Computers in Biology and Medicine 165 (2023) 107338

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Predicting lung cancer survival based on clinical data using machine learning: A review

Fatimah Abdulazim Altuhaifa [a,b,*], Khin Than Win [a], Guoxin Su [a]

[a] School of Computing and Information Technology, University of Wollongong, NSW, 2500, Australia
[b] Saudi Arabia Ministry of Higher Education, Riyadh, Saudi Arabia

ARTICLE INFO

Keywords:
Data mining
Machine learning
Artificial intelligence
Lung cancer
Survival prediction
Feature selection

ABSTRACT

Machine learning has gained popularity in predicting survival time in the medical field. This review examines studies utilizing machine learning and data-mining techniques to predict lung cancer survival using clinical data. A systematic literature review searched MEDLINE, Scopus, and Google Scholar databases, following reporting guidelines and using the COVIDENCE system. Studies published from 2000 to 2023 employing machine learning for lung cancer survival prediction were included. Risk of bias assessment used the prediction model risk of bias assessment tool. Thirty studies were reviewed, with 13 (43.3%) using the surveillance, epidemiology, and end results database. Missing data handling was addressed in 12 (40%) studies, primarily through data transformation and conversion. Feature selection algorithms were used in 19 (63.3%) studies, with age, sex, and N stage being the most chosen features. Random forest was the predominant machine learning model, used in 17 (56.6%) studies. While the number of lung cancer survival prediction studies is limited, the use of machine learning models based on clinical data has grown since 2012. Consideration of diverse patient cohorts and data pre-processing are crucial. Notably, most studies did not account for missing data, normalization, scaling, or standardized data, potentially introducing bias. Therefore, a comprehensive study on lung cancer survival prediction using clinical data is needed, addressing these challenges.

1. Introduction

Lung cancer is the leading cause of cancer-related deaths, accounting for approximately 1.8 million deaths in 2020 [1,2]. Additionally, it accounted for the highest number of disability-adjusted life years (DALYs), affecting 40.9 million individuals in 2017. The cancer survival period refers to the period between diagnosis and death as a result of cancer [3]. Moreover, survival analysis is valuable for clinicians, researchers, patients, and policymakers.

Developing an accurate and robust model is crucial for identifying lung cancer survival rates [4]. Various ML algorithms have been developed for clinical applications, including random forests (RFs), ensemble algorithms, Naive Bayesian (NB) classifiers, Support vector machines (SVMs), neural networks (NNs), Decision Trees (DTs), and a number of proprietary algorithms [5]. It is possible to link various clinical attributes of cancer patients to their survival rates using ML techniques. Additionally, ML has the advantage of reducing the workload of medical practitioners and the risk of human mistakes. The high performance of ML has made it an exciting and motivating tool for healthcare providers. ML techniques allow the development of predictive models based on cancer data to predict survival. Despite this, no current technique qualifies for application to a specific dataset [6].

This review aims to explore studies that employ these techniques to predict lung cancer survival based on clinical data. It provides an overview of the survival prediction processes used, including the data sources, patient cohort, data preprocessing, feature selection methods, selected features, ML algorithms, and validation methods (see Table 1).

2. Methods

To identify the search terms used for "Lung cancer survival using machine learning" search queries. We first identified a related topic, "Predicting lung cancer survival using machine learning and clinical data". Thereafter, we split the topic into five terms: "predict," "lung cancer," "survival," "machine learning", and "clinical data" and searched for their synonyms. For all these processes, we used the University of Wollongong Library Search Words Generator for finding the synonyms and Medical Subject Headings thesaurus (MeSH) tool. The search

* Corresponding author. School of Computing and Information Technology, University of Wollongong, NSW, 2500, Australia.
E-mail address: faaa272@uowmail.edu.au (F.A. Altuhaifa).

https://doi.org/10.1016/j.compbiomed.2023.107338
Received 10 May 2023; Received in revised form 31 July 2023; Accepted 7 August 2023
Available online 9 August 2023

Table 1
Statement of significance.

Problem	Reviewing articles about machine learning model that predicts lung cancer survival based on clinical data.
What is Already Known?	Although researchers have explored data mining and machine learning for predicting survival, to date there is no such technique effective at pre-processing data and accurately predicting lung cancer survival.
What This Paper Adds?	In this paper, we review and explore studies using machine learning to predict lung cancer survival based on clinical data. It provides an overview of data sources, patient cohort, data pre-processing, feature selection algorithms, selected features, ML algorithms, and validation methods.

algorithms were implemented with assistance from a librarian. Three digital databases were searched: MEDLINE, Scopus, and Google Scholar. The search was conducted from April 02, 2022–July 05, 2022. The search was based on the keywords "lung cancer," lung neoplasms", "lung tumor", "carcinoma," "survival," "clinical," "machine learning," "deep learning," "algorithm," and" artificial intelligence." The search algorithms used in this systematic review are presented in (see Table 2).

The search of the three digital databases involved screening the article titles. We excluded studies whose titles indicated that different cancer than lung cancer was addressed, or an image-based database was used. Three independent reviewers evaluated the selected studies using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines through COVIDENCE [7,8].

Four inclusion criteria were employed for the papers: (1) peer-reviewed; (2) published between 2000 and 2022; (3) based on modeling lung cancer survival rates using ML algorithms (models that learn from data without utilizing traditional programming techniques and focus on predicting survival rates rather than calculating confidence intervals) [9]; and (4) published in English. Additionally, three exclusion criteria were employed: (1) not using an ML model; (2) published before 2000; (3) and utilizing computed tomography scan (CT) and image databases.

The reviewers screened the titles and abstracts of the potentially relevant studies by using COVIDENCE and any opinion differences were resolved through meetings. Following the abstract screening and full review of each study, multiple meetings were conducted among the authors to identify the concepts and data to be extracted for the systemic review. The data from each study is extracted using the NVivo system. For the analysis, the studies were categorized as follows: Surveillance, Epidemiology, and End Results (SEER) databases, other databases, survival time periods (one, two, three, and five years), and used accuracy (Area under the ROC Curve (AUC), Root mean square error (RMSE), and concordance statistic (C-statistic)).

Table 2
Search algorithms.

Digital databases	Search algorithms
MEDLINE	(TI (patient* AND ("lung cancer" OR "lung neoplasms" OR "lung tumor" OR "lung adenocarcinoma") OR (MM "Lung Neoplasms+")) OR AB (patient* AND ("lung cancer" OR "lung neoplasms" OR "lung tumor" or "lung adenocarcinoma") OR (MM "Lung Neoplasms+")) OR SU (patient* AND ("lung cancer" OR "lung neoplasms" OR "lung tumor" or "lung adenocarcinoma") OR (MM "Lung Neoplasms+"))) AND (TI ("survival rate" OR "survival outcomes" OR "clinical outcomes") OR AB ("survival rate" OR "survival outcomes" OR "clinical outcomes") OR SU ("survival rate" OR "survival outcomes" OR "clinical outcomes")) AND (TI ("artificial intelligence" OR "AI" OR "machine learning" OR algorithm*) OR AB ("artificial intelligence" OR "AI" OR "machine learning" OR algorithm*) OR SU ("artificial intelligence" OR "AI" OR "machine learning" OR algorithm*)) ((MH "Carcinoma, Non-Small-Cell Lung") OR (MH "Lung Neoplasms") OR (MH "Cell Adhesion Molecule-1")) AND (machine learning OR artificial intelligence or deep learning or neural network or machine learning in cancer) AND (survival rate or survival outcomes or clinical outcomes)
Scopus	TITLE-ABS-KEY (*"patient* "* AND *"lung cancer"* AND *"survival*"* AND *"machine learning")* AND (LIMIT-TO (SUBJAREA, *"COMP")*) TITLE-ABS-KEY (*"patient* "* AND *"lung cancer"* AND *"survival*"* AND *"machine learning")*
Google Scholar	"patient" AND "lung cancer survival" AND "machine learning"

3. Results

From the digital databases, 2506 studies were found: 1248 from MEDLINE, 420 from Scopus, and 838 from Google Scholar. Among these, 2084 were excluded because their titles did not meet the inclusion criteria, and 422 were retained and uploaded to COVIDENCE. Thereafter, duplicates were removed by COVIDENCE, and 71 studies remained, out of which 26 were excluded after their titles and abstracts were screened: seven (approximately 26%) did not employ ML algorithms, thirteen (approximately 50%) relied on image-based datasets, two (approximately 10%) had the wrong objectives, one (approximately 4%) was a preprint, and two (approximately 10%) examined different cancers. After reviewing the full texts of the remaining 45 studies, 15 were excluded: 6 (approximately 40%) did not employ ML algorithms, and 9 (approximately 60%) relied on image-based datasets, 30 studies were included in this systematic review (Fig. 1).

All studies were published between 2012 and 2023. However, none were conducted between 2013 and 2016. Furthermore, the interest in the development of ML algorithms for lung cancer survival based on clinical data increased between 2019 and 2023. Nine studies were published in 2020, which represented a peak in the number of published studies. Additionally, while extracting information in Nvivo, six more aspects related to this systemic review emerged: (1) patient cohort, which is a group of individuals with the same lung cancer histology, treatment, stage, or metastases; (2) data preprocessing, which is subdivided into missing value processes, data transformation, and data conversion; (3) feature selection, which contains studies describing the feature selection process; (4) selected features, which contain the features used as predictors in each study; (5) ML, which is subdivided into ML algorithms and the best models in each study; (6) and validation methods, which include information regarding the validation method used in each study, including k-fold cross-validation or data splitting.

3.1. Data collection

A review revealed that 13 used the US-SEER database and 2 used data from the Cancer Registry. Moreover, one study combined SEER data with local hospital data (Indian Medical Hospitals) [27]. Furthermore, four used data from local hospitals, and two did not identify the data source. Additionally, five used data from local hospitals for external validation. However, one indicated that the authors were unable to obtain external validation data that were similar to the SEER data [10]. The datasets in the 30 studies ranged in size from 150 to 57,254 patients. Even though several studies used the same dataset, such as SEER, the sizes of the datasets differed. These differences in dataset sizes can be attributed to differences in patient cohorts. Further details regarding the data characteristics of each study are presented in Table 3.

To use the vast amount of medical data produced via clinical data, it is imperative to select an appropriate patient cohort for the specific research question being investigated [40]. In each of the 30 studies, a specific patient cohort was included: eight (26.6%) selected patients with non-small cell lung cancer (NSCLC) [13,15,18,19,25,30,32,37]. [10,26,27,29,34,35] selected only patients with adenocarcinoma, whereas [22] selected patients with adenocarcinoma and squamous cell carcinoma in addition to one malignant primary lesion. In contrast [15, 36], selected patients with bone metastases. 10 (33.3%) studies selected cancer patients with different stages: stage I [21], stage III [24,25,37], stage I–IIIB [15–17], stages I–IV [10,22], and stages IB–IIA [32].

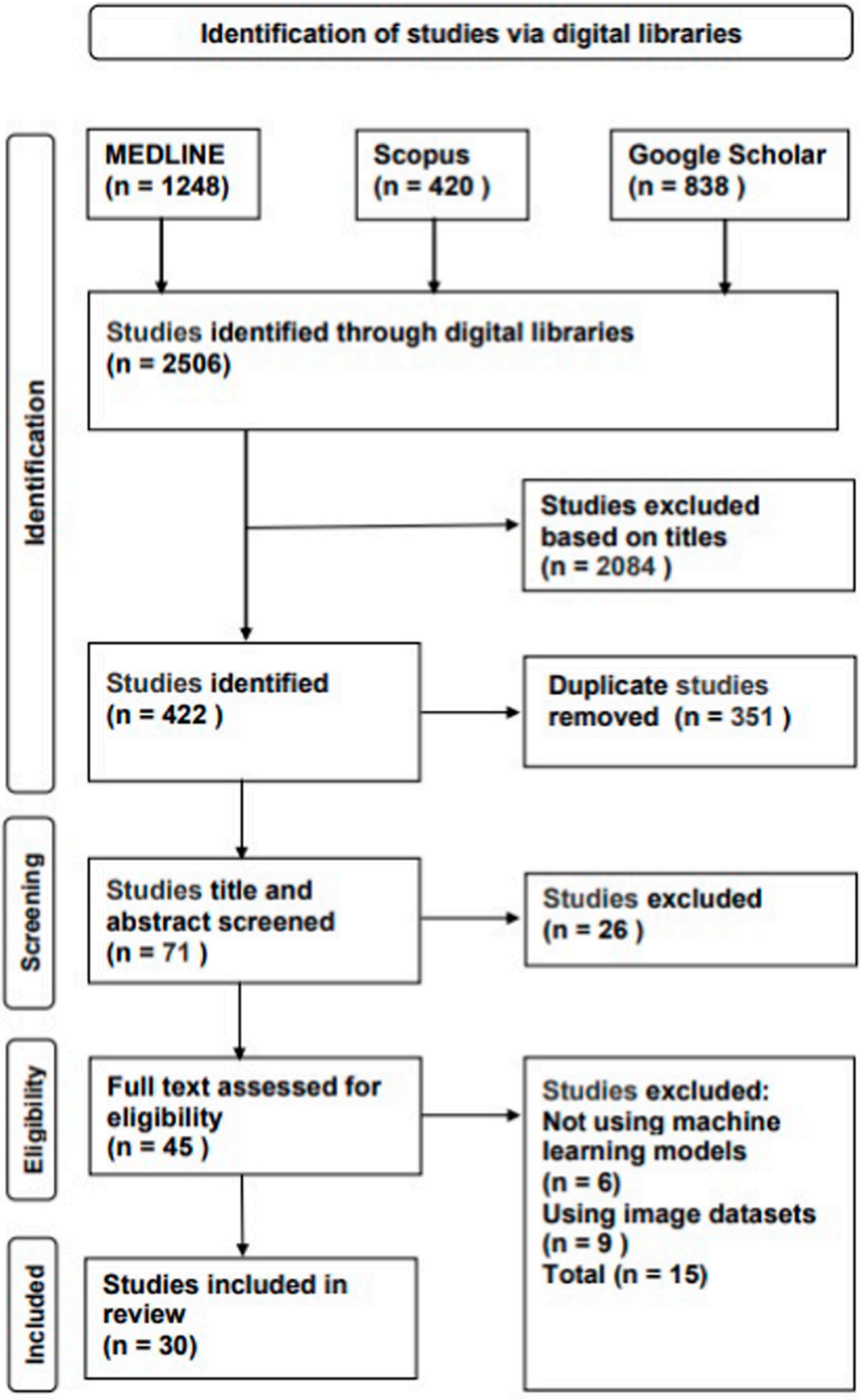

Fig. 1. PRISMA flowchart of the study selection process.

Table 3
Data characteristics.

Reference	Source of the data	Time	Size of the data	Country	Source availability	External validation
[11]	SEER November 2008	1998–2001	57,254	The USA	https://seer.cancer.gov/	Not applicable
[12]	SEER	2004–2009	10,442	The USA	https://seer.cancer.gov/	Not applicable
[13]	Cancer data	Not reported	239	Netherlands, UK, and the USA	https://www.cancerdata.org/publication/developing-and-validating-survival-prediction-model-nsclc-patients-through-distributed	Not applicable
[14]	SEER November 2018	Only 2014	Not reported	The USA	https://seer.cancer.gov/	Not applicable
[10]	SEER	2010–2015	50,687	The USA	https://seer.cancer.gov/	Not applicable
[15]	SEER	2010–2015	5973	The USA	https://seer.cancer.gov/	Affiliated Hospital of Chengde Medical University (AHOCMU)
[16]	SEER	2004–2009	10,442	The USA	https://seer.cancer.gov/	Not applicable
[17]	SEER	2004–2009	10,442	The USA	https://seer.cancer.gov/	Not applicable
[18]	Chinese Multi-institutional Registry (CMIR)	Not reported	5123	China	Not reported	Not applicable
[19]	Cancer data	Not reported	509	Not reported	https://www.cancerdata.org	Not applicable
[20]	SEER	2006–2011	683	The USA	https://seer.cancer.gov/	Not applicable
[21]	SEER and Indian Medical Hospitals	Not reported	321	The USA	https://seer.cancer.gov/	Not applicable
[22]	SEER	2010–2015	17,322	The USA	https://seer.cancer.gov/	Shanghai Pulmonary Hospital
[23]	Not reported	Not reported	809		Not reported	Not applicable
[24]	Patients at Eskişehir Osmangazi University Faculty of Medicine.	2007–2018	585	Turki	Not reported	Not applicable
[25]	SEER	Not reported	16,613	The USA	https://seer.cancer.gov/	Shandong Cancer Hospital and Institute
[26]	The Cancer Genome Atlas (TCGA)	Not reported	371	The USA	https://www.cancer.gov/ccg/research/genome-sequencing/tcga	Not applicable
[27]	(TCGA)	Not reported	291	The USA	https://www.cancer.gov/ccg/research/genome-sequencing/tcga	Not applicable
[28]	Medical canters in China	2018–2021	150	China	Not reported	Medical canters in China
[29]	(TCGA)	Not reported	1563	The USA	https://www.cancer.gov/ccg/research/genome-sequencing/tcga	Not applicable (TCGA)
[30]	National Center for Biotechnology Information (NCBI) GEO database	Not reported	704	The USA	https://www.ncbi.nlm.nih.gov/geo/	MTAB-923 dataset
[31]	SEER	2004–2009	38,262	The USA	https://seer.cancer.gov/	Not applicable
[32]	Not reported	2005–2018	1137	The USA	Not reported	SEER
[33]	SEER	2998–2001	17,484	The USA	https://seer.cancer.gov/	Not applicable
[34]	TCGA cohort GSE72094 cohort	Not reported	506	The USA	https://www.cancer.gov/ccg/research/genome-sequencing/tcga	GSE72094 and GSE11969 cohorts
[35]	TCGA	Not reported	739	The USA	https://www.cancer.gov/ccg/research/genome-sequencing/tcga	GEO and ICGC mutation dataset
[36]	SEER	2000–2019	10,001	The USA	https://seer.cancer.gov/	Not applicable
[37]	SEER 18 Registries	2000–2017	4517	The USA	https://seer.cancer.gov/	Shaanxi Provincial People's Hospital, China
[38]	TCGA and GEO	Not reported	2166	The USA	https://www.cancer.gov/ccg/research/genome-sequencing/tcga	GSE42127 and GSE37745 data
[39]	SEER	2000–2015	28,458	The USA	https://seer.cancer.gov/	Not applicable

Moreover, four (13.3%) selected patients that had undergone treatment [13,18,23,25]. Moreover [24], selected patients without a diagnosis of distant metastases, and [14] selected patients who had undergone the first course of treatment.

3.2. Data preprocessing

Among the 30 studies, only twelve (40%) reported the processes used for handling missing values, which ranged from removing the features with missing values to imputation. Six handled the missing values by removing the rows with missing values [13,22,25,27,34,37]. [13] stated that the mean binned values were only estimates; therefore, they could not be completely relied upon for training the ML model. In contrast, four adopted imputed methods for handling missing values [19,23,28,33]. Specifically [19,33], imputed missing values using median/mode and mean imputation, whereas [23,28] imputed them based on information provided by some critical candidates in the dataset [14]. applied both methods (removing missing values and imputation) to handle missing values and removed features with over 20% of missing values. If the missing values in a feature were less than 20%, they were imputed using median/mode imputation. The last study [18] employed the ReliefF algorithm to address the problem of missing values during the feature selection process. Only seven (23.3%) studies described the processes used for data transformation, which varied among normalization, scaling, and standardization. Among these seven studies, four used normalizations [13,20,37,39]. In Ref. [13] normalization was performed using the min–max method, whereas in Ref. [20] the numerical values were normalized in the range of −1 to 1 [37,39]. did not report the method of normalization. Alternatively, the scaling variable method was used in Refs. [14,17]. Using scaling variables [17], prevented overly sensitive models from being affected by skewed distance measures. To speed up the training process [15], adopted a standardization method for age and size. Additionally, data conversion was reported in four (13.3%) studies; they adopted one-hot encoding [14,15,

37,39], which allows for calculating distance more accurately. In contrast [17], did not adopt one-hot encoding because the dataset they used contained sparsely populated data, and to minimize the amount of preprocessing required.

3.3. Features selection and engineering

In total, 19 (approximately 63.3%) studies considered feature selection during data preprocessing, including ANOVA, ReliefF, least absolute shrinkage and selection operator (LASSO), random survival forest, extreme gradient boosting (XGBoost), Cox proportional-hazards model, chi-squared test, *t*-test, correlation matrix, and integrated gradients [10,12–16,18,19,23–25,27,29,30,33,34,37–39]. In Ref. [12], features were selected using ANOVA, a two-sided confidence interval measuring 95% confidence, and any feature not reported by ANOVA was eliminated. Among the 19 studies that employed feature selection methods, two adopted ReliefF, which reduces dimensionality and makes the data more interpretable [13,18]. While [27], selected Relief [14]. adopted LASSO. In Ref. [10], the RF method was used to assess the feature importance and the relationship between the number of trees and the error rate. In Refs. [15,39], XGBoost was used to estimate the feature importance. In Refs. [16,29,30,34,38], the Cox proportional-hazards model was adopted, whereas in Ref. [23], chi-squared and *t*-test methods were adopted. Finally, four studies adopted correlation matrix methods [19,24,25,33] and one study [37] adopted Integrated gradients. (Table 4).

Among the 30 studies, 32 features were selected to predict lung cancer survival. The number of features selected in each study varied between 4 and 15. Age (24 studies (80%)), sex (21 studies (70%)) and N stage (16 studies (53.3%)) were the most selected features, followed by stage (15 studies (50%)) and T stage (13 studies (43.3%)) [11]. reported that 11 features (birthplace, age, grade, surgery, radiation therapy, N stage, stage, malignancy, tumors, number of primaries, and histology) encoded the information reasonably, which prevented any substantial accuracy loss. Moreover, birthplace appeared to be a crucial feature among the 11 features [14]. discovered that some features (sex, surgery, radiation therapy, tumor size, and N stage) impact local prediction; however, the global level is not significantly affected. This implies that the impact of these features is primarily determined by their interactions with some other features [15]. found that tumor size was the most significant survival factor, followed by age, chemotherapy, liver metastases, surgery, grade, and race, which moderately influenced survival. Moreover, N stage, histologic type, and sex negligibly influenced survival [18]. found that age, sex, and histology are useful and meaningful features for predictions, and reduce computation costs and enhance analysis accuracy. In contrast [19], found a strong correlation between the total tumor dose and overall treatment time (Table 5a and Table 5b). Moreover, five studies reported that the most significant features for predicting survival time are surgery, stage [10,11], histology type [28], T stage [15], treatment time [24], total tumor size [19], and N lymph nodes [24]. [21] found that hemoglobin level in conjunction with TNM stage 1 can aid in predicting lung cancer survival [24]. identified T stage, surgery, lymph node site, and histopathological factors as crucial features affecting survival time. Furthermore, 7 studies have identified genetic features to predict lung cancer survival, including TP53 [26,29,30,34], KRAS [26,27], EGFR, PTEN, AURKA, AURKB, CDKN2A, ERBB2, MDM2, RB1, NFKB1 [26,27,34], XIAP [27,29], BRCA1, BRCA2, PIK3CA, SMAD4, STK11, NF1, ATM, IDH1, IDH2, TET2, ALK [26], ALDH1A1, APC, BAX, BIRC5, CASP3, EIF4A2, ESR1, FADD, FOXO1, TP53 [27], ACTR2, ALDH2, FBP1, HIRA, ITGB2, MLF1, P4HA1, S100A10, S100B, SARS, SCGB1A1, SERPIND1, STAB1 [29], EPCAM, HIF1A, PKM, PTK7, ALCAM, CADM1, SLC2A1 [30], ESR1, FADD, FOXO1, AKT1, BIRC5 [34], RHOV, CSMD3, FBN2, MAGEL2, SMIM4, BCKDHB, and GANC [35].

Table 4
Feature selection algorithm.

Features selection method	References
ANOVA	[12]
ReliefF	[13,18]
Relief	[27]
LASSO	[14]
Random survival forest method	[10]
XGBoost	[15,39]
Cox proportional hazards model	[16,29,30,34,38]
Chi-square and *t*-test	[23]
Correlation matrix	[19,24,25,33]
Integrated gradients	[37]

3.4. ML algorithms

Forty-seven different ML algorithms were used in the 30 studies to predict lung cancer survival (Table 6). However, the prediction varied from six months to five years in each study. Overall, RF was the most used model, accounting for 17 (36.1%) of 47 models. It was used in 17 (56.6%) of the 30 studies [10–12,14–16,18,19,23–25,27,33–35,38,39]. Some models, such as hierarchical clustering [17] and LogitBoost [11], were used in only one study. Moreover, 3 studies (10%) used 10 different ML algorithms to predict lung cancer survival at 6 months [11,12,14], 6 (20%) used 14 different ML algorithms to predict survival at 1 year [10,11,15], 6 (20%) used 22 different ML algorithms to predict survival at 2 years [11–14,19,24], 4 (13.3%) used 13 different ML algorithms to predict survival at 3 years [10,22,26,27,29,39], and 22 (73.3%) used 39 different ML algorithms to predict survival at 5 years [10–12,14,16–18,20–24,29–37,39].

3.5. Model evaluation and performance metrics

Among the 30 studies, 12 (40%) used k-fold cross-validation, 16 (53.3%) employed data splitting, and 4 (13.3%) did not report the validation method (Appendix B). In two of these studies, both k-fold cross-validation and data splitting were implemented [23,24]. K-fold cross-validation varied between 3- and 10-cross validations and data splitting ratios also varied (8:2, 9:1, and 7:3). The primary objective of k-fold cross-validation in Ref. [16] was to reduce overfitting rather than parameter optimization [24]. employed k-fold cross-validation to determine whether the trained model could generalize new data, and to identify any selection bias or over-compliance issues. In Ref. [23], data splitting was adopted to avoid model overfitting, which was necessary to accurately estimate model performance.

The accuracies of the ML algorithms varied across the 30 studies, including the RMSE (7 studies (23.3%)), AUC (20 studies (60.6%)), and the C-statistic (4 study (13.3%)). These metrics provide insights into different aspects of model performance. RMSE (Root Mean Square Error), which measures the average difference between predicted and actual values, is commonly used in regression tasks. This measure quantifies the prediction error and is commonly used to determine whether regression models are accurate [41] The AUC (Area Under the ROC Curve) is a commonly used metric in binary classification tasks. It measures the area under the Receiver Operating Characteristic (ROC) curve in to assess the classifier's ability to differentiate between positive and negative samples. In general, a higher AUC indicates better discrimination performance [42]. A C-statistic is a metric used in survival analysis, also known as the concordance index or the area under the ROC curve for survival analysis. It evaluates the predictive accuracy of survival models by measuring the ability to rank subjects based on their survival times. A higher C-statistic indicates better predictive performance [43].

Studies using RMSE focus on the accuracy of continuous survival time prediction. These studies typically utilize the time to event and survival time as the target variables, which are continuous numerical variables. The goal is to develop machine learning models that can

Table 5a
Selected features by studies.

Features/Studies	[11]	[12]	[13]	[14]	[10]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]	[25]
Age	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sex				✓	✓	✓			✓	✓		✓	✓	✓	✓	✓
Race						✓										✓
Surgery	✓			✓	✓								✓			✓
Radiation therapy	✓			✓		✓					✓	✓				✓
Chemotherapy					✓	✓										✓
Stage	✓	✓			✓		✓	✓			✓					
Malignant	✓											✓				
Number of primaries	✓	✓						✓			✓					
Histology	✓					✓			✓				✓		✓	✓
Tumor size		✓		✓		✓	✓	✓				✓	✓		✓	
Grade	✓	✓				✓	✓	✓			✓		✓			
T stage		✓	✓		✓	✓					✓	✓	✓	✓	✓	✓
N stage	✓	✓	✓		✓	✓	✓				✓	✓	✓	✓	✓	✓
M stage (lung, liver, brain metastases)					✓	✓	✓				✓	✓	✓		✓	
Laterality						✓										✓
Total tumor dose			✓							✓						
Tumor load			✓													
Overall treatment time			✓							✓						
Primary site						✓					✓		✓			✓
Genetic features																
Insurance status						✓										
Marital status						✓								✓		
Hemoglobin level												✓				
Red blood cell count												✓				
Immune system for lung cancer patients												✓				
Marriage status													✓			
Educational level														✓		
Monthly family income														✓		
Body mass index														✓		
Physical activity														✓		
Birthplace	✓															

Table 5b
Selected features by studies.

Features/Studies	[26]	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34]	[35]	[36]	[37]	[38]	[39]
Age			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Sex			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Race											✓			✓
Surgery								✓				✓		
Radiation therapy												✓		
Chemotherapy											✓	✓		
Stage	✓			✓	✓	✓		✓		✓		✓	✓	✓
Malignant														
Number of primaries														
Histology						✓	✓					✓		✓
Tumor size							✓		✓			✓		
Grade								✓						
T stage										✓	✓		✓	
N stage								✓		✓	✓		✓	
M stage (lung, liver, brain metastases)			✓							✓	✓		✓	
Primary site							✓					✓		
Genetic features	✓	✓		✓	✓				✓	✓			✓	
Body mass index			✓											

accurately predict the exact duration of survival for individual patients. The RMSE metric is used to assess the performance of these models by measuring the average difference between the predicted survival time values and the actual survival time values. A lower RMSE indicates better accuracy in predicting continuous survival time. In contrast, studies that use AUC evaluate the model's ability to classify patients into different survival categories based on a specific time frame, such as whether a patient survives for 1 year, 2 years, 3 years, or 5 years. In these studies, the target variable is transformed into a binary variable, indicating whether a patient belongs to a specific survival category (e.g., survived for a certain period or did not survive beyond that period). The AUC (Area Under the Curve) metric is used to assess the model's ability to distinguish between different survival categories. It evaluates the ability of the model to rank the survival probabilities of patients correctly and to classify patients into different survival groups based on predetermined time intervals. Furthermore, the study uses C-statistics to evaluate the ability of the model to rank patients based on their survival times. Throughout the study, the target variable has remained as the time to event or survival time, which represents the time until death occurs. A C-statistic, also known as a concordance index or area under the ROC curve, is used to evaluate the model's discriminatory power in ranking patients based on their predicted survival times. Essentially, it is a measure of the model's capability to order patients at the correct time based on their actual survival rates.

An overview of the best ML algorithms used in each study for predicting lung cancer survival is presented in Appendix A, including their accuracies, survival times, and dataset sizes. An overview of the ML algorithms used in the 16 studies to predict lung cancer survival,

Table 6
ML types and algorithms.

ML types	ML algorithms
Tree-based	16 RF, 1 IRF, 1 J48 DT, 5 DT, 1 AdaBoost, 2 gradient boost tree (GBT), 3 XGBoost, 1 light gradient boost (Light GBT), 1 logitBoost, 1 bagging, 1 Random subspace.
Neural Network-based	3 artificial neural network (ANN), 2 single perceptron neural network (SPNN), 1 multilayer neural network (MNN), 1 convolution neural network (CNN), 1, Recurrent neural networks (RNN), 1, generalized regression neural network (GRNN), 5 Deep Learning (DL), 1 Multilayer perceptron (MLP), 1 deep neural network (DNN).
Bayesian	4 naïve bayes (NB), 1 Gaussian K-base NB, 1 Gaussian NB, 2 Bayes Net.
Linear-based	8 support vector machine (SVM), 3 linear regression (LR), 3 logistic regression, 1 multitask logistic regression (MTLR), 1 Ridge Regression, 1 Line support vector regression (SVR), 2 Least Absolute Shrinkage and Selection Operator (LASSO).
Clustering-based	1 hierarchical clustering (HC), 1 model based clustering (MBC), 1 K means, 1 non-negative matrix factorization, 1 self-organizing map, 1 support rector cluster (SVC).
Other	1 lazy classifier (LWL), 1 meta-classifier (ASC), 1 Rule Learner (OneR), 2 Skeletal Oncology Research Group (SORG), 1 k-Nearest Neighbors (KNN), 4 cox r egression, 1 cox neural network (Cox NN).

including their average accuracies and the accuracy reported in each study, is presented in Appendix C.

4. Discussion

This review was conducted to summarize the applications of data mining and ML methods in predicting lung cancer survival using clinical data (structured data) to guide future research.

Risk of bias was assessed using the prediction model risk of bias assessment tool (PROBAST). The PROBAST tool is designed to assess prognostic and diagnostic prediction model risk of bias and applicability (see Appendix D) [44]. A PROBAST evaluated the quality of the machine learning models developed and validated for predicting lung cancer survival. Overall, the quality of the included studies in this review is moderate, and further well-designed studies are needed to validate machine learning predictions of lung cancer survival. In particular, models need to be improved in pre-processing, as sample size and missing data sections are unclear in 17 (56.6% of the studies), and high risk in 5 (16.6% of the studies).

According to the results of this review, the interest in research related to lung cancer survival using ML algorithms based on clinical datasets has increased; it was the highest in 2020 when nine studies were published. However, the number of studies from 2023 is limited because we conducted this review in 2022–2023 and could only extract studies that were published before July 2023. Moreover, the number of studies related to "Predicting lung cancer survival using machine learning and clinical data" was low because most studies used image-based datasets and very few researchers have conducted research on lung cancer. This is consistent with the findings of [6,45], which indicated that most studies employ image-based datasets and primarily focus on breast cancer. This could be because breast cancer is the most common type of cancer. However, lung cancer is the leading cause of cancer-related deaths [1]. Thus, the annual number of publications predicting lung cancer survival using clinical data is expected to increase in the future.

4.1. Data collection

The results of this review also indicate that most studies (approximately 43.3%) used primarily US-based SEER datasets. Researchers are expected to employ SEER data because of availability. However, the results of this review indicate that ML algorithms based on private datasets perform almost as well as those based on SEER data. However, local hospital datasets comprise a smaller sample size than public datasets. Additionally, the minimum sample size used in the reviewed studies was 150, and thirteen studies used datasets that include less than 1000 patients. Using small datasets for model training results in overfitting and reduced the model's generalizability to new data when it is used with unsuitable models [46,47]. Future studies should use appropriate ML algorithms for small datasets, such as NB or RF [48,49]. Each SEER data differs in geography, months in dates, some demographic fields, and treatment information such as chemotherapy. As a result, each study in this review uses data for training its model that differs from the others even though they are using SEER data [50]. This review also found that most selected patients with lung cancer in general without specifying any type of lung cancer. However, eight (26.6.2%) studies specifically targeted NSCLC patients. This may be because 84% of lung cancer cases worldwide are NSCLC [51]. Additionally, this review discovered that no studies have employed ML based on clinical data for survival prediction of patients with small lung cancers. Moreover, no studies have targeted specific age.

In accordance with the research article by Ref. [11], surgical interventions and radiation therapy are the two treatment options available for their study. Based on the findings of the study, the most appropriate treatment approach for lung cancer patients depends on factors such as the stage of the disease and the patient's general health. A significant finding of the study indicates that patients who underwent surgical intervention had a higher survival rate than those who did not. Despite this, the study does not provide explicit guidance regarding the precise application of surgery or radiation therapy according to the stage of lung cancer. According to the [14] patients who underwent surgical intervention had a higher response to a longer survival period compared to those who did not have surgery. This suggests that surgery plays a significant role in improving the chances of survival for lung cancer patients. Despite the study's acknowledgment of the importance of surgical interventions and their association with better survival rates, it does not offer explicit guidance regarding the precise application of surgery or radiation therapy based on the lung cancer stage.

These findings of the study [10] suggest that patients who received chemotherapy had a higher likelihood of survival compared to those who did not receive chemotherapy. However, the study does not provide specific information about the type of chemotherapy used, the duration of treatment, or the specific patient characteristics that influenced the effectiveness of chemotherapy. These details would require further investigation and analysis. Additionally, the study suggests that it may not be a significant independent risk factor for prognosis in lung adenocarcinoma patients. However, considering the limitations of the study, further research and analysis are needed to draw more definitive conclusions about the impact of surgery on patient outcomes.

The study [15] indicated that chemotherapy had a beneficial effect on patient prognosis in NSCLC patients with BM. However, it is important to note that the specific chemotherapy drugs and protocols used in the treatment were not mentioned in the study, which limited further exploration of the relationship between treatment methods and prognosis. Additionally, it mentions that Radiation therapy was performed in 59.0% of the patients but does not provide specific information about its impact on prognosis or survival outcomes. On the other hand, the studies [20–22,25,28,29] mentioned that the used datasets include radiation therapy or surgery, but it does not provide specific details or findings about the predictive value or impact of radiation therapy or surgery on patient survival.

4.2. Data preprocessing

High-quality data are imperative for the proper operation and performance of ML algorithms [52]. Therefore, it is essential that researchers preprocess the data before applying regression or classification models. Missing values are often encountered when

analyzing and interpreting datasets during the development of an ML model. This review found that more than 50% of studies (approximately 60%) did not report the process used for handling missing data. Although there is no perfect method for handling missing values, it is essential to treat missing data prior to data analysis to avoid bias or misguided conclusions [53]. Thus, data processing should consider missing values. Several algorithms, such as imputation, can be used to handle missing values [54]. There are several imputation methods, including single, regression, maximum/minimum, and multiple imputations, and imputation using ML, such as the KNN model. However, using single imputation methods for high-dimensional or big data may adversely affect the performance of the ML model, and in many cases, lead to data bias [53]. However, two studies [19,23] that used single imputation comprised small data sizes, and another [14] did not include information regarding the size of the data used. Additionally, as indicated in Ref. [55], complicated imputations, such as multiple imputations, are more effective in reproducing missing values. By using multiple imputations, the missing values can be estimated in an unbiased manner, and accurate estimates of standard errors can be obtained [56]. Therefore, using more complex imputation methods for missing values can benefit future studies that employ ML for lung cancer survival prediction [55]. One study [18] used ReliefF, which is based on KNN imputation [57]. KNN imputation performs well for both continuous and discrete data and can also handle multiple missing values [53]. However, it has some limitations, including low accuracy. Another approach for handling missing values is deletion, which involves removing all missing entries from the dataset during the analysis or training stages of an ML model [53]. Among the studies that handled missing values in their datasets, six [13,22,25,27,34,37] used the deletion approach. Although this is an easy way to handle missing values, deleting entries along with missing values changes the data distribution, which may lead to bias in the ML model results [53]. Because the performance of ML algorithms depends on the distribution and noise of the training data [58], changes in the data distribution before training the model must be avoided to maintain prediction reliability. If a researcher adopts the deletion approach, they must ensure that the data distribution remains unaffected or use ML algorithms robust to bias.

Data transformation is performed during data preprocessing, wherein data is converted from one format or structure to another [59]. However, 23 studies (approximately 76.6%) did not report the feature transformation method used. ML algorithms cannot perform effectively if the data are not transformed because they are likely to endure from inconsistency and bias. Normalization is one of the methods used to improve data analysis [60]. Because normalization improves the overall database and consistency of the data and reduces redundant data, studies [13,20,37,39] that applied normalization were robust to data redundancy [60]. Likewise, data standardization and scaling can improve ML algorithms and make the flow of data more efficient [61]. Standardization is the process of scaling data by converting them to standard normal variables (mean = 0, variance = 1) [50]. Because standardization is useful when the data follow a Gaussian or normal distribution [50], it is recommended to visualize the data or examine the data distribution to determine whether normalization is suitable. However, data normalization and standardization do not resolve the issue of large outliers. Therefore, outliers must be removed before normalization or standardization.

One-hot encoding was employed in four studies [13,20,37,39], and neither of them achieved satisfactory results; however, it cannot be considered ineffective as other parameters and settings can affect the performance of ML algorithms. The authors did not explain the reason for using one-hot encoding or the extent to which it affected performance. However, it may have adverse consequences when used for features with a high degree of cardinality as it will result in high dimensionality, which may cause memory and computing problems [62]. Moreover, the study [17] only used numeric features or those that could be converted to numeric easily. Thus, the developed models in this study could be biased due to their omission of categorical features, which could have an important role in predicting lung cancer survival. A researcher may convert categorical features into numeric features by coding each variable instead of eliminating them.

4.3. Features selection and engineering

Using features that are not related to survival is unlikely to increase the prediction accuracy and will instead increase the computational load. Therefore, selecting features that are significantly related to lung cancer survival can increase the prediction accuracy of ML algorithms. The results of this review indicate that eleven studies (approximately 36.6%) did not employ any feature selection algorithm. Nevertheless, the ML algorithms from these eleven studies performed almost as well as those that employed feature selection methods. Some studies used specific features as predictors of lung cancer survival, such as hemoglobin level, red blood cell count [21], and body mass index (BMI) [23]; however, the performances of these ML algorithms did not show any improvement compared to those of other studies. Nevertheless, Agrawal et al. [17] demonstrated that although "birthplace" can improve lung cancer survival prediction, "surgery" is the most important predictor of survival. Furthermore, each study reported different features as the most significant; however, this can be attributed to the differences in the ML algorithms or patient cohorts adopted in each study. In the future, researchers should explore and compare the features that are significant for survival prediction for each lung cancer histology. Thus, researchers can identify universal features that may be applicable to all types of lung cancer histology to predict survival.

4.4. ML algorithms

RF was the most used ML method among all studies. However, ML performances across the studies were not compared owing to differences in the data cleaning process, databases, patient cohorts, features, and validation procedures employed. The performance of ML algorithms depends on the quality of the data used, and when the data quality is poor, the ML model performs poorly, resulting in inaccurate predictions [63]. Nevertheless, no study performed an in-depth examination of the data quality and impact of the steps performed during data preprocessing on the ML model. Furthermore, each study had a different best-performing ML model; however, RF was determined to be the best in the seven studies [10,12,19,30,33–35]. Moreover, RF can be trained well on a small dataset [48], and it can handle data structures with complex and large-scale feature spaces [64]. Nevertheless, the hyperparameters of RF are complex, which may result in overfitting [65]. DT, RF, bagging, and XGBoost are tree-based models [66]. A tree-based model can use qualitative predictors in its prediction process without creating dummy variables [66]; hence, tree-based models are considered more effective for detecting non-monotonic or non-linear relationships between dependent variables and predictors. They are also capable of handling many high-order interactions and moderate dataset sizes more effectively than regression models [67]. Although tree-based models have many advantages, they also have some disadvantages. When a small dataset is used to train tree-based models based on the high correlation among predictors, the detection of interactions between predictors is impeded, which may lead to overfitting. However, this effect can be reduced by employing RF [68].

The RF algorithm is widely recognized and popular in the field of machine learning due to its excellent performance and ability to address certain challenges associated with other tree-based approaches, such as decision trees, GBM, Light GBM, and XGBoost. One of the key advantages of RF is its capability to reduce overfitting compared to a single DT algorithm. By combining multiple decision trees and aggregating their predictions, RF can mitigate the variance and enhance generalization performance. In contrast, a single DT is more susceptible to overfitting as it can learn intricate details and noise present in the training data.

Light GBT has been applied with data size 585, while it is more sensitive to overfitting on small datasets [24]. According to that, its performance was low. The fundamental idea behind AdaBoost is to combine multiple weak learners to create a strong ensemble model that performs better than the individual weak learners alone. However, if the weak learners used in AdaBoost are too powerful or have high capacity compared to the limited amount of data available, there is a risk of overfitting. The study [23] employed an Adaboost algorithm with a small-size dataset (809), wherein the utilization of an explicit weak learner was not clearly stated. It is possible that a complex weak learner has utilized alongside the small-size dataset, which may have contributed to low performance. These findings suggest that caution should be exercised when applying algorithms like Light GBT and AdaBoost to small datasets, and alternative approaches or regularization techniques should be considered to mitigate the risk of overfitting.

The study conducted by Ref. [15] utilized XGBoost in conjunction with one-hot encoding. While one-hot encoding can introduce dummy variables for each category, it is important to acknowledge that some of these variables may have limited or no predictive power. Consequently, the inclusion of such variables has the potential to introduce noise into the dataset, thereby increasing the complexity of the learning process for XGBoost. Moreover, when faced with noisy data and outliers, XGBoost may exhibit increased susceptibility. The presence of noise within a dataset can give rise to overfitting or suboptimal model performance. Additionally, one-hot encoding may affect the interpretation of the importance of features in XGBoost. Further, the importance values associated with individual one-hot encoded binary features may not directly reflect the importance of the original categorical variable. Considering this is important when interpreting the importance of features derived from one-hot encoding.

In survival data, time-to-event analysis is often conducted in order to determine when a specific event occurs, such as death or the recurrence of a disease. Consequently, the relationship between the predictor variables (e.g., patient characteristics, treatment factors) and survival outcomes may be non-linear. A RF is an effective tool for capturing non-linear relationships between features and the target variable. With RF, each decision tree is independent of the other, enabling it to capture a variety of patterns in the data based on different subsets of features. Conversely, the sequential nature of GBT may make it difficult to capture complex non-linear relationships as it attempts to correct previous errors sequentially.

While other tree-based models like Light GBT, XGBoost, AdaBoost, GBT, and decision trees have their strengths and use cases, RF has gained popularity due to its overall performance, simplicity, and ability to handle a wide range of data types and problems. However, it is important to note that the choice of model depends on the specific problem, dataset characteristics, and tuning requirements. Other models may outperform RF in certain scenarios, and it is always advisable to experiment and compare different models to identify the best fit for a given task.

A neural network-based model performs better when the dataset is large [69]. But a multilayer NN was applied in Ref. [13] to a small dataset comprising 239 samples to construct the model. However, they should have used ML algorithms that are more accurate with small datasets, such as Bayesian or RF models [48,70,71] to improve the performance of their model. In future studies, researchers must employ models that are suitable for the size of the datasets used. Also, the neural network-based model is sensitive to missing values. The Bayesian model is robust to irrelevant variables (those that have a small significant effect on the data). Moreover, it is efficient for missing values and high-dimensional data but is sensitive to correlated data [68]. In studies with small datasets and without imputed or handling missing data, one of the Bayesian-based models is recommended.

The performance of SVM in predicting survival between six months to two years was low [11,19]. Although SVM performs well with balanced datasets, its performance is poor with imbalanced datasets. This is because the separating hyperplane of SVM is biased toward majority classes, which may negatively impact minority classification accuracy [72]. Studies [11,19] that failed to achieve good performance with SVM applied it without handling the dataset imbalance. A significant amount of workflow must be accomplished before ML can be applied to predict lung cancer survival. For instance, imbalanced datasets must be addressed before developing ML algorithms. As [11] failed to address imbalanced data, the accuracies of its models were unacceptable. Classification models are more likely to be biased toward large classes and may completely ignore small classes. Most classification models, including NNs and DTs, perform better when the response variables are evenly distributed within the dataset [73]. Therefore, it is recommended that future studies handle imbalanced datasets using a suitable approach such as the synthetic minority oversampling technique before applying ML algorithms [74]. Additionally, we found that the reported ML algorithms and development processes in all studies were not detailed enough for replication on other datasets or external validation. For example, no study identified the features of the data that had missing data and neither the number nor percentage of missing data.

Moreover, three studies [11,14,24] showed their models performed better in predicting longer survival times. It is possible that these results are caused by datasets consisting of only a small number of patients who survived for a short period of time. Thus, none of the reviewed studies included ML algorithms that could simultaneously predict lung cancer survival for short-, medium-, and long-term periods (one, two, three, and five years). In future studies, researchers should develop ML algorithms capable of accurately predicting survival for both short and long periods.

Our recommendation is to discover the characteristics of the dataset prior to developing the machine learning model. To accomplish this, it may be necessary to find the descriptive statistics for the dataset and visualize them for identification of the size, distribution, and noise of the data. The identification of these characteristics will assist the data cleansing process, the decision between normalization and standardization, and the selection of the appropriate machine learning model for the size of the data set.

4.5. Model evaluation and performance metrics

Model validation is essential as it allows applying it to a population larger than the sample population. K-fold cross-validation allows sequential training and testing of each sample within a dataset during validation. Therefore, the amount of training data increases, which results in a lower pessimistic bias. Using different parts of the training dataset results in variance, which is a variation in ML algorithms [66]. However, the variance can be overcome using bagging and RF models because these models average a set of observations [66]. Moreover, the test folds do not overlap in k-fold cross-validation [75]; therefore, it is suitable for small sample sizes [76]. However, nine studies [13,19,23,24,29,30,32,34,35] with small datasets used training/test splits instead of k-fold cross-validation. Moreover, the train/test method for validating the models may result in misleading estimates of error rates if the selected split is inappropriate [77]. None of the studies reported reasons for selecting their respective validation methods. In the future, the validation method should be selected according to the dataset size and ML method used. It is recommended that researchers report the reason for selecting the validation method, which will allow other researchers to follow the same process if they have data with similar characteristics. Furthermore, only three studies used external validation, and one did not report the validation method used. The importance of external validation can be attributed to the fact that each population has distinctive, unmeasurable, site-specific features that may critically impact both patient outcomes and complications. Thus, training a model without external validation may result in significantly different performances for different populations [78].

Overall, This study focuses on the incorporation of machine learning into predicting lung cancer survival time. The study identifies various

challenges that need to be addressed, including missing data, outliers, class imbalance, and predicting long-term survival times. Datasets containing medical information often present these challenges, necessitating appropriate resolutions. The quality of the included studies in this review was found to be moderate. However, further well-designed research is required to validate the predictions made by machine learning. Key areas that need improvement include pre-processing techniques, particularly in terms of sample size and handling missing data, which were either unclear or considered high risk in a significant portion of the studies.

Several negative consequences can arise when addressing missing data, employing improper feature transformation methods, or mishandling categorical features in the context of predicting lung cancer survival using machine learning. ML models trained on incomplete or improperly transformed data may provide inaccurate predictions, leading to incorrect assessments of lung cancer survival. This can result in misguided decisions for patients and potentially compromise their treatment. Mishandling missing data or categorical features can also introduce bias into the ML models. Additionally, ML models developed without adequately addressing missing data or employing appropriate feature transformation methods may struggle to generalize well to new, unseen data.

The inclusion of irrelevant or non-predictive features in the model due to improper feature selection can decrease prediction accuracy. These unrelated features not only increase computational load but also introduce noise, making it harder for the ML algorithm to identify meaningful patterns related to lung cancer survival. The choice of ML algorithms is crucial, as selecting an inappropriate one for a specific dataset can lead to suboptimal performance and inaccurate predictions. Each algorithm has its own strengths and weaknesses, and understanding these characteristics is essential to ensure optimal model selection.

Furthermore, the issue of imbalanced datasets, where one class (e.g., short-term survival) is significantly underrepresented compared to another class, needs to be addressed. Failure to account for this imbalance can bias the ML model towards the majority class, resulting in poor prediction accuracy for the minority class, which may be important for identifying high-risk individuals or those requiring specific interventions. Lastly, neglecting to understand the dataset's characteristics, such as its size, distribution, and noise, can hinder effective data preprocessing and model development. Inadequate data cleaning, normalization, or standardization can introduce biases or distort the data distribution, ultimately leading to unreliable predictions.

5. Limitation

This systematic review has some limitations that must be considered. This review aimed to identify all studies conducted on predicting lung cancer survival based on clinical data using ML. Additionally, the studies included in this review had significant differences, which limited our ability to compare them. Further, since some studies failed to provide a quantitative analysis of model performance, such as specificity, sensitivity, negative prediction rate, or positive prediction rate, this review does not include a quantitative analysis. Despite these limitations, most reviewed studies used similar criteria to predict lung cancer survival, such as clinical data and ML algorithms, which contribute to a strong framework for building ML algorithms to predict lung cancer survival.

6. Conclusion

To select the most suitable treatment for a specific patient with lung cancer, it is crucial that healthcare providers are able to predict their survival time. An ML model can be used to predict survival, as it is currently used in many other fields to predict outcomes. Based on the findings of this review, it can be concluded that interest in using ML algorithms to predict lung cancer survival based on clinical data has been growing since 2012. An analysis of various data sources showed that the SEER database is the most widely used. Although the target patient cohort varied between studies, a detailed analysis of other patient cohorts, such as those with small-cell lung cancer, is required. The results of the analysis of preprocessing methods indicated that, although many studies employed feature selection methods to identify important features, most reviewed studies did not account for missing data, normalization, scaling, and standardized data. Thus, it is possible that the ML algorithms used in these studies may be biased. Therefore, a feature study is still necessary to predict lung cancer survival based on clinical data by paying more attention to preprocessing methods and providing more detail on these methods.

Authors' contributions

Altuhaifa, Dr. Win, and Dr. Su conceptualized and designed the study. Altuhaifa, Dr. Win, and Dr. Su collected the reviewed articles and filtering Them., and drafted the manuscript. Fatimah Altuhaifa conducted the analysis The reviewed articles.

Funding

Not applicable.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

We are grateful for the scholarship provided by the Saudi Arabia Ministry of Higher Education to Fatimah Altuhaifa. Also, we are grateful to the University of Wollongong Australia for providing access to most of the digital libraries and scientific search engines used in this study. We are deeply grateful to Rachel Lawson, Librarian for Science, Medicine, and Health Liaison Services at Building 16 Library, for her invaluable assistance in finding the reviewed article. Her expertise and dedication were instrumental to our research.

Appendices.

Appendix A

Table A1
Best models in each study based on RMSE accuracy

Model	Accuracy	Predicted period of survival time	Publication size	Reference
RF	10.52	6 months	10,442	[12]
Custom Ensemble	15.30	5 years	10,442	[16]
Self-Ordering Maps	15.59	5 years	10,442	[17]
GRNN	0.60	5 years	683	[20]

(*continued on next page*)

Table A1 (*continued*)

Model	Accuracy	Predicted period of survival time	Publication size	Reference
Ridge Regression	2.70	3 years	291	[27]
ANN for male	2.32	5 years	38,262	[31]
DNN	0.314	5 years	10,001	[36]

Table A2
Best models in each study based on AUC accuracy

Model	Accuracy	Predicted period of survival	Publication size	Reference
ADTree	93.8	5 years	57,254	[11]
MNN	78.30	2 years	239	[13]
ANN	92.0	5 years	Not reported	[14]
CNN	92.0	5 years	Not reported	[14]
RF	90.1	3 years	50,687	[10]
XGBoost	78.6	1 year	5973	[15]
RF	80.2	5 years	5123	[18]
IRF	98.0	2 years	509	[19]
Gaussian K-base NB	88.1	5 years	321	[21]
AdaBoost	71.3	2 years	809	[23]
Logistic regression	76.0	5 years	585	[24]
DL	74.4	5 years	16,613	[25]
SVM	71.0	3 years	371	[26]
NB	81.0	3 years	291	[27]
SORG ML	71.4	1 year	150	[28]
LASSO	75.3	1 year	1563	[29]
DL	81.3	5 years	704	[30]
RF	95.1	5 years	17,484	[33]
RF	68.0	5 years	739	[35]
Cox Regression	72.2	5 years	2166	[38]
XGBoost	Female 93.0	1 year	28,458	[39]

Table A3
Best model based on C-statistic accuracy

Model	Accuracy	Predicted period time of survival	Publication size	Reference
DL	73.9	5 years	17,322	[22]
DL	63.6	5 years	1137	[32]
RF	67.2	5 years	506	[34]
DL	83.4	5 years	4617	[37]

Appendix B

Table B1
Models in each study based on RMSE accuracy for 6 months survival time

ML algorithms	Accuracy for each reference
RF	10.52 [12]
LR	10.63 [12]
GBT	10.65 [12]
Custom Ensemble	10.84 [12]

Table B2
Models in each study based on RMSE accuracy for 2 years survival time

ML algorithms	Accuracy for each reference
RF	15.70 [12]
LR	15.77 [12]
GBT	15.65 [12]
Custom Ensemble	16.26 [12]

Table B3
Models in each study based on RMSE accuracy for 3 years survival time

ML algorithms	Accuracy for each reference
LR	3.01 [27]
Ridge Regression	2.70 [27]
Line SVR	2.78 [27]
Poly SVM	2.81 [27]

Table B4
Models in each study based on RMSE accuracy for 5 years survival time

ML algorithms	Accuracy for each reference
RF	20.51 [12], 15.63 [16]
LR	21.37 [12], 15.38 [16]
GBT	21.14 [12], 15.32 [16]
Custom Ensemble	21.18 [12], 15.30 [16]
SVM	15.82 [16]
DT	15.81 [16]
HC	16.202 [17]
MBC	16.250 [17]
K-Means Clustering	16.193 [17]
SOMs	15.591 [17]
Non-negative Matrix Factorization	16.589 [17]
GRNN	0.60 [20]
ANN	Male 2.32 [31], female 2.52 [31]
Bayes Net	0.315 [36]
DNN	0.314 [36]
Logistic Regression	0.314 [36]
Lazy Classifier LWL	0.318 [36]
J48 DT	0.32 [36]
Meta-Classifier (ASC)	0.316 [36]
Rule Learner (OneR)	0.339 [36]

Table B5
Models in each study based on AUC accuracy for 6 months survival time

ML algorithms	Accuracy for each reference
SVM	60.5 [11], 79.0 [14]
ANN	74.0 [11], 82.0 [14]
J48 DT	78.1 [11]
RF	78.1 [11], 80.0 [14]
LogitBoost	80.5 [11]
Random Subspace	80.8 [11]
ADTree	80.8 [11]
CNN	83.0 [14]
RNN	81.0 [14]
NB	77.0 [14]

Table B6
Models in each study based on AUC accuracy for 1 year survival time

ML algorithms	Accuracy for each reference
SVM	63.4 [11], 73.0 [15], male 86.4 [39], 91.6 [39]
ANN	80.2 [11]
J48 DT	78.6 [11]
RF	81.1 [11], 85.2 [10], 73.6 [15], male 86.0 [39], female 91.2 [39]
MTLR	82.1 [10]
Logit Boost	82.1 [11]
Random Subspace	82.4 [11]
ADTree	83.0 [11]
Logistic Algorithms	71.0 [15], male 85.5 [39], female 92.1 [39]
DT	Male 84.3 [39], female 90.4 [39]
XGBoost	78.6 [15], male 87.8 [39], female 93.0 [39]
SORG	71.4 [28]
LASSO	75.3 [29]
KNN	Male 83.6 [39], female 89.8 [39]

Table B7
Models in each study based on AUC accuracy for 2 years survival time

ML algorithms	Accuracy for each reference
SVM	57.0 [11], 64.0 [14], 51.6 [19]
ANN	81.6 [11], 86.0 [14]
J48 DT	81.8 [11]
RF	86.7 [11], 66.0 [14], 96.8 [19], 55.0 [24]
Logit Boost	87.7 [11]
Random Subspace	87.8 [11]
ADTree	88.4 [11]
CNN	86.0 [14]
RNN	86.0 [14]
NB	63.0 [14], 54.0 [19]
Logistic Regression	64.2 [13], 49.0 [24]
SPNN	65.0 [13]
MNN	78.3 [13]
DT	47.4 [19]
IRF	98.0 [19]
MLP	65.0 [24]
XGBoost	58.0 [24]
Gaussian NB	52.0 [24]
Light GBT	54.0 [24]
support vector clustering SVC	55.0 [24]

Table B8
Models in each study based on AUC accuracy for 3 years survival time

ML algorithms	Accuracy for each reference
SVM	71.0 [26], 73.0 [27], male 70.6 [39], female 81.1 [39]
RF	90.1 [10], 76.0 [27], male 70.2 [39], female 80.7 [39]
NB	81.0 [27]
MTLR	86.4 [10]
LR	74.0 [27]
LASSO	73.6 [29]
Logistic regression	Male 69.9 [39], female 81.8 [39]
DT	Male [39], female 80.1 [39]
XGBoost	Male 72.4 [39], female 82.9 [39]
KNN	Male 67.2 [39], female 79.5 [39]

Table B9
Models in each study based on AUC accuracy for 5 years survival time

ML algorithms	Accuracy for each reference
SVM	56.4 [11], 84.0 [14], male 72.0 [39], female 82.3 [39]
ANN	92.3 [11], 92.0 [14]
J48 DT	84.7 [11], 94.4 [33]
RF	90.5 [11], 86.0 [14], 89.9 [10], 69.2 [23], 68.8 [25], 80.259 [18], 95.1 [33], 68.0 [35], male 71.6 [39], female 82.0 [39]
Logit Boost	93.0 [11]
Random Subspace	93.2 [11], 56.0 [24]
MTLR	87.0 [10]
ADTree	93.7 [11]
CNN	92.0 [14]
RNN	91.0 [14]
NB	83.0 [14], 59.7 [21]
Logistic Regression	63.2 [23], 76.0 [24], 92.7 [33], male 71.0 [39], female 83.0 [39]
SPNN	71.0 [24]
DT	69.2 [23], 78.166 [18], male 69.4 [39], female 81.4 [39]
MLP	71.0 [24]
XGBoost	54.0 [24], male 73.5 [39], female 84.2 [39]
Gaussian NB	73.0 [24]
Light GBT	67.0 [24]
MTLR	87.0 [10]
LR	62.2 [21]
Gaussian K base NB	88.1 [21]
Bagging	70.6 [23]
AdaBoost	71.3 [23]
support vector clustering SVC	67.0 [24]
DL	74.4 [25], 81.6 [30]
LASSO	65.6 [29]
Bayes Net	94.2 [33]
Cox Regression	72.2 [38]
KNN	Male 68.6 [39], female 80.7 [39]

Table B10
Model based on C-statistic accuracy for 5 years survival time

ML model	Accuracy for each reference
DL	73.9 [22], 63.6 [32], 83.4 [37]
Cox Regression	56.1 [32], GSE72094 cohort 63.0 [34], TCGA cohort 64.5 [34], 64.0 [37]
Cox NN	56.2 [32]
RF	GSE72094 cohort 67.2 [34], TCGA cohort 65.6 [34], 67.8 [37]

Appendix C

Table C1
Validation

	K-fold cross-validation	Data splitting	Not reported	External validation
[11]	✓			
[12]	✓			
[13]		✓		
[14]		✓		
[10]			✓	
[15]		✓		✓
[16]	✓			
[17]			✓	
[18]		✓		
[19]		✓		
[20]	✓			
[21]			✓	
[22]		✓		✓
[23]	✓	✓		
[24]	✓	✓		
[25]		✓		✓
[26]	✓			
[27]	✓			
[28]			✓	✓
[29]		✓		✓
[30]		✓		✓
[31]		✓		
[32]		✓		✓
[33]	✓			
[34]		✓		✓
[35]		✓		✓
[36]	✓			
[37]	✓			
[38]		✓		✓
[39]	✓			

Appendix D

Table D1
PROBAST assessment

Reference	Study participant	Predictors	Outcomes	Sample size and missing data	Statistical analysis
[11]	Low	Unclear	Low	Unclear	Unclear
[12]	Low	Low	Low	Unclear	Low
[13]	Low	Low	Low	Unclear	Low
[14]	Low	Low	Low	Unclear	Unclear
[10]	Low	Low	Low	Unclear	Unclear
[15]	Low	Low	Low	Unclear	Unclear
[16]	Low	Low	Low	Unclear	Unclear
[17]	Low	Unclear	Low	Unclear	Unclear
[18]	Low	Low	Low	Low	Unclear
[19]	Low	Low	Low	Unclear	Low
[20]	Low	Unclear	Low	High risk	Unclear
[21]	Low	Unclear	Low	High risk	Low
[22]	Low	Unclear	Low	Low	Unclear
[23]	Low	Low	Low	Unclear	Unclear
[24]	Low	Low	Low	High risk	Low
[25]	Low	Low	Low	Low	Low
[26]	Low	Unclear	Low	High risk	Unclear
[27]	Low	Unclear	Low	Unclear	Low
[28]	Low	Unclear	Low	Unclear	Unclear
[29]	Low	Unclear	Low	Unclear	Unclear
[30]	Low	Unclear	Low	High risk	Unclear

(continued on next page)

Table D1 (*continued*)

Reference	Study participant	Predictors	Outcomes	Sample size and missing data	Statistical analysis
[31]	Low	Low	Low	Unclear	Unclear
[32]	Low	Low	Low	Unclear	Unclear
[33]	Low	Low	Low	Low	Unclear
[34]	Low	Low	Low	Unclear	Unclear
[35]	Low	Low	Low	High risk	Unclear
[36]	Low	Low	Low	Unclear	Low
[37]	Low	Low	Low	Low	Unclear
[38]	Low	Low	Low	Unclear	Unclear
[39]	Low	Low	Low	Low	Unclear

References

[1] WHO, "Cancer [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer, 2022.

[2] S. Tomassini, N. Falcionelli, P. Sernani, L. Burattini, A.F. Dragoni, Lung nodule diagnosis and cancer histology classification from computed tomography data by convolutional neural networks: a survey, Comput. Biol. Med. 146 (Jul 2022), 105691, https://doi.org/10.1016/j.compbiomed.2022.105691.

[3] L.A. Vale-Silva, K. Rohr, Long-term cancer survival prediction using multimodal deep learning, in English, Sci Rep-Uk 11 (1) (Jun 29 2021). ARTN 1350510.1038/s41598-021-92799-4.

[4] Y. Yang, L. Xu, L. Sun, P. Zhang, S.S. Farid, Machine learning application in personalised lung cancer recurrence and survivability prediction, Comput. Struct. Biotechnol. J. 20 (2022) 1811–1820, https://doi.org/10.1016/j.csbj.2022.03.035.

[5] E.M. Nwanosike, B.R. Conway, H.A. Merchant, S.S. Hasan, Potential applications and performance of machine learning techniques and algorithms in clinical practice: a systematic review, Int. J. Med. Inf. 159 (Mar 2022), 104679, https://doi.org/10.1016/j.ijmedinf.2021.104679.

[6] I. Kaur, M.N. Doja, T. Ahmad, Data mining and machine learning in cancer survival research: an overview and future recommendations, J. Biomed. Inf. 128 (Apr 2022), 104026, https://doi.org/10.1016/j.jbi.2022.104026.

[7] A. Liberati, et al., The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration, BMJ 339 (Jul 21 2009) b2700, https://doi.org/10.1136/bmj.b2700.

[8] Covidence systematic review software." Veritas Health Innovation. https://www.covidence.org/(accessed.

[9] D. Bzdok, N. Altman, M. Krzywinski, Statistics versus machine learning, Nat. Methods 15 (4) (Apr 2018) 233–234, https://doi.org/10.1038/nmeth.4642.

[10] T. He, J. Li, P. Wang, Z. Zhang, Artificial intelligence predictive system of individual survival rate for lung adenocarcinoma, Comput. Struct. Biotechnol. J. 20 (2022) 2352–2359, https://doi.org/10.1016/j.csbj.2022.05.005.

[11] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, A. Choudhary, Lung cancer survival prediction using ensemble data mining on SEER data, Sci. Program. 20 (1) (2012) 29–42, https://doi.org/10.1155/2012/920245, in English.

[12] J.A. Bartholomai, H.B. Frieboes, Lung cancer survival prediction via machine learning regression, classification, and statistical techniques, 2018, in: Proc IEEE Int Symp Signal Proc Inf Tech, Dec 2018, pp. 632–637, https://doi.org/10.1109/ISSPIT.2018.8642753.

[13] Y. Dagli, S. Choksi, S. Roy, Prediction of two year survival among patients of non-small cell lung cancer, L N Comput. Vis. Biomed. 31 (2019) 169–177, https://doi.org/10.1007/978-3-030-04061-1_17, in English.

[14] S. Doppalapudi, R.G. Qiu, Y. Badr, Lung cancer survival period prediction and understanding: deep learning approaches, Int. J. Med. Inf. 148 (Apr 2021), 104371, https://doi.org/10.1016/j.ijmedinf.2020.104371.

[15] Z. Huang, C. Hu, C. Chi, Z. Jiang, Y. Tong, C. Zhao, An artificial intelligence model for predicting 1-year survival of bone metastases in non-small-cell lung cancer patients based on XGBoost algorithm, 2020, BioMed Res. Int. (2020), 3462363, https://doi.org/10.1155/2020/3462363.

[16] C.M. Lynch, et al., Prediction of lung cancer patient survival via supervised machine learning classification techniques, Int. J. Med. Inf. 108 (Dec 2017) 1–8, https://doi.org/10.1016/j.ijmedinf.2017.09.013.

[17] C.M. Lynch, V.H. van Berkel, H.B. Frieboes, Application of unsupervised analysis techniques to lung cancer patient data, PLoS One 12 (9) (2017), e0184370, https://doi.org/10.1371/journal.pone.0184370.

[18] X.Y. Mei, Predicting five-year overall survival in patients with non-small cell lung cancer by ReliefF algorithm and random forests, in: 2017 Ieee 2nd Advanced Information Technology, Electronic and Automation Control Conference (Iaeac), 2017, pp. 2527–2530, https://doi.org/10.1109/IAEAC.2017.8054479, in English.

[19] P. Nanda, N. Duraipandian, Prediction of survival rate from non-small cell lung cancer using improved random forest, in: Proceedings of the 5th International Conference on Inventive Computation Technologies (Icict-2020), 2020, pp. 93–97, https://doi.org/10.1109/ICICT48043.2020.9112558, in English.

[20] K. Qaddoum, Lung Cancer Patient's Survival Prediction Using GRNN-CP, 2020.

[21] K. R, G.R. R, Accuracy enhanced lung cancer prognosis for improving patient survivability using proposed Gaussian classifier system, J. Med. Syst. 43 (7) (May 24 2019), https://doi.org/10.1007/s10916-019-1297-2, 201.

[22] Y. She, et al., Development and validation of a deep learning model for non-small cell lung cancer survival, JAMA Netw. Open 3 (6) (Jun 1 2020), e205842, https://doi.org/10.1001/jamanetworkopen.2020.5842.

[23] J.A. Sim, Y.H. Yun, Predicting disease-free lung cancer survival using patient reported outcome (PRO) measurements with comparisons of five machine learning techniques (MLT), Stud. Health Technol. Inf. 264 (Aug 21 2019) 1588–1589, https://doi.org/10.3233/SHTI190548.

[24] M. Yakar, D. Etiz, S. Yilmaz, O. Celik, A.K. Guntulu, M. Metintas, Prediction of survival and progression-free survival using machine learning in stage III lung cancer: a pilot study, Turk. Oncol. Derg. 36 (4) (2021) 446–458, https://doi.org/10.5505/tjo.2021.2788, in English.

[25] L. Yang, et al., A novel deep learning prognostic system improves survival predictions for stage III non-small cell lung cancer, Cancer Med. (May 2 2022), https://doi.org/10.1002/cam4.4782.

[26] J. Yu, et al., LUADpp: an effective prediction model on prognosis of lung adenocarcinomas based on somatic mutational features, BMC Cancer 19 (1) (Mar 22 2019) 263, https://doi.org/10.1186/s12885-019-5433-7.

[27] Y. Liu, et al., Developing prognostic gene panel of survival time in lung adenocarcinoma patients using machine learning, Transl. Cancer Res. 9 (6) (Jun 2020) 3860–3869, https://doi.org/10.21037/tcr-19-2739.

[28] G. Zhong, et al., External validation of the SORG machine learning algorithms for predicting 90-day and 1-year survival of patients with lung cancer-derived spine metastases: a recent bi-center cohort from China, Spine J. 23 (5) (May 2023) 731–738, https://doi.org/10.1016/j.spinee.2023.01.008.

[29] Y. Li, D. Ge, J. Gu, F. Xu, Q. Zhu, C. Lu, A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies, BMC Cancer 19 (1) (Sep 5 2019) 886, https://doi.org/10.1186/s12885-019-6101-7.

[30] Y.H. Lai, W.N. Chen, T.C. Hsu, C. Lin, Y. Tsao, S. Wu, Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning, Sci. Rep. 10 (1) (Mar 13 2020) 4679, https://doi.org/10.1038/s41598-020-61588-w.

[31] H.R.C.P. Tsokos, Artificial neural network model for predicting lung cancer survival, J. Data Anal. Inf. Process. (2017), https://doi.org/10.4236/jdaip.2017.51003.

[32] J. Wang, N. Chen, J. Guo, X. Xu, L. Liu, Z. Yi, SurvNet: a novel deep neural network for lung cancer survival analysis with missing values, Front. Oncol. 10 (2020), 588990, https://doi.org/10.3389/fonc.2020.588990.

[33] A. Safiyari, R. Javidan, Predicting lung cancer survivability using ensemble learning methods, in English, in: Proceedings of the 2017 Intelligent Systems Conference (Intellisys), 2017, pp. 684–688. <Go to ISI>://WOS:000456827800090.

[34] B. Ma, Y. Geng, F. Meng, G. Yan, F. Song, Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method, J. Cancer 11 (5) (2020) 1288–1298, https://doi.org/10.7150/jca.34585.

[35] S. Zhang, X. Zeng, S. Lin, M. Liang, H. Huang, Identification of seven-gene marker to predict the survival of patients with lung adenocarcinoma using integrated multi-omics data analysis, J. Clin. Lab. Anal. 36 (2) (Feb 2022), e24190, https://doi.org/10.1002/jcla.24190.

[36] S. Huang, I. Arpaci, M. Al-Emran, S. Kılıçarslan, M.A. Al-Sharafi, A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability, Multimed. Tool. Appl. (2023), https://doi.org/10.1007/s11042-023-16349-y.

[37] L. Jin, Q. Zhao, S. Fu, F. Cao, B. Hou, J. Ma, Development and validation of machine learning models to predict survival of patients with resected stage-III NSCLC, Front. Oncol. 13 (2023), 1092478, https://doi.org/10.3389/fonc.2023.1092478.

[38] H. Ma, L. Tong, Q. Zhang, W. Chang, F. Li, Identification of 5 gene signatures in survival prediction for patients with lung squamous cell carcinoma based on integrated multiomics data analysis, 2020, BioMed Res. Int. (2020), 6427483, https://doi.org/10.1155/2020/6427483.

[39] Y. Wang, et al., A machine learning-based investigation of gender-specific prognosis of lung cancers, Medicina (Kaunas) 57 (2) (Jan 22 2021), https://doi.org/10.3390/medicina57020099.

[40] A. Moskowitz, K. Chen, Defining the patient cohort, in: Secondary Analysis of Electronic Health Records, CH), Cham, 2016, pp. 93–100, ch. 93–100.

[41] D.S.K. Karunasingha, Root mean square error or mean absolute error? Use their ratio as well, Inf. Sci. 585 (2022) 609–629, https://doi.org/10.1016/j.ins.2021.11.036.
[42] F. Melo, Area under the ROC curve, in: Encyclopedia of Systems Biology, 2013. Springer Nature.
[43] D. Westreich, S.R. Cole, M.J. Funk, M.A. Brookhart, T. Sturmer, The role of the c-statistic in variable selection for propensity score models, Pharmacoepidemiol. Drug Saf. 20 (3) (Mar 2011) 317–320, https://doi.org/10.1002/pds.2074.
[44] R.F. Wolff, et al., PROBAST: a tool to assess the risk of bias and applicability of prediction model studies, Ann. Intern. Med. 170 (1) (Jan 1 2019) 51–58, https://doi.org/10.7326/M18-1376.
[45] Y. Kumar, S. Gupta, R. Singla, Y.C. Hu, A systematic review of artificial intelligence techniques in cancer prediction and diagnosis, Arch. Comput. Methods Eng. 29 (4) (2022) 2043–2070, https://doi.org/10.1007/s11831-021-09648-w.
[46] E.W. Steyerberg, M.J. Eijkemans, F.E. Harrell Jr., J.D. Habbema, Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets, Stat. Med. 19 (8) (Apr 30 2000) 1059–1079, https://doi.org/10.1002/(sici)1097-0258(20000430)19:8<1059::aid-sim412>3.0.co;2-0.
[47] T. van der Ploeg, P.C. Austin, E.W. Steyerberg, Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints, BMC Med. Res. Methodol. 14 (Dec 22 2014) 137, https://doi.org/10.1186/1471-2288-14-137.
[48] V. Khadse, P.N. Mahalle, S.V. Biraris, An empirical comparison of supervised machine learning algorithms for internet of things data, in English, in: 2018 Fourth International Conference on Computing Communication Control and Automation (Iccubea), 2018. <Go to ISI>://WOS:000493801500117.
[49] D.M. Rice, Causal reasoning, in: Calculus of Thought, 2014, pp. 95–123.
[50] Scikit-learn developers (BSD License), 6.3. Preprocessing data, https://scikit-learn.org/stable/modules/preprocessing.html.
[51] AmericanCancerSociety. "Key Statistics for Lung Cancer." American Cancer Society. https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html (accessed.
[52] K. Maharana, S. Mondal, B. Nemade, A review: data pre-processing and data augmentation techniques, Global Transit. Proc. 3 (1) (2022) 91–99, https://doi.org/10.1016/j.gltp.2022.04.020.
[53] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, J. Big Data 8 (1) (2021) 140, https://doi.org/10.1186/s40537-021-00516-9.
[54] A.R. Donders, G.J. van der Heijden, T. Stijnen, K.G. Moons, Review: a gentle introduction to imputation of missing values, J. Clin. Epidemiol. 59 (10) (Oct 2006) 1087–1091, https://doi.org/10.1016/j.jclinepi.2006.01.014.
[55] D.B. Rubin, Multiple imputation after 18+ years, J. Am. Stat. Assoc. 91 (434) (Jun 1996) 473–489, https://doi.org/10.1080/01621459.1996.10476908, in English.
[56] T. Kose, S. Ozgur, E. Cosgun, A. Keskinoglu, P. Keskinoglu, Effect of missing data imputation on deep learning prediction performance for vesicoureteral reflux and recurrent urinary tract infection clinical study, 2020, BioMed Res. Int. (2020), 1895076, https://doi.org/10.1155/2020/1895076.
[57] N. S, E.A. C, M.C. M, H.D. Lee, ReliefF for Multi-Label Feature Selection, 2013, https://doi.org/10.1109/BRACIS.2013.10 [Online]. Available: https://ieeexplore.ieee.org/document/6726418.
[58] N.G. Gyori, M. Palombo, C.A. Clark, H. Zhang, D.C. Alexander, Training data distribution significantly impacts the estimation of tissue microstructure with machine learning, Magn. Reson. Med. 87 (2) (Feb 2022) 932–947, https://doi.org/10.1002/mrm.29014.
[59] M.S. Chen, J.W. Han, P.S. Yu, Data mining: an overview from a database perspective, in English, IEEE Trans. Knowl. Data Eng. 8 (6) (Dec 1996) 866–883, https://doi.org/10.1109/69.553155.
[60] M.M. Siraj, N.A. Rahmat, M.M. Din, A survey on privacy preserving data mining approaches and techniques, in English, in: 2019 8th International Conference on Software and Computer Applications (Icsca 2019), 2019, pp. 65–69, https://doi.org/10.1145/3316615.3316632.
[61] M. Gal, D.L. Rubinfeld, Data standardization, SSRN Electron. J. (2018), https://doi.org/10.2139/ssrn.3326377.
[62] C. Seger, An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, SWEDEN, in: KTH ROYAL INSTITUTE OF TECHNOLOGY, 2018 [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf.
[63] P.A. Noseworthy, et al., Assessing and mitigating bias in medical artificial intelligence: the effects of race and ethnicity on a deep learning model for ECG analysis, Circ. Arrhythmia Electrophysiol. 13 (3) (Mar 2020), e007988, https://doi.org/10.1161/CIRCEP.119.007988.
[64] E. Scornet, G. Biau, J.-P. Vert, Consistency of random forests, Ann. Stat. 43 (4) (2015), https://doi.org/10.1214/15-aos1321.
[65] Y. Ao, H. Li, L. Zhu, S. Ali, Z. Yang, The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling, J. Petrol. Sci. Eng. 174 (2019) 776–789, https://doi.org/10.1016/j.petrol.2018.11.067.
[66] T. Hastie. "Tree-based Methods." Stanford University. https://hastie.su.domains/MOOC-Slides/trees.pdf (accessed.
[67] M. Schweinberger. "Tree-Based Models in R." Lang. Technol. Data Anal. Lab. (LADAL). https://ladal.edu.au/tree.html#References (accessed.
[68] E.E. Sevices, Data Science and Big Data Analytics: Discovering, Analyzing, 2015. Visualizing and Presenting Data.
[69] J.V. Tu, Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes, J. Clin. Epidemiol. 49 (11) (1996) 1225–1231, https://doi.org/10.1016/s0895-4356(96)00002-9.
[70] R. van de Schoot, J.J. Broere, K.H. Perryck, M. Zondervan-Zwijnenburg, N.E. van Loey, Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors, in English, Eur. J. Psychotraumatol. 6 (2015). ARTN 2521610.3402/ejpt.v6.25216.
[71] D. McNeish, On using bayesian methods to address small sample problems, Struct. Equ. Model.: A Multidiscip. J. 23 (5) (2016) 750–773, https://doi.org/10.1080/10705511.2016.1186549.
[72] F.Q. Han, M. Lei, W.J. Zhao, J.X. Yang, A new Support vector machine for imbalance data classification, Intell. Autom. Soft Comput. 18 (6) (2012) 679–686, https://doi.org/10.1080/10798587.2012.10643277, in English.
[73] P. Kumar, R. Bhatnagar, K. Gaur, A. Bhatnagar, Classification of imbalanced data: review of methods and applications, IOP Conf. Ser. Mater. Sci. Eng. 1099 (1) (2021), https://doi.org/10.1088/1757-899x/1099/1/012077.
[74] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953, in English.
[75] S. Raschka, STAT 479: Machine Learning Lecture Notes, University of Wisconsin–Madison, 2018.
[76] S. Yadav, S. Shukla, in English, in: Analysis of K-fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification, 2016, pp. 78–83, https://doi.org/10.1109/Iacc.2016.25. Int Conf Adv Compu.
[77] R. Gutierrez-Osuna, Introduction to Pattern Analysis, Texas A&M University, 2005.
[78] T.C. Canturk, et al., A scoping review of complication prediction models in spinal surgery: an analysis of model development, validation and impact, North Am. Spine Soc. J. 11 (Sep 2022), 100142, https://doi.org/10.1016/j.xnsj.2022.100142.