanced disease rarely reach the five-year survival mark. Staging becomes more relevant in long-term survival, as early-stage patients benefit significantly from curative treatments such as surgery and radiation. Among treatment factors, surgery has increasing importance over time, especially for early-stage patients. Radiation delay shows minimal short-term impact but gains significance over time as it affects recurrence and long-term outcomes.

Age has little impact on short-term survival but becomes a dominant factor in the long term due to chronic health issues and treatment tolerability.

In summary, short-term survival is dominated by acute clinical factors such as metastases and socioeconomic barriers. Intermediate-term survival reflects the increasing importance of clinical interventions such as surgery, while long-term survival is shaped by socioeconomic and demographic factors, emphasizing sustained care, early diagnosis, and optimized treatment strategies. Short-term outcomes

require aggressive management of metastatic disease, while long-term outcomes benefit from consistent follow-up care and addressing age-related challenges.

6 FUTURE WORK

This study, conducted as part of a master's single-subject coursework project, represents an initial exploration and serves as a foundation for more extensive future research, acknowledging the constraints of time and resources. Despite these limitations, it provides valuable insights into the potential of using ML for lung cancer survival prediction. Future work can add value by including a more extensive exploration of additional factors within the SEER dataset and further optimization of predictive models through advanced tuning and validation techniques. These efforts will be essential to improve the accuracy and generalizability of survival predictions and ensure their applicability in diverse clinical settings.

To comprehensively account for the diverse factors influencing survival outcomes, future prediction modelling can benefit from including a wider range of risk factors that influence patient survival. Further, integrating heterogeneous data types, including imaging data, and leverage advanced algorithms such as CNNs [32]. Furthermore, the application of large language models (LLMs) for analysing clinical notes has demonstrated significant utility in survival prediction and should be incorporated into future models. Robust predictive modelling requires careful consideration of the interplay between multiple health determinants. Rigorous validation of models, including the use of external validation strategies, will be crucial to ensure reliability and applicability across different datasets and populations.

Data modelling is inherently sensitive to the characteristics of the sampled population. While the SEER database offers the advantage of a large dataset, it predominantly represents the demographics, healthcare practices, and access patterns of the United States, which may not be reflective of populations in other regions. Variations in healthcare systems, socioeconomic conditions, cultural attitudes, and genetic diversity can influence cancer incidence, treatment patterns, and survival outcomes. As such, findings based on SEER data can potentially have limited generalizability to non-U.S. populations and should be interpreted cautiously when applied elsewhere.

7 REFERENCES

[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA: A Cancer Journal for Clinicians*, vol. 73, no. 1, pp. 17–48, 2023. https://doi.org/10.3322/caac.21763

[2] D. S. Dizon and A. H. Kamal, "Cancer statistics 2024: All hands-on deck," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 1, pp. 8–9, 2024. https://doi.org/10.3322/caac.21824

[3] C. Y. Shao *et al.*, "Online decision tools for personalized survival prediction and treatment optimization in elderly patients with lung squamous cell carcinoma: A retrospective cohort study," *BMC Cancer*, vol. 23, p. 920, 2023. https://doi.org/10.1186/s12885-023-11309-z

[4] Y. Li, X. Wu, P. Yang, G. Jiang, and Y. Luo, "Machine learning for lung cancer diagnosis, treatment, and prognosis," *Genomics, Proteomics and Bioinformatics*, vol. 20, no. 5, pp. 850–866, 2022. https://doi.org/10.1016/j.gpb.2022.11.003

[5] V. Mhasawade, Y. Zhao, and R. Chunara, "Machine learning and algorithmic fairness in public and population health," *Nature Machine Intelligence*, vol. 3, pp. 659–666, 2021. https://doi.org/10.1038/s42256-021-00373-4

[6] G. Huang, S. Song, J. N. Gupta, and C. Wu, "Semi-supervised and unsupervised extreme learning machines," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2405–2417, 2014. https://doi.org/10.1109/TCYB.2014.2307349

[7] M. F. Gensheimer *et al.*, "Automated model versus treating physician for predicting survival time of patients with metastatic cancer," *Journal of the American Medical Informatics Association*, vol. 28, no. 6, pp. 1108–1116, 2021. https://doi.org/10.1093/jamia/ocaa290

[8] Q. N. Tran, M.-K. Le, T. Kondo, and T. Moriguchi, "A machine learning-based model to predict in-hospital mortality of lung cancer patients: A population-based study of 523,959 cases," *Advances in Respiratory Medicine*, vol. 91, no. 4, pp. 310–323, 2023. https://doi.org/10.3390/arm91040025

[9] S. Dubey, G. Tiwari, S. Singh, S. Goldberg, and E. Pinsky, "Using machine learning for healthcare treatment planning," *Frontiers in Artificial Intelligence*, vol. 6, p. 1124182, 2023. https://doi.org/10.3389/frai.2023.1124182

[10] M. B. Sesen, T. Kadir, R.-B. Alcantara, J. Fox, and S. Michael Brady, "Survival prediction and treatment recommendation with Bayesian techniques in lung cancer," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, vol. 2012, 2012, p. 838.

[11] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, p. 100907, 2021. https://doi.org/10.1016/j.tranon.2020.100907

[12] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2018, pp. 632–637. https://doi.org/10.1109/ISSPIT.2018.8642753

[13] R. Patra, "Prediction of lung cancer using machine learning classifier," in *Computing Science, Communication and Security, COMS2 2020, Communications in Computer and Information Science*, Springer, vol. 1235, 2020, pp. 132–142. https://doi.org/10.1007/978-981-15-6648-6_11

[14] M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, p. 91, 2023. https://doi.org/10.3390/computers12050091

[15] S. Doppalapudi, R. G. Qiu, and Y. Badr, "Lung cancer survival period prediction and understanding: Deep learning approaches," *International Journal of Medical Informatics*, vol. 148, p. 104371, 2021. https://doi.org/10.1016/j.ijmedinf.2020.104371

[16] L. A. Vale-Silva and K. Rohr, "Long-term cancer survival prediction using multimodal deep learning," *Scientific Reports*, vol. 11, p. 13505, 2021. https://doi.org/10.1038/s41598-021-92799-4

[17] H. Alkhadar, M. Macluskey, S. White, I. Ellis, and A. Gardner, "Comparison of machine learning algorithms for the prediction of five-year survival in oral squamous cell carcinoma," *Journal of Oral Pathology & Medicine*, vol. 50, no. 4, pp. 378–384, 2021. https://doi.org/10.1111/jop.13135

[18] P. Radhika, R. A. Nair, and G. Veena, "A comparative study of lung cancer detection using machine learning algorithms," in *IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, India, 2019, pp. 1–4. https://doi.org/10.1109/ICECCT.2019.8869001

[19] M. Behring, K. Hale, B. Ozaydin, W. E. Grizzle, S. O. Sodeke, and U. Manne, "Inclusiveness and ethical considerations for observational, translational, and clinical cancer health disparity research," *Cancer*, vol. 125, no. 24, pp. 4452–4461, 2019. https://doi.org/10.1002/cncr.32495

[20] R. Nooreldeen and H. Bach, "Current and future development in lung cancer diagnosis," *International Journal of Molecular Sciences*, vol. 22, no. 16, p. 8661, 2021. https://doi.org/10.3390/ijms22168661

[21] C. Quantin *et al.*, "Variation over time of the effects of prognostic factors in a population-based study of colon cancer: Comparison of statistical models," *American Journal of Epidemiology*, vol. 150, no. 11, pp. 1188–1200, 1999. https://doi.org/10.1093/oxfordjournals.aje.a009945

[22] E. B. Cone *et al.*, "Assessment of time-to-treatment initiation and survival in a cohort of patients with common cancers," *JAMA Network Open*, vol. 3, no. 12, p. e2030072, 2020. https://doi.org/10.1001/jamanetworkopen.2020.30072

[23] A. Hazra, N. Bera, and A. Mandal, "Predicting lung cancer survivability using SVM and logistic regression algorithms," *International Journal of Computer Applications*, vol. 174, no. 2, pp. 19–24, 2017. https://doi.org/10.5120/ijca2017915325

[24] V. Krishnaiah, G. Narsimha, and N. S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," *International Journal of Computer Science and Information Technologies*, vol. 4, no. 1, pp. 39–45, 2013.

[25] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Louisville, KY, USA, 2018, pp. 632–637. https://doi.org/10.1109/ISSPIT.2018.8642753

[26] B. Ma, G. Yan, B. Chai, and X. Hou, "XGBLC: An improved survival prediction model based on XGBoost," *Bioinformatics*, vol. 38, no. 2, pp. 410–418, 2022. https://doi.org/10.1093/bioinformatics/btab675

[27] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Systems with Applications*, vol. 164, p. 113981, 2021. https://doi.org/10.1016/j.eswa.2020.113981

[28] F. A. Altuhaifa, K. T. Win, and G. Su, "Predicting lung cancer survival based on clinical data using machine learning: A review," *Computers in Biology and Medicine*, vol. 165, p. 107338, 2023. https://doi.org/10.1016/j.compbiomed.2023.107338

[29] A. Safiyari and R. Javidan, "Predicting lung cancer survivability using ensemble learning methods," in *Intelligent Systems Conference (IntelliSys)*, London, UK, 2017, pp. 684–688. https://doi.org/10.1109/IntelliSys.2017.8324368

[30] G. Nath *et al.*, "An interactive web-based tool for predicting and exploring brain cancer survivability," *Healthcare Analytics*, vol. 3, p. 100132, 2023. https://doi.org/10.1016/j.health.2022.100132

[31] A. J. Didier, A. Nigro, Z. Noori, M. A. Omballi, S. M. Pappada, and D. M. Hamouda, "Application of machine learning for lung cancer survival prognostication—A systematic review and meta-analysis," *Frontiers in Artificial Intelligence*, vol. 7, p. 1365777, 2024. https://doi.org/10.3389/frai.2024.1365777

[32] Y. Kumar, S. Gupta, R. Singla, and Y.-C. Hu, "A systematic review of artificial intelligence techniques in cancer prediction and diagnosis," *Archives of Computational Methods in Engineering*, vol. 29, pp. 2043–2070, 2022. https://doi.org/10.1007/s11831-021-09648-w

8 AUTHORS

Rooshan Ghous's background is in healthcare, having participated in various research and administrative projects both in New Zealand and overseas. She trained as a dentist from Pakistan shortly after which she moved to NZ and stepped into public health. She is currently pursuing a Master's in Data Science. Her research focuses on the exciting potential of artificial intelligence (AI) within healthcare, specifically its applications in cancer research.

Dr Seyed Ebrahim Hosseini has extensive teaching and research experience at the Universities and PTE sectors. Curriculum development experience as subject

matter expert for BSc & MSc, NZQA Level 5/6/7/8 IT and Networking as per NZQA standards. He has authored over 55 international research publications in premier conferences and peer-reviewed journals. Seyed has also been actively working with international industry leaders including Cisco and Microsoft, he has also participated in international research projects in regards to social media technologies for elderly people. His specialisations are: Network and System Engineering, Systems Administration, IT Infrastructure planning, designing, implementation & management, Communication & Networks, Wireless Mesh Networks, Operating Systems, Cloud Computing, and Project Management (E-mail: seyedh@whitecliffe.ac.nz).

Shahbaz Pervez Chattha is a Professor at Whitecliffe, New Zealand, and a distinguished ICT leader with extensive experience across international organizations. He has authored over 60 international research publications in premier conferences and peer-reviewed journals. With 15+ active international certifications from industry leaders such as Cisco, ISACA, EC-Council, and Microsoft, he brings a wealth of expertise in cybersecurity, AI, Machine learning, networking, and ICT governance. Dr. Chattha has held senior technical and managerial positions in both the corporate sector and higher education institutions. He is a Certified Master Trainer from Cisco Networking Academy USA for expert-level professional training and certifications, as well as a Certified Master Trainer from EC-Council USA specializing in Information and Cyber Security certifications and training. He has secured multiple research projects and travel grants. His work includes the successful design, deployment, and management of ICT projects, with a strong focus on smart cities infrastructure and services.