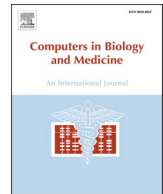


Daftar isi tersedia di [Sains Langsung](https://www.elsevier.com/locate/compbiomed)

Komputer dalam Biologi dan Kedokteran

beranda jurnal: www.elsevier.com/locate/compbiomed

Memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis menggunakan pembelajaran mesin: Sebuah tinjauan

Fatimah Abdulazim Altuhaifa^{A,B,*}, Khin Daripada Menang^A, Guoxin Su^A

^ASekolah Komputasi dan Teknologi Informasi, Universitas Wollongong, NSW, 2500, Australia

^BKementerian Pendidikan Tinggi Arab Saudi, Riyadh, Arab Saudi

INFO PASAL

Kata kunci:

Penambangan data
Pembelajaran mesin
Kecerdasan buatan
Kanker paru-paru
Prediksi kelangsungan hidup
Pemilihan fitur

ABSTRAK

Pembelajaran mesin telah mendapatkan popularitas dalam memprediksi waktu kelangsungan hidup di bidang medis. Ulasan ini mengkaji penelitian yang memanfaatkan pembelajaran mesin dan teknik pengumpulan data untuk memprediksi kelangsungan hidup kanker paru-paru menggunakan data klinis. Tinjauan literatur sistematis menelusuri database MEDLINE, Scopus, dan Google Cendekia, mengikuti pedoman pelaporan dan menggunakan sistem COVIDENCE. Studi yang diterbitkan dari tahun 2000 hingga 2023 yang menggunakan pembelajaran mesin untuk prediksi kelangsungan hidup kanker paru-paru juga disertakan. Penilaian risiko bias menggunakan model prediksi alat penilaian risiko bias. Tiga puluh penelitian telah ditinjau, dengan 13 (43,3%) menggunakan database surveilans, epidemiologi, dan hasil akhir. Penanganan data yang hilang telah diatasi dalam 12 (40%) penelitian, terutama melalui transformasi dan konversi data. Algoritme pemilihan fitur digunakan dalam 19 (63,3%) penelitian, dengan usia, jenis kelamin, dan tahap N menjadi fitur yang paling banyak dipilih. Random forest adalah model pembelajaran mesin yang dominan, digunakan dalam 17 (56,6%) penelitian. Meskipun jumlah studi prediksi kelangsungan hidup kanker paru-paru terbatas, penggunaan model pembelajaran mesin berdasarkan data klinis telah berkembang sejak tahun 2012. Pertimbangan terhadap kelompok pasien yang beragam dan pra-pemrosesan data sangatlah penting. Khususnya, sebagian besar penelitian tidak memperhitungkan data yang hilang, normalisasi, penskalaan, atau data standar, yang berpotensi menimbulkan bias. Oleh karena itu, diperlukan studi komprehensif tentang prediksi kelangsungan hidup kanker paru-paru menggunakan data klinis untuk mengatasi tantangan ini.

1. Perkenalan

Kanker paru-paru adalah penyebab utama kematian terkait kanker, terhitung sekitar 1,8 juta kematian pada tahun 2020 [1,2]. Selain itu, penyakit ini juga merupakan angka tertinggi dari tahun hidup yang disesuaikan dengan kecacatan (DALYs), yang mempengaruhi 40,9 juta orang pada tahun 2017. Periode kelangsungan hidup kanker mengacu pada periode antara diagnosis dan kematian akibat kanker [3]. Selain itu, analisis kelangsungan hidup sangat berharga bagi dokter, peneliti, pasien, dan pembuat kebijakan.

Mengembangkan model yang akurat dan kuat sangat penting untuk mengidentifikasi tingkat kelangsungan hidup kanker paru-paru [4]. Berbagai algoritme ML telah dikembangkan untuk aplikasi klinis, termasuk hutan acak (RF), algoritme ansambel, pengklasifikasi Naive Bayesian (NB), Mesin vektor dukungan (SVM), jaringan saraf (NN), Pohon Keputusan (DT), dan sejumlah algoritma kepemilikan [5]. Berbagai atribut klinis pasien kanker dengan tingkat kelangsungan hidup mereka dapat dihubungkan menggunakan teknik ML. Selain itu, ML memiliki keuntungan dalam mengurangi beban kerja praktisi medis dan risiko kesalahan manusia. Performa ML yang tinggi menjadikannya alat yang menarik dan memotivasi

penyedia layanan kesehatan. Teknik ML memungkinkan pengembangan model prediktif berdasarkan data kanker untuk memprediksi kelangsungan hidup. Meskipun demikian, tidak ada teknik saat ini yang memenuhi syarat untuk diterapkan pada kumpulan data tertentu [6].

Tinjauan ini bertujuan untuk mengeksplorasi penelitian yang menggunakan teknik ini untuk memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis. Ini memberikan gambaran umum tentang proses prediksi kelangsungan hidup yang digunakan, termasuk sumber data, kelompok pasien, pemrosesan awal data, metode pemilihan fitur, fitur yang dipilih, algoritma ML, dan metode validasi (lihat [Tabel 1](#)).

2. Metode

Untuk mengidentifikasi istilah penelusuran yang digunakan untuk kueri penelusuran "Kelangsungan hidup kanker paru-paru menggunakan pembelajaran mesin". Kami pertama kali mengidentifikasi topik terkait, "Memprediksi kelangsungan hidup kanker paru-paru menggunakan pembelajaran mesin dan data klinis". Setelah itu, kami membagi topik menjadi lima istilah: "memprediksi", "kanker paru-paru", "kelangsungan hidup", "pembelajaran mesin", dan "data klinis" serta menelusuri sinonimnya. Untuk semua proses ini, kami menggunakan Generator Kata Pencarian Perpustakaan Universitas Wollongong untuk menemukan sinonim dan alat Tesaurus Judul Subjek Medis (MeSH). Pencarian

* Penulis yang sesuai. Sekolah Komputasi dan Teknologi Informasi, Universitas Wollongong, NSW, 2500, Australia. *Alamat email:* faaa272@uowmail.edu.au (FA Altuhaifa).

Tabel 1
Pernyataan penting.

Masalah	Meninjau artikel tentang model pembelajaran mesin yang memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis.
Apa yang Sudah Diketahui?	Meskipun para peneliti telah mengeksplorasi penambahan data dan pembelajaran mesin untuk memprediksi kelangsungan hidup, hingga saat ini belum ada teknik yang efektif dalam pra-pemrosesan data dan secara akurat memprediksi kelangsungan hidup kanker paru-paru.
Apa Makalah Ini Menambahkan?	Dalam makalah ini, kami meninjau dan mengeksplorasi penelitian yang menggunakan pembelajaran mesin untuk memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis. Ini memberikan gambaran umum tentang sumber data, kelompok pasien, pra-pemrosesan data, algoritma pemilihan fitur, fitur yang dipilih, algoritma ML, dan metode validasi.

algoritma diimplementasikan dengan bantuan dari pustakawan. Tiga database digital yang dicari: MEDLINE, Scopus, dan Google Scholar. Penelusuran dilakukan pada tanggal 02 April 2022 – 05 Juli 2022. Penelusuran dilakukan berdasarkan kata kunci “kanker paru”, “neoplasma paru”, “tumor paru”, “karsinoma”, “kelangsungan hidup”, “klinis”, “mesin pembelajaran”, “pembelajaran mendalam”, “algoritma”, dan “kecerdasan buatan”. Algoritma pencarian yang digunakan dalam tinjauan sistematis ini disajikan di (lihat [Meja 2](#)).

Pencarian ketiga database digital tersebut melibatkan penyaringan judul artikel. Kami mengecualikan penelitian yang judulnya menunjukkan bahwa kanker yang ditangani berbeda dari kanker paru-paru, atau database berbasis gambar digunakan. Tiga peninjau independen mengevaluasi studi terpilih menggunakan pedoman Item Pelaporan Pilihan untuk Tinjauan Sistematis dan Analisis Meta (PRISMA) melalui COVidence [\[7,8\]](#).

Empat kriteria inklusi digunakan untuk makalah ini: (1) ditinjau sejawat; (2) diterbitkan antara tahun 2000 dan 2022; (3) berdasarkan pemodelan tingkat kelangsungan hidup kanker paru-paru menggunakan algoritma ML (model yang belajar dari data tanpa menggunakan teknik pemrograman tradisional dan fokus pada prediksi tingkat kelangsungan hidup daripada menghitung interval kepercayaan) [\[9\]](#); dan (4) diterbitkan dalam bahasa Inggris. Selain itu, tiga kriteria eksklusif digunakan: (1) tidak menggunakan model ML; (2) diterbitkan sebelum tahun 2000; (3) dan memanfaatkan computerized tomography scan (CT) dan database gambar.

Para peninjau menyaring judul dan abstrak penelitian yang berpotensi relevan dengan menggunakan COVidence dan setiap perbedaan pendapat diselesaikan melalui pertemuan. Setelah penyaringan abstrak dan tinjauan lengkap setiap studi, beberapa pertemuan dilakukan di antara para peneliti

Meja 2

Algoritma pencarian.

Digital database	Algoritma pencarian
MEDLINE	(TI (pasien* DAN ("kanker paru-paru" ATAU "neoplasma paru-paru" ATAU "tumor paru-paru" ATAU "adenokarsinoma paru-paru") ATAU (MM "Neoplasma Paru*")) ATAU AB (pasien* DAN ("kanker paru-paru" ATAU "paru-paru neoplasma" ATAU "tumor paru-paru" ATAU "adenokarsinoma paru-paru") ATAU (MM "Neoplasma Paru*")) ATAU SU (pasien* DAN ("kanker paru-paru" ATAU "neoplasma paru-paru" ATAU "tumor paru-paru" ATAU "adenokarsinoma paru-paru") ATAU (MM "Neoplasma Paru*")) AND (TI ("tingkat kelangsungan hidup" ATAU "hasil kelangsungan hidup" ATAU "hasil klinis") OR AB ("tingkat kelangsungan hidup" ATAU "hasil kelangsungan hidup" ATAU "hasil klinis") ATAU SU ("tingkat kelangsungan hidup" ATAU "hasil kelangsungan hidup" ATAU "hasil klinis")) AND (TI ("kecerdasan buatan" ATAU "AI" ATAU "pembelajaran mesin" ATAU algoritme*) OR AB ("kecerdasan buatan" ATAU "AI" ATAU "mesin pembelajaran" ATAU algoritme*) ATAU SU ("kecerdasan buatan" ATAU "AI" ATAU "pembelajaran mesin" ATAU algoritme*))
Scopus	((MH "Karsinoma, Paru-Paru Non-Sel Kecil") ATAU (MH "Neoplasma Paru") ATAU (MH "Molekul Adhesi Sel-1")) DAN (pembelajaran mesin ATAU kecerdasan buatan atau pembelajaran mendalam atau jaringan saraf atau pembelajaran mesin pada kanker) DAN (tingkat kelangsungan hidup atau hasil kelangsungan hidup atau hasil klinis)
Scopus	JUDUL-KUNCI-ABS ("sabar* DAN "kanker paru-paru" DAN "bertahan hidup*" DAN "pembelajaran mesin" DAN (BATAS-KEPADA (SUBJAREA, "KOMP")) JUDUL-KUNCI-ABS ("sabar* DAN "kanker paru-paru" DAN "bertahan hidup*" DAN "pembelajaran mesin")
Google Sarjana	"pasien" DAN "kelangsungan hidup kanker paru-paru" DAN "pembelajaran mesin"

penulis untuk mengidentifikasi konsep dan data yang akan diekstraksi untuk tinjauan sistemik. Data dari setiap penelitian diekstraksi menggunakan sistem NVivo. Untuk analisisnya, studi-studi tersebut dikategorikan sebagai berikut: database Surveilans, Epidemiologi, dan Hasil Akhir (SIER), database lain, periode waktu kelangsungan hidup (satu, dua, tiga, dan lima tahun), dan akurasi yang digunakan (Area di bawah Kurva ROC (AUC), Root mean square error (RMSE), dan statistik konkordansi (C-statistik)).

3. Hasil

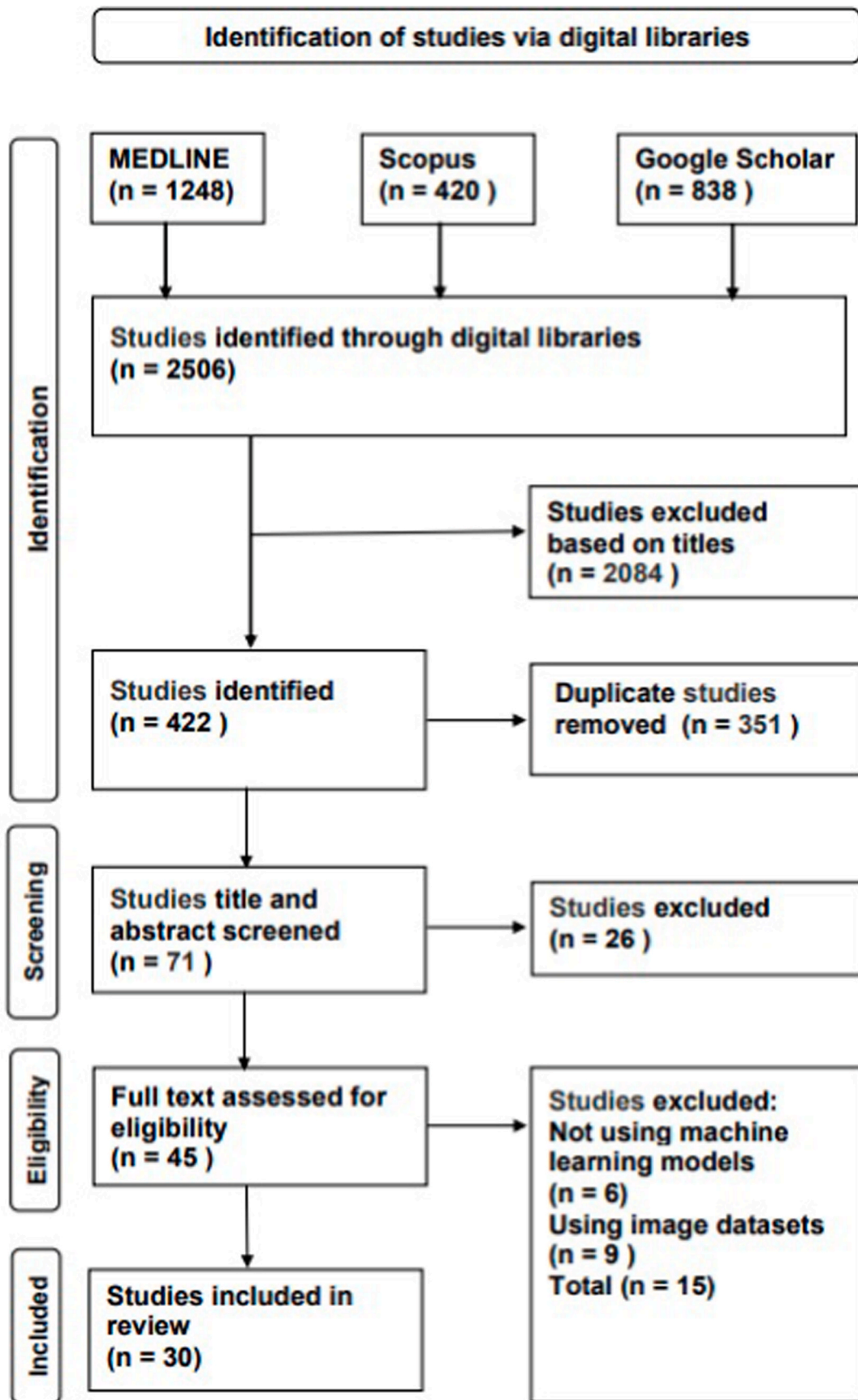
Dari database digital ditemukan 2506 penelitian: 1248 dari MEDLINE, 420 dari Scopus, dan 838 dari Google Scholar. Diantaranya, 2.084 judul dikeluarkan karena judulnya tidak memenuhi kriteria inklusi, dan 422 judul dipertahankan dan diunggah ke COVidence. Setelah itu, duplikat dihapus oleh COVidence, dan 71 penelitian tetap ada, 26 diantaranya dikeluarkan setelah judul dan abstraknya disaring: tujuh (sekitar 26%) tidak menggunakan algoritma ML, tiga belas (sekitar 50%) mengandalkan algoritma berbasis gambar. kumpulan data, dua (kira-kira 10%) memiliki tujuan yang salah, satu (kira-kira 4%) adalah pracetak, dan dua (kira-kira 10%) meneliti jenis kanker yang berbeda. Setelah meninjau teks lengkap dari 45 penelitian yang tersisa, 15 penelitian dikeluarkan: 6 (sekitar 40%) tidak menggunakan algoritma ML, dan 9 (sekitar 60%) mengandalkan kumpulan data berbasis gambar, 30 penelitian dimasukkan dalam tinjauan sistematis ini ([Gambar 1](#)).

Semua penelitian diterbitkan antara tahun 2012 dan 2023. Namun, tidak ada satupun yang dilakukan antara tahun 2013 dan 2016. Selain itu, minat terhadap pengembangan algoritma ML untuk kelangsungan hidup kanker paru-paru berdasarkan data klinis meningkat antara tahun 2019 dan 2023. Sembilan penelitian diterbitkan pada tahun 2020, yang mewakili puncak dalam jumlah penelitian yang dipublikasikan. Selain itu, saat menggali informasi di Nvivo, muncul enam aspek lagi terkait tinjauan sistemik ini: (1) kohort pasien, yaitu sekelompok individu dengan histologi, pengobatan, stadium, atau metastasis kanker paru-paru yang sama; (2) prapemrosesan data, yang dibagi lagi menjadi proses nilai yang hilang, transformasi data, dan konversi data; (3) seleksi fitur, berisi kajian-kajian yang menjelaskan proses seleksi fitur; (4) fitur terpilih, yaitu berisi fitur-fitur yang digunakan sebagai prediktor pada setiap penelitian; (5) ML, yang dibagi lagi menjadi algoritma ML dan model terbaik di setiap penelitian; (6) dan metode validasi, yang memuat informasi mengenai metode validasi yang digunakan pada setiap penelitian, termasuk k-fold cross validation atau pemisahan data.

3.1. Pengumpulan data

Sebuah tinjauan mengungkapkan bahwa 13 menggunakan database US-SEER dan 2 menggunakan data dari Cancer Registry. Selain itu, satu penelitian menggabungkan data SIER dengan data rumah sakit lokal (Rumah Sakit Medis India) [\[27\]](#). Selain itu, empat orang menggunakan data dari rumah sakit setempat, dan dua orang tidak menyebutkan sumber datanya. Selain itu, lima rumah sakit menggunakan data dari rumah sakit setempat untuk validasi eksternal. Namun, ada satu indikasi bahwa penulis tidak dapat memperoleh data validasi eksternal yang serupa dengan data SIER [\[10\]](#). Kumpulan data dalam 30 penelitian berkisar antara 150 hingga 57.254 pasien. Meskipun beberapa penelitian menggunakan dataset yang sama, seperti SEER, namun ukuran datasetnya berbeda. Perbedaan ukuran kumpulan data ini dapat disebabkan oleh perbedaan kelompok pasien. Rincian lebih lanjut mengenai karakteristik data masing-masing penelitian disajikan pada [Tabel 3](#).

Untuk menggunakan sejumlah besar data medis yang dihasilkan melalui data klinis, sangat penting untuk memilih kelompok pasien yang sesuai untuk pertanyaan penelitian spesifik yang sedang diselidiki [\[40\]](#). Dalam masing-masing dari 30 penelitian, kohort pasien tertentu dimasukkan: delapan (26,6%) pasien terpilih dengan kanker paru-paru non-sel kecil (NSCLC) [\[13,15,18,19,25,30,32,37\]](#). [\[10,26,27,29,34,35\]](#) hanya memilih pasien dengan adenokarsinoma, sedangkan [\[22\]](#) pasien terpilih dengan adenokarsinoma dan karsinoma sel skuamosa selain satu lesi primer ganas. Sebaliknya [\[15,36\]](#), pasien terpilih dengan metastasis tulang. 10 (33,3%) penelitian memilih pasien kanker dengan stadium berbeda: stadium I [\[21\]](#), tahap III [\[24,25,37\]](#), tahap I-IIIb [\[15-17\]](#), tahap I-IV [\[10,22\]](#), dan tahapan IB-IIA [\[32\]](#)].



Gambar 1. Diagram alir proses seleksi studi PRISMA.

Tabel 3

Karakteristik data.

Referensi	Sumber data	Waktu	Ukuran data	Negara	Ketersediaan sumber	Validasi eksternal
[11]	MELIHAT November 2008	1998–2001	57.254	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[12]	PERAMAL	2004–2009	10.442	Amerika Serikat	https://www.cancerdata.org/publicati on/	Tak dapat diterapkan
[13]	Data kanker	Bukan dilaporkan	239	Belanda, Inggris, dan Amerika Serikat	https://seer.cancer.gov/ yang didistribusikan melalui	Tak dapat diterapkan
[14]	MELIHAT November 2018	Hanya tahun 2014	Bukan dilaporkan	Amerika Serikat		Tak dapat diterapkan
[10]	PERAMAL	2010–2015	50.687	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[15]	PERAMAL	2010–2015	5973	Amerika Serikat	https://seer.cancer.gov/	Rumah Sakit Afiliasi Medis Chengde Universitas (AHOCMU)
[16]	PERAMAL	2004–2009	10.442	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[17]	PERAMAL	2004–2009	10.442	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[18]	Pendaftaran Multi-Institusi Tiongkok (CMIR)	Bukan dilaporkan	5123	Cina	Tidak dilaporkan	Tak dapat diterapkan
[19]	Data kanker	Bukan dilaporkan	509	Tidak dilaporkan	https://www.cancerdata.org	Tak dapat diterapkan
[20]	PERAMAL	2006–2011	683	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[21]	SIER dan Rumah Sakit Medis India	Bukan dilaporkan	321	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[22]	PERAMAL	2010–2015	17.322	Amerika Serikat	https://seer.cancer.gov/	Paru Shanghai RSUD
[23]	Tidak dilaporkan	Bukan dilaporkan	809		Tidak dilaporkan	Tak dapat diterapkan
[24]	Pasien di Eskişehir Fakultas Kedokteran Universitas Osmangazi.	2007–2018	585	Turki	Tidak dilaporkan	Tak dapat diterapkan
[25]	PERAMAL	Bukan dilaporkan	16.613	Amerika Serikat	https://seer.cancer.gov/	Rumah Sakit dan Institut Kanker Shandong
[26]	Atlas Genom Kanker (TCGA)	Bukan dilaporkan	371	Amerika Serikat	https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	Tak dapat diterapkan
[27]	(TCGA)	Bukan dilaporkan	291	Amerika Serikat	https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	Tak dapat diterapkan
[28]	Canter medis di Tiongkok	2018–2021	150	Cina	Tidak dilaporkan	Canter medis di Tiongkok
[29]	(TCGA)	Bukan dilaporkan	1563	Amerika Serikat	https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	Tidak berlaku (TCGA)
[30]	Pusat Nasional untuk Basis data GEO Informasi Bioteknologi (NCBI).	Bukan dilaporkan	704	Amerika Serikat	https://www.ncbi.nlm.nih.gov/geo/	Kumpulan data MTAB-923
[31]	PERAMAL	2004–2009	38.262	Amerika Serikat	https://seer.cancer.gov/ Tidak	Tak dapat diterapkan
[32]	Tidak dilaporkan	2005–2018	1137	Amerika Serikat	dilaporkan	PERAMAL
[33]	PERAMAL	2998–2001	17.484	Amerika Serikat	https://seer.cancer.gov/ https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	Tak dapat diterapkan
[34]	Kelompok TCGA Kelompok GSE72094	Bukan dilaporkan	506	Amerika Serikat	https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	GSE72094 dan GSE11969 kelompok
[35]	TCGA	Bukan dilaporkan	739	Amerika Serikat	https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	Kumpulan data mutasi GEO dan ICGC
[36]	PERAMAL	2000–2019	10.001	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan
[37]	LIHAT 18 Registri	2000–2017	4517	Amerika Serikat	https://seer.cancer.gov/	Provinsi Shaanxi Rumah Sakit Rakyat, Cina
[38]	TCGA dan GEO	Bukan dilaporkan	2166	Amerika Serikat	https://www.cancer.gov/ccg/research/genome-pengurutan/tcga	GSE42127 dan GSE37745 data
[39]	PERAMAL	2000–2015	28.458	Amerika Serikat	https://seer.cancer.gov/	Tak dapat diterapkan

Selain itu, empat (13,3%) pasien terpilih yang telah menjalani pengobatan [13,18,23,25]. Lebih-lebih lagi [24], pasien terpilih tanpa diagnosis metastasis jauh, dan [14] pasien terpilih yang telah menjalani pengobatan pertama.

3.2. Pemrosesan awal data

Di antara 30 penelitian, hanya dua belas (40%) yang melaporkan proses yang digunakan untuk menangani nilai yang hilang, mulai dari menghilangkan fitur dengan nilai yang hilang hingga imputasi. Enam menangani nilai yang hilang dengan menghapus baris dengan nilai yang hilang [13,22,25,27,34,37]. [13] menyatakan bahwa nilai rata-rata binned hanyalah perkiraan; oleh karena itu, mereka tidak dapat sepenuhnya diandalkan untuk melatih model ML. Sebaliknya, empat metode yang diperhitungkan diadopsi untuk menangani nilai-nilai yang hilang [19,23,28, 33]. Khususnya [19,33], nilai yang hilang diperhitungkan menggunakan median/mode dan imputasi mean, sedangkan [23,28] memperhitungkannya berdasarkan informasi yang diberikan oleh beberapa kandidat penting dalam kumpulan data [14]. terapan

kedua metode (menghapus nilai yang hilang dan imputasi) untuk menangani nilai yang hilang dan menghapus fitur dengan lebih dari 20% nilai yang hilang. Jika nilai yang hilang dalam suatu fitur kurang dari 20%, nilai tersebut diperhitungkan menggunakan imputasi median/mode. Studi terakhir [18] menggunakan algoritma ReliefF untuk mengatasi masalah nilai yang hilang selama proses pemilihan fitur. Hanya tujuh (23,3%) studi yang menjelaskan proses yang digunakan untuk transformasi data, yang bervariasi antara normalisasi, penskalaan, dan standardisasi. Di antara tujuh penelitian ini, empat menggunakan normalisasi [13,20,37,39]. Dalam Ref. [13] normalisasi dilakukan dengan menggunakan metode min-max, sedangkan dalam Ref. [20] nilai numerik dinormalisasi dalam rentang -1 hingga 1 [37,39]. tidak melaporkan metode normalisasi. Alternatifnya, metode variabel penskalaan digunakan dalam Referensi. [14,17]. Menggunakan variabel penskalaan [17], mencegah model yang terlalu sensitif terpengaruh oleh pengukuran jarak yang miring. Untuk mempercepat proses pelatihan [15], mengadopsi metode standarisasi untuk umur dan ukuran. Selain itu, konversi data dilaporkan dalam empat (13,3%) penelitian; mereka mengadopsi pengkodean one-hot [14,15,

37,39],Yang memungkinkan untuk menghitung jarak lebih akurat. Sebaliknya [17], tidak mengadopsi pengkodean one-hot karena kumpulan data yang mereka gunakan berisi data dengan populasi yang jarang, dan untuk meminimalkan jumlah pemrosesan awal yang diperlukan.

3.3. Fitur seleksi dan rekayasa

Secara total, 19 (sekitar 63,3%) studi mempertimbangkan pemilihan fitur selama prapemrosesan data, termasuk ANOVA, ReliefF, operator penyusutan dan seleksi absolut terkecil (LASSO), hutan kelangsungan hidup acak, peningkatan gradien ekstrim (XGBoost), model bahaya proporsional Cox, chi -tes kuadrat, T-test, matriks korelasi, dan gradien terintegrasi [10,12–16,18,19,23–25,27,29,30,33,34,37–39]. Dalam Ref. [12], fitur dipilih menggunakan ANOVA, interval kepercayaan dua sisi yang mengukur tingkat kepercayaan 95%, dan fitur apa pun yang tidak dilaporkan oleh ANOVA dihilangkan. Di antara 19 penelitian yang menggunakan metode pemilihan fitur, dua diantaranya mengadopsi ReliefF, yang mengurangi dimensi dan membuat data lebih dapat diinterpretasikan [13,18]. Ketika [27], dipilih Bantuan [14]. mengadopsi LASSO. Dalam Ref. [10], metode RF digunakan untuk menilai pentingnya fitur dan hubungan antara jumlah pohon dan tingkat kesalahan. Dalam Referensi. [15,39], XGBoost digunakan untuk memperkirakan pentingnya fitur. Dalam Referensi. [16,29,30,34,38], model bahaya proporsional Cox diadopsi, sedangkan dalam Pustaka. [23], chi-kuadrat dan T-metode tes diadopsi. Akhirnya, empat penelitian mengadopsi metode matriks korelasi [19,24,25,33] dan satu penelitian [37] mengadopsi gradien terintegrasi. (Tabel 4).

Di antara 30 penelitian, 32 fitur dipilih untuk memprediksi kelangsungan hidup kanker paru-paru. Jumlah fitur yang dipilih dalam setiap studi bervariasi antara 4 dan 15. Usia (24 studi (80%)), jenis kelamin (21 studi (70%)) dan tahap N (16 studi (53,3%)) adalah fitur yang paling banyak dipilih, diikuti oleh tahap (15 studi (50%)) dan tahap T (13 studi (43,3%)) [11]. melaporkan bahwa 11 fitur (tempat lahir, usia, tingkatan, pembedahan, terapi radiasi, stadium N, stadium, keganasan, tumor, jumlah primer, dan histologi) mengkodekan informasi secara wajar, yang mencegah hilangnya akurasi yang substansial. Selain itu, tempat lahir tampaknya menjadi fitur penting di antara 11 fitur [14]. menemukan bahwa beberapa ciri (jenis kelamin, pembedahan, terapi radiasi, ukuran tumor, dan stadium N) berdampak pada prediksi lokal; namun, tingkat global tidak terpengaruh secara signifikan. Hal ini menyiratkan bahwa dampak fitur-fitur ini terutama ditentukan oleh interaksinya dengan beberapa fitur lainnya [15]. menemukan bahwa ukuran tumor adalah faktor kelangsungan hidup yang paling signifikan, diikuti oleh usia, kemoterapi, metastasis hati, pembedahan, tingkatan, dan ras, yang memiliki pengaruh sedang terhadap kelangsungan hidup. Selain itu, stadium N, tipe histologis, dan jenis kelamin tidak terlalu mempengaruhi kelangsungan hidup. [18]. menemukan bahwa usia, jenis kelamin, dan histologi adalah fitur yang berguna dan bermakna untuk prediksi, serta mengurangi biaya komputasi dan meningkatkan akurasi analisis. Sebaliknya [19], menemukan korelasi yang kuat antara dosis total tumor dan waktu pengobatan secara keseluruhan (Tabel 5aDanTabel 5b). Selain itu, lima penelitian melaporkan bahwa fitur paling signifikan untuk memprediksi waktu kelangsungan hidup adalah pembedahan, stadium [10,11], tipe histologi [28], tahap T [15], waktu perawatan [24], ukuran total tumor [19], dan N kelenjar getah bening [24]. [21] menemukan bahwa kadar hemoglobin bersamaan dengan TNM stadium 1 dapat membantu memprediksi kelangsungan hidup kanker paru-paru [24]. diidentifikasi T

stadium, pembedahan, lokasi kelenjar getah bening, dan faktor histopatologis sebagai fitur penting yang mempengaruhi waktu kelangsungan hidup. Lebih lanjut, 7 penelitian telah mengidentifikasi fitur genetik untuk memprediksi kelangsungan hidup kanker paru-paru, termasuk 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, dan GANC [35].

3.4. Algoritma ML

Empat puluh tujuh algoritma ML yang berbeda digunakan dalam 30 penelitian untuk memprediksi kelangsungan hidup kanker paru-paru (Tabel 6). Namun, prediksinya bervariasi dari enam bulan hingga lima tahun di setiap penelitian. Secara keseluruhan, RF adalah model yang paling banyak digunakan, yaitu 17 (36,1%) dari 47 model. Itu digunakan dalam 17 (56,6%) dari 30 penelitian [10–12,14–16,18,19,23–25,27,33–35,38,39]. Beberapa model, seperti pengelompokan hierarki [17] dan LogitBoost [11], digunakan hanya dalam satu penelitian. Selain itu, 3 penelitian (10%) menggunakan 10 algoritma ML yang berbeda untuk memprediksi kelangsungan hidup kanker paru-paru dalam 6 bulan [11, 12,14], 6 (20%) menggunakan 14 algoritma ML berbeda untuk memprediksi kelangsungan hidup dalam 1 tahun [10,11,15], 6 (20%) menggunakan 22 algoritma ML yang berbeda untuk memprediksi kelangsungan hidup dalam 2 tahun [11–14,19,24], 4 (13,3%) menggunakan 13 algoritma ML yang berbeda untuk memprediksi kelangsungan hidup dalam 3 tahun [10,22,26,27,29,39], dan 22 (73,3%) menggunakan 39 algoritma ML yang berbeda untuk memprediksi kelangsungan hidup dalam 5 tahun [10–12,14,16–18,20–24,29–37,39].

3.5. Evaluasi model dan metrik kinerja

Di antara 30 penelitian, 12 (40%) menggunakan validasi silang k-fold, 16 (53,3%) menggunakan pemisahan data, dan 4 (13,3%) tidak melaporkan metode validasi (Lampiran B). Dalam dua penelitian ini, validasi silang k-fold dan pemisahan data diterapkan [23,24]. Validasi silang K-fold bervariasi antara validasi 3 dan 10 silang dan rasio pemisahan data juga bervariasi (8:2, 9:1, dan 7:3). Tujuan utama validasi silang k-fold di Pustaka. [16] adalah untuk mengurangi overfitting daripada optimasi parameter [24]. menggunakan validasi silang k-fold untuk menentukan apakah model yang dilatih dapat menggeneralisasi data baru, dan untuk mengidentifikasi bias seleksi atau masalah kepatuhan yang berlebihan. Dalam Ref. [23], pemisahan data diadopsi untuk menghindari overfitting model, yang diperlukan untuk memperkirakan performa model secara akurat.

Akurasi algoritme ML bervariasi di 30 penelitian, termasuk RMSE (7 penelitian (23,3%)), AUC (20 penelitian (60,6%)), dan C-statistik (4 penelitian (13,3%)). Metrik ini memberikan wawasan tentang berbagai aspek performa model. RMSE (Root Mean Square Error), yang mengukur perbedaan rata-rata antara nilai prediksi dan nilai aktual, biasanya digunakan dalam tugas regresi. Ukuran ini mengkuantifikasi kesalahan prediksi dan biasanya digunakan untuk menentukan apakah model regresi akurat [41] AUC (Area Di Bawah Kurva ROC) adalah metrik yang umum digunakan dalam tugas klasifikasi biner. Ini mengukur area di bawah kurva Karakteristik Operasi Penerima (ROC) untuk menilai kemampuan pengklasifikasi dalam membedakan antara sampel positif dan negatif. Secara umum, AUC yang lebih tinggi menunjukkan kinerja diskriminasi yang lebih baik [42]. Statistik C adalah metrik yang digunakan dalam analisis kelangsungan hidup, juga dikenal sebagai indeks konkordansi atau area di bawah kurva ROC untuk analisis kelangsungan hidup. Ini mengevaluasi keakuratan prediksi model kelangsungan hidup dengan mengukur kemampuan untuk menentukan peringkat subjek berdasarkan waktu kelangsungan hidup mereka. Statistik C yang lebih tinggi menunjukkan kinerja prediktif yang lebih baik [43].

Studi menggunakan RMSE fokus pada keakuratan prediksi waktu kelangsungan hidup berkelanjutan. Studi-studi ini biasanya menggunakan waktu terjadinya peristiwa dan waktu kelangsungan hidup sebagai variabel target, yang merupakan variabel numerik kontinu. Tujuannya adalah untuk mengembangkan model pembelajaran mesin yang bisa

Tabel 4
Algoritma pemilihan fitur.

Fitur metode pemilihan	Referensi
ANOVA	[12]
BantuanF	[13,18]
Lega	[27]
LASO	[14]
Metode hutan kelangsungan hidup acak	[10]
XGBoost	[15,39]
Model bahaya proporsional Cox	[16,29,30,34,38] [
Chi-kuadrat dan T-uji matriks korelasi	23]
Gradien terintegrasi	[19,24,25,33] [37]

Tabel 5a

Fitur yang dipilih berdasarkan studi.

Fitur/Studi	[11]	[12]	[13]	[14]	[10]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]	[25]
Usia	✓							✓		✓		✓		✓	✓	✓
Seks							✓	✓		✓		✓		✓	✓	✓
Balapan											✓					✓
Operasi	✓						✓	✓								✓
Terapi radiasi	✓						✓			✓				✓		✓
Kemoterapi								✓		✓						✓
Panggung	✓		✓						✓			✓				
Ganas	✓													✓		
Jumlah pemilihan pendahuluan	✓		✓											✓		
Histologi	✓									✓						✓
Ukuran tumor						✓				✓		✓				✓
Nilai	✓		✓							✓		✓		✓		
tahap T			✓		✓				✓	✓		✓		✓		✓
tahap N	✓		✓		✓				✓	✓		✓		✓		✓
Stadium M (metastasis paru-paru, hati, otak)									✓	✓		✓		✓		✓
Lateralitas										✓						✓
Dosis total tumor					✓						✓					
Beban tumor					✓											
Waktu perawatan secara keseluruhan					✓						✓					
Situs utama										✓		✓				✓
Fitur genetik																
Status asuransi										✓						
Status pernikahan										✓						
Tingkat hemoglobin												✓				
Jumlah sel darah merah												✓				
Sistem kekebalan tubuh pasien kanker paru												✓				
Status pernikahan													✓			
Jenjang pendidikan															✓	
Pendapatan keluarga bulanan															✓	
Indeks massa tubuh															✓	
Aktivitas fisik															✓	
Tempat lahir	✓															

Tabel 5b

Fitur yang dipilih berdasarkan studi.

Fitur/Studi	[26]	[27]	[28]	[29]	[30]	[31]	[32]	[33]	[34]	[35]	[36]	[37]	[38]	[39]
Usia			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Seks			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Balapan											✓			✓
Operasi								✓				✓		
Terapi radiasi												✓		
Kemoterapi											✓	✓		
Panggung	✓				✓	✓		✓		✓		✓	✓	✓
Ganas													✓	
Jumlah pemilihan pendahuluan														
Histologi						✓	✓					✓		✓
Ukuran tumor							✓		✓			✓		
Nilai								✓						
tahap T										✓	✓		✓	
tahap N								✓		✓	✓		✓	
Tahap M (metastasis ke paru-paru, hati, otak)			✓							✓	✓		✓	
Tempat utama							✓					✓		
Fitur genetik	✓	✓		✓	✓				✓	✓			✓	
Indeks massa tubuh			✓											

secara akurat memprediksi durasi kelangsungan hidup yang tepat untuk masing-masing pasien. Metrik RMSE digunakan untuk menilai kinerja model-model ini dengan mengukur perbedaan rata-rata antara nilai waktu kelangsungan hidup yang diprediksi dan nilai waktu kelangsungan hidup sebenarnya. RMSE yang lebih rendah menunjukkan akurasi yang lebih baik dalam memprediksi waktu kelangsungan hidup berkelanjutan. Sebaliknya, penelitian yang menggunakan AUC mengevaluasi kemampuan model untuk mengklasifikasikan pasien ke dalam kategori kelangsungan hidup yang berbeda berdasarkan jangka waktu tertentu, seperti apakah pasien dapat bertahan hidup selama 1 tahun, 2 tahun, 3 tahun, atau 5 tahun. Dalam penelitian ini, variabel target diubah menjadi variabel biner, yang menunjukkan apakah pasien termasuk dalam kategori kelangsungan hidup tertentu (misalnya, bertahan hidup selama jangka waktu tertentu atau tidak bertahan hidup setelah jangka waktu tersebut). Metrik AUC (Area Under the Curve) digunakan untuk menilai kemampuan model dalam membedakan berbagai kategori kelangsungan hidup. Ini mengevaluasi kemampuan model untuk menentukan peringkat probabilitas kelangsungan hidup pasien

dengan benar dan untuk mengklasifikasikan pasien ke dalam kelompok kelangsungan hidup yang berbeda berdasarkan interval waktu yang telah ditentukan. Selain itu, penelitian ini menggunakan C-statistik untuk mengevaluasi kemampuan model dalam menentukan peringkat pasien berdasarkan waktu kelangsungan hidup mereka. Sepanjang penelitian, variabel target tetap berupa waktu terjadinya peristiwa atau waktu bertahan hidup, yang mewakili waktu hingga kematian terjadi. Statistik C, juga dikenal sebagai indeks kesesuaian atau area di bawah kurva ROC, digunakan untuk mengevaluasi kekuatan diskriminatif model dalam menentukan peringkat pasien berdasarkan perkiraan waktu kelangsungan hidup mereka. Pada dasarnya, ini adalah ukuran kemampuan model untuk mengurutkan pasien pada waktu yang tepat berdasarkan tingkat kelangsungan hidup mereka yang sebenarnya.

Ikhtisar algoritma ML terbaik yang digunakan dalam setiap penelitian untuk memprediksi kelangsungan hidup kanker paru-paru disajikan dalam [Lampiran A](#), termasuk akurasi, waktu bertahan, dan ukuran kumpulan datanya. Ikhtisar algoritma ML yang digunakan dalam 16 penelitian untuk memprediksi kelangsungan hidup kanker paru-paru,

Tabel 6
Jenis dan algoritma ML.

Tipe ML	Algoritma ML
Berbasis pohon	16 RF, 1 IRF, 1 J48 DT, 5 DT, 1 AdaBoost, 2 pohon peningkatan gradien (GBT), 3 XGBoost, 1 peningkatan gradien cahaya (GBT Ringan), 1 logitBoost, 1 bagging, 1 Subruang acak.
Jaringan syaraf- berdasarkan	3 jaringan saraf tiruan (JST), 2 jaringan saraf perceptron tunggal (SPNN), 1 jaringan saraf multilayer (MNN), 1 jaringan saraf konvolusi (CNN), 1, Jaringan saraf berulang (RNN), 1, jaringan saraf regresi umum (GRNN), 5 Deep Learning (DL), 1 Multilayer perceptron (MLP), 1 jaringan saraf dalam (DNN).
Bayesian	4 naïve bayes (NB), 1 NB Gaussian K-base, 1 Gaussian NB, 2 Bayes Net.
Berbasis linier	8 mesin vektor pendukung (SVM), 3 regresi linier (LR), 3 regresi logistik, 1 regresi logistik multitask (MLTLR), 1 Regresi Ridge, 1 Regresi vektor pendukung garis (SVR), 2 Operator Penyusutan dan Seleksi Absolut Terkecil (LASSO), 1 pengelompokan hierarki (HC), 1
Berbasis clustering	pengelompokan berbasis model (MBC), 2 rata-rata K, 1 faktorisasi matriks non-negatif, 1 mandiri- peta penyelenggaraan, 1 kluster pendukung rektor (SVC).
Lainnya	1 pengklasifikasi malas (LWL), 1 pengklasifikasi meta (ASC), 1 Pembelajaran Aturan (OneR), 2 Kelompok Penelitian Onkologi Kerangka (SORG), 1 k-Nearest Neighbors (KNN), 4 regresi cox, 1 jaringan saraf cox (Cox NN).

termasuk akurasi rata-ratanya dan akurasi yang dilaporkan dalam setiap penelitian, disajikan dalam [Lampiran C](#).

4. Diskusi

Tinjauan ini dilakukan untuk merangkum penerapan metode data mining dan ML dalam memprediksi kelangsungan hidup kanker paru-paru menggunakan data klinis (data terstruktur) untuk memandu penelitian di masa depan.

Risiko bias dinilai menggunakan model prediksi risiko alat penilaian bias (PROBAST). Alat PROBAST dirancang untuk menilai risiko bias dan penerapan model prediksi prognostik dan diagnostik (lihat [Lampiran D](#)) [44]. PROBAST mengevaluasi kualitas model pembelajaran mesin yang dikembangkan dan divalidasi untuk memprediksi kelangsungan hidup kanker paru-paru. Secara keseluruhan, kualitas penelitian yang dimasukkan dalam ulasan ini adalah sedang, dan diperlukan penelitian lebih lanjut yang dirancang dengan baik untuk memvalidasi prediksi pembelajaran mesin terhadap kelangsungan hidup kanker paru-paru. Secara khusus, model perlu ditingkatkan dalam pra-pemrosesan, karena ukuran sampel dan bagian data yang hilang tidak jelas di 17 model (56,6% studi), dan berisiko tinggi di 5 model (16,6% studi).

Berdasarkan hasil review ini, minat terhadap penelitian terkait kelangsungan hidup kanker paru-paru menggunakan algoritma ML berdasarkan kumpulan data klinis telah meningkat; itu adalah yang tertinggi pada tahun 2020 ketika sembilan penelitian diterbitkan. Namun, jumlah penelitian dari tahun 2023 terbatas karena kami melakukan tinjauan ini pada tahun 2022–2023 dan hanya dapat mengekstrak penelitian yang dipublikasikan sebelum Juli 2023. Selain itu, jumlah penelitian terkait "Memprediksi kelangsungan hidup kanker paru-paru menggunakan pembelajaran mesin dan data klinis " rendah karena sebagian besar penelitian menggunakan kumpulan data berbasis gambar dan sangat sedikit peneliti yang melakukan penelitian tentang kanker paru-paru. Hal ini konsisten dengan temuan [6,45], yang menunjukkan bahwa sebagian besar penelitian menggunakan kumpulan data berbasis gambar dan terutama berfokus pada kanker payudara. Hal ini mungkin terjadi karena kanker payudara merupakan jenis kanker yang paling umum terjadi. Namun, kanker paru-paru adalah penyebab utama kematian terkait kanker [1]. Oleh karena itu, jumlah publikasi tahunan yang memperkirakan kelangsungan hidup kanker paru-paru menggunakan data klinis diperkirakan akan meningkat di masa depan.

4.1. Pengumpulan data

Hasil tinjauan ini juga menunjukkan bahwa sebagian besar penelitian (sekitar 43,3%) terutama menggunakan kumpulan data SIER yang berbasis di AS. Peneliti diharapkan menggunakan data SIER karena ketersediaannya. Namun hasil review ini menunjukkan bahwa algoritma ML berbasis private

kinerja kumpulan data hampir sama baiknya dengan kumpulan data yang didasarkan pada data SIER. Namun, kumpulan data rumah sakit lokal memiliki ukuran sampel yang lebih kecil dibandingkan kumpulan data publik. Selain itu, ukuran sampel minimum yang digunakan dalam penelitian yang ditinjau adalah 150, dan tiga belas penelitian menggunakan kumpulan data yang mencakup kurang dari 1000 pasien. Menggunakan kumpulan data kecil untuk pelatihan model mengakibatkan overfitting dan mengurangi kemampuan generalisasi model ke data baru bila digunakan dengan model yang tidak sesuai [46,47]. Penelitian di masa depan harus menggunakan algoritma ML yang sesuai untuk kumpulan data kecil, seperti NB atau RF [48,49]. Setiap data SIER berbeda dalam geografi, bulan dalam tanggal, beberapa bidang demografi, dan informasi pengobatan seperti kemoterapi. Hasilnya, setiap penelitian dalam review ini menggunakan data untuk melatih modelnya yang berbeda dari penelitian lainnya meskipun menggunakan data SEER [50]. Tinjauan ini juga menemukan bahwa sebagian besar pasien kanker paru-paru dipilih secara umum tanpa menyebutkan jenis kanker paru-paru apa pun. Namun, delapan (26,6,2%) penelitian secara khusus menargetkan pasien NSCLC. Hal ini mungkin karena 84% kasus kanker paru-paru di seluruh dunia adalah NSCLC [51]. Selain itu, tinjauan ini menemukan bahwa belum ada penelitian yang menggunakan ML berdasarkan data klinis untuk memprediksi kelangsungan hidup pasien dengan kanker paru-paru kecil. Selain itu, belum ada penelitian yang menargetkan usia tertentu.

Sesuai dengan artikel penelitian Ref. [11], intervensi bedah dan terapi radiasi adalah dua pilihan pengobatan yang tersedia untuk penelitian mereka. Berdasarkan temuan penelitian, pendekatan pengobatan yang paling tepat untuk pasien kanker paru-paru bergantung pada faktor-faktor seperti stadium penyakit dan kesehatan umum pasien. Temuan signifikan dari penelitian ini menunjukkan bahwa pasien yang menjalani intervensi bedah memiliki tingkat kelangsungan hidup yang lebih tinggi dibandingkan mereka yang tidak. Meskipun demikian, penelitian ini tidak memberikan panduan eksplisit mengenai penerapan pembedahan atau terapi radiasi yang tepat sesuai dengan stadium kanker paru-paru. Menurut [14] pasien yang menjalani intervensi bedah memiliki respons lebih tinggi terhadap masa kelangsungan hidup yang lebih lama dibandingkan dengan mereka yang tidak menjalani operasi. Hal ini menunjukkan bahwa pembedahan memainkan peran penting dalam meningkatkan peluang kelangsungan hidup pasien kanker paru-paru. Meskipun penelitian ini mengakui pentingnya intervensi bedah dan hubungannya dengan tingkat kelangsungan hidup yang lebih baik, penelitian ini tidak memberikan panduan eksplisit mengenai penerapan operasi atau terapi radiasi yang tepat berdasarkan stadium kanker paru-paru.

Temuan penelitian ini [10] menunjukkan bahwa pasien yang menerima kemoterapi memiliki kemungkinan bertahan hidup yang lebih tinggi dibandingkan dengan mereka yang tidak menerima kemoterapi. Namun penelitian tersebut tidak memberikan informasi spesifik mengenai jenis kemoterapi yang digunakan, durasi pengobatan, atau karakteristik spesifik pasien yang memengaruhi efektivitas kemoterapi. Rincian ini memerlukan penyelidikan dan analisis lebih lanjut. Selain itu, penelitian ini menunjukkan bahwa hal ini mungkin bukan merupakan faktor risiko independen yang signifikan terhadap prognosis pada pasien adenokarsinoma paru. Namun, mengingat keterbatasan penelitian, penelitian dan analisis lebih lanjut diperlukan untuk menarik kesimpulan yang lebih pasti mengenai dampak pembedahan terhadap hasil akhir pasien.

Pembelajaran [15] menunjukkan bahwa kemoterapi memiliki efek menguntungkan pada prognosis pasien pada pasien NSCLC dengan BM. Namun, penting untuk dicatat bahwa obat kemoterapi spesifik dan protokol yang digunakan dalam pengobatan tidak disebutkan dalam penelitian ini, sehingga membatasi eksplorasi lebih lanjut mengenai hubungan antara metode pengobatan dan prognosis. Selain itu, disebutkan bahwa Terapi radiasi dilakukan pada 59,0% pasien tetapi tidak memberikan informasi spesifik tentang dampaknya terhadap prognosis atau hasil kelangsungan hidup. Di sisi lain, penelitian [20–22, 25,28,29] menyebutkan bahwa dataset yang digunakan mencakup terapi radiasi atau pembedahan, namun tidak memberikan rincian atau temuan spesifik mengenai nilai prediksi atau dampak terapi radiasi atau pembedahan terhadap kelangsungan hidup pasien.

4.2. Pemrosesan awal data

Data berkualitas tinggi sangat penting untuk pengoperasian dan kinerja algoritma ML yang tepat [52]. Oleh karena itu, penting bagi peneliti untuk memproses data terlebih dahulu sebelum menerapkan model regresi atau klasifikasi. Nilai yang hilang sering ditemui ketika

menganalisis dan menafsirkan kumpulan data selama pengembangan model ML. Tinjauan ini menemukan bahwa lebih dari 50% penelitian (sekitar 60%) tidak melaporkan proses yang digunakan untuk menangani data yang hilang. Meskipun tidak ada metode yang sempurna untuk menangani nilai yang hilang, penting untuk menangani data yang hilang sebelum analisis data untuk menghindari bias atau kesimpulan yang salah arah. [53]. Dengan demikian, pemrosesan data harus mempertimbangkan nilai-nilai yang hilang. Beberapa algoritma, seperti imputasi, dapat digunakan untuk menangani nilai yang hilang [54]. Ada beberapa metode imputasi, antara lain imputasi tunggal, regresi, imputasi maksimum/minimum, dan imputasi ganda, serta imputasi menggunakan ML, seperti model KNN. Namun, penggunaan metode imputasi tunggal untuk data berdimensi tinggi atau data besar dapat berdampak buruk terhadap performa model ML, dan dalam banyak kasus, menyebabkan bias data [53]. Namun, dua penelitian [19,23] yang menggunakan imputasi tunggal terdiri dari ukuran data kecil, dan lainnya [14] tidak menyertakan informasi mengenai ukuran data yang digunakan. Selain itu, seperti yang ditunjukkan dalam Ref. [55], imputasi yang rumit, seperti imputasi ganda, lebih efektif dalam mereproduksi nilai yang hilang. Dengan menggunakan beberapa imputasi, nilai yang hilang dapat diperkirakan dengan cara yang tidak bias, dan perkiraan kesalahan standar yang akurat dapat diperoleh. [56]. Oleh karena itu, penggunaan metode imputasi yang lebih kompleks untuk nilai yang hilang dapat bermanfaat bagi penelitian di masa depan yang menggunakan ML untuk prediksi kelangsungan hidup kanker paru-paru [55]. Satu studi [18] menggunakan ReliefF, yang didasarkan pada imputasi KNN [57]. Imputasi KNN bekerja dengan baik untuk data kontinu dan diskrit dan juga dapat menangani beberapa nilai yang hilang [53]. Namun, ia memiliki beberapa keterbatasan, termasuk akurasi yang rendah. Pendekatan lain untuk menangani nilai yang hilang adalah penghapusan, yang melibatkan penghapusan semua entri yang hilang dari kumpulan data selama tahap analisis atau pelatihan model ML [53]. Di antara penelitian yang menangani nilai yang hilang dalam kumpulan datanya, enam [13,22,25,27,34,37] menggunakan pendekatan penghapusan. Meskipun ini adalah cara mudah untuk menangani nilai yang hilang, menghapus entri bersama dengan nilai yang hilang akan mengubah distribusi data, yang dapat menyebabkan bias pada hasil model ML [53]. Karena performa algoritma ML bergantung pada distribusi dan noise data pelatihan [58], perubahan distribusi data sebelum melatih model harus dihindari untuk menjaga keandalan prediksi. Jika seorang peneliti mengadopsi pendekatan penghapusan, mereka harus memastikan bahwa distribusi data tetap tidak terpengaruh atau menggunakan algoritma ML yang kuat terhadap bias.

Transformasi data dilakukan selama prapemrosesan data, dimana data diubah dari satu format atau struktur ke format atau struktur lainnya [59]. Namun, 23 penelitian (sekitar 76,6%) tidak melaporkan metode transformasi fitur yang digunakan. Algoritme ML tidak dapat bekerja secara efektif jika data tidak diubah karena kemungkinan besar akan mengalami inkonsistensi dan bias. Normalisasi adalah salah satu metode yang digunakan untuk meningkatkan analisis data [60]. Karena normalisasi meningkatkan keseluruhan database dan konsistensi data serta mengurangi data yang berlebihan, penelitian [13,20,37,39] yang menerapkan normalisasi kuat terhadap redundansi data [60]. Demikian pula, standarisasi dan penskalaan data dapat meningkatkan algoritma ML dan membuat aliran data lebih efisien [61]. Standarisasi adalah proses penskalaan data dengan mengubahnya menjadi variabel normal standar ($\text{mean} = 0$, $\text{variance} = 1$) [50]. Karena standarisasi berguna ketika data mengikuti distribusi Gaussian atau normal [50], disarankan untuk memvisualisasikan data atau memeriksa distribusi data untuk menentukan apakah normalisasi sesuai. Namun normalisasi dan standarisasi data tidak menyelesaikan masalah outlier yang besar. Oleh karena itu, outlier harus dihilangkan sebelum normalisasi atau standarisasi.

Pengkodean one-hot digunakan dalam empat penelitian [13,20,37,39], dan tidak satupun dari mereka mencapai hasil yang memuaskan; namun, hal ini tidak dapat dianggap tidak efektif karena parameter dan pengaturan lain dapat memengaruhi performa algoritma ML. Penulis tidak menjelaskan alasan penggunaan pengkodean one-hot atau sejauh mana hal itu mempengaruhi kinerja. Namun, hal ini dapat menimbulkan konsekuensi buruk bila digunakan untuk fitur dengan tingkat kardinalitas yang tinggi karena akan menghasilkan dimensi yang tinggi, yang dapat menyebabkan masalah memori dan komputasi [62]. Selain itu, penelitian [17] hanya menggunakan fitur numerik atau yang dapat diubah menjadi numerik dengan mudah. Dengan demikian, dikembangkan model dalam hal ini

Penelitian ini bisa menjadi bias karena tidak mencantumkan fitur-fitur kategorikal, yang mungkin mempunyai peran penting dalam memprediksi kelangsungan hidup kanker paru-paru. Seorang peneliti dapat mengubah fitur kategorikal menjadi fitur numerik dengan mengkodekan setiap variabel alih-alih menghilangkannya.

4.3. Fitur seleksi dan rekayasa

Penggunaan fitur yang tidak terkait dengan kelangsungan hidup kemungkinan besar tidak akan meningkatkan akurasi prediksi dan malah akan meningkatkan beban komputasi. Oleh karena itu, memilih fitur yang secara signifikan berhubungan dengan kelangsungan hidup kanker paru-paru dapat meningkatkan akurasi prediksi algoritma ML. Hasil tinjauan ini menunjukkan bahwa sebelas penelitian (sekitar 36,6%) tidak menggunakan algoritma pemilihan fitur apa pun. Namun demikian, algoritma ML dari sebelas penelitian ini memiliki kinerja yang hampir sama baiknya dengan algoritma yang menggunakan metode pemilihan fitur. Beberapa penelitian menggunakan gambaran spesifik sebagai prediktor kelangsungan hidup kanker paru-paru, seperti kadar hemoglobin, jumlah sel darah merah [21], dan indeks massa tubuh (BMI) [23]; namun, performa algoritma ML ini tidak menunjukkan peningkatan apa pun dibandingkan dengan penelitian lainnya. Namun demikian, Agrawal dkk. [17] menunjukkan bahwa meskipun "tempat lahir" dapat meningkatkan prediksi kelangsungan hidup kanker paru-paru, "pembedahan" adalah prediktor kelangsungan hidup yang paling penting. Selain itu, setiap penelitian melaporkan fitur berbeda sebagai yang paling signifikan; namun, hal ini dapat dikaitkan dengan perbedaan dalam algoritma ML atau kelompok pasien yang diadopsi dalam setiap penelitian. Di masa depan, para peneliti harus mengeksplorasi dan membandingkan fitur-fitur yang signifikan untuk prediksi kelangsungan hidup setiap histologi kanker paru-paru. Dengan demikian, peneliti dapat mengidentifikasi ciri-ciri universal yang mungkin berlaku untuk semua jenis histologi kanker paru-paru untuk memprediksi kelangsungan hidup.

4.4. Algoritma ML

RF adalah metode ML yang paling banyak digunakan di antara semua penelitian. Namun, performa ML di seluruh penelitian tidak dibandingkan karena perbedaan dalam proses pembersihan data, database, kelompok pasien, fitur, dan prosedur validasi yang digunakan. Performa algoritme ML bergantung pada kualitas data yang digunakan, dan jika kualitas data buruk, performa model ML buruk, sehingga menghasilkan prediksi yang tidak akurat [63]. Namun demikian, belum ada penelitian yang melakukan pemeriksaan mendalam terhadap kualitas data dan dampak langkah-langkah yang dilakukan selama prapemrosesan data pada model ML. Selain itu, setiap penelitian memiliki model ML dengan performa terbaik yang berbeda; namun, RF bertekad menjadi yang terbaik dalam tujuh studi [10,12,19,30,33–35]. Selain itu, RF dapat dilatih dengan baik pada kumpulan data kecil [48], dan dapat menangani struktur data dengan ruang fitur yang kompleks dan berskala besar [64]. Namun demikian, hyperparameter RF bersifat kompleks, yang dapat mengakibatkan overfitting [65]. DT, RF, bagging, dan XGBoost adalah model berbasis pohon [66]. Model berbasis pohon dapat menggunakan prediktor kualitatif dalam proses prediksinya tanpa membuat variabel dummy [66]; Oleh karena itu, model berbasis pohon dianggap lebih efektif untuk mendeteksi hubungan non-monotonik atau non-linier antara variabel dependen dan prediktor. Mereka juga mampu menangani banyak interaksi tingkat tinggi dan ukuran kumpulan data yang moderat dengan lebih efektif dibandingkan model regresi [67]. Meskipun model berbasis pohon mempunyai banyak kelebihan, model ini juga mempunyai beberapa kelemahan. Ketika kumpulan data kecil digunakan untuk melatih model berbasis pohon berdasarkan korelasi tinggi antar prediktor, deteksi interaksi antar prediktor akan terhambat, yang dapat menyebabkan overfitting. Namun, efek ini dapat dikurangi dengan menggunakan RF [68].

Algoritme RF dikenal luas dan populer di bidang pembelajaran mesin karena kinerjanya yang sangat baik dan kemampuannya untuk mengatasi tantangan tertentu yang terkait dengan pendekatan berbasis pohon lainnya, seperti pohon keputusan, GBM, Light GBM, dan XGBoost. Salah satu keunggulan utama RF adalah kemampuannya mengurangi overfitting dibandingkan dengan algoritma DT tunggal. Dengan menggabungkan beberapa pohon keputusan dan menggabungkan prediksinya, RF dapat mengurangi varians dan meningkatkan kinerja generalisasi. Sebaliknya, satu DT lebih rentan terhadap overfitting karena dapat mempelajari detail rumit dan noise yang ada dalam data pelatihan.

GBT ringan telah diterapkan dengan ukuran data 585, namun lebih sensitif terhadap overfitting pada kumpulan data kecil [24]. Oleh karena itu, kinerjanya rendah. Ide mendasar di balik AdaBoost adalah menggabungkan beberapa pembelajar yang lemah untuk menciptakan model ansambel yang kuat yang berkinerja lebih baik daripada hanya satu pembelajar yang lemah saja. Namun, jika pembelajar lemah yang digunakan di AdaBoost terlalu kuat atau memiliki kapasitas tinggi dibandingkan dengan terbatasnya jumlah data yang tersedia, maka terdapat risiko overfitting. Pembelajaran [23] menggunakan algoritma Adaboost dengan kumpulan data berukuran kecil (809), dimana pemanfaatan pembelajar lemah yang eksplisit tidak dinyatakan dengan jelas. Ada kemungkinan bahwa pelajar yang lemah dan kompleks telah memanfaatkan kumpulan data berukuran kecil, yang mungkin berkontribusi terhadap rendahnya kinerja. Temuan ini menunjukkan bahwa kehati-hatian harus dilakukan saat menerapkan algoritme seperti Light GBT dan AdaBoost pada kumpulan data kecil, dan pendekatan alternatif atau teknik regularisasi harus dipertimbangkan untuk mengurangi risiko overfitting.

Penelitian yang dilakukan oleh Ref. [15] menggunakan XGBoost bersama dengan pengkodean one-hot. Meskipun pengkodean one-hot dapat memperkenalkan variabel dummy untuk setiap kategori, penting untuk diketahui bahwa beberapa variabel ini mungkin memiliki kekuatan prediksi yang terbatas atau tidak ada sama sekali. Akibatnya, penyertaan variabel-variabel tersebut berpotensi menimbulkan gangguan ke dalam kumpulan data, sehingga meningkatkan kompleksitas proses pembelajaran untuk XGBoost. Selain itu, ketika dihadapkan dengan data yang berisik dan outlier, XGBoost mungkin menunjukkan peningkatan kerentanan. Kehadiran noise dalam kumpulan data dapat meningkatkan performa model yang overfitting atau kurang optimal. Selain itu, pengkodean one-hot dapat mempengaruhi interpretasi pentingnya fitur di XGBoost. Lebih lanjut, nilai kepentingan yang terkait dengan fitur biner yang dikodekan one-hot mungkin tidak secara langsung mencerminkan pentingnya variabel kategori asli. Mengingat hal ini penting ketika menafsirkan pentingnya fitur yang berasal dari pengkodean one-hot.

Dalam data kelangsungan hidup, analisis waktu-ke-peristiwa sering kali dilakukan untuk menentukan kapan peristiwa tertentu terjadi, seperti kematian atau kambuhnya suatu penyakit. Akibatnya, hubungan antara variabel prediktor (misalnya karakteristik pasien, faktor pengobatan) dan hasil kelangsungan hidup mungkin tidak linier. RF adalah alat yang efektif untuk menangkap hubungan nonlinier antara fitur dan variabel target. Dengan RF, setiap pohon keputusan bersifat independen satu sama lain, sehingga memungkinkannya menangkap berbagai pola dalam data berdasarkan subkumpulan fitur yang berbeda. Sebaliknya, sifat sekuensial GBT mungkin mempersulit penangkapan hubungan non-linier yang kompleks saat mencoba memperbaiki kesalahan sebelumnya secara berurutan.

Meskipun model berbasis pohon lainnya seperti Light GBT, XGBoost, AdaBoost, GBT, dan pohon keputusan memiliki kekuatan dan kasus penggunaannya masing-masing, RF telah mendapatkan popularitas karena kinerjanya secara keseluruhan, kesederhanaan, dan kemampuannya untuk menangani berbagai jenis data dan masalah. Namun, penting untuk dicatat bahwa pilihan model bergantung pada masalah spesifik, karakteristik kumpulan data, dan persyaratan penyetalan. Model lain mungkin mengungguli RF dalam skenario tertentu, dan selalu disarankan untuk bereksperimen dan membandingkan model yang berbeda untuk mengidentifikasi model yang paling sesuai untuk tugas tertentu.

Model berbasis jaringan saraf berperforma lebih baik ketika kumpulan datanya besar [69]. Tapi NN multilayer diterapkan di Ref. [13] ke kumpulan data kecil yang terdiri dari 239 sampel untuk membuat model. Namun, mereka seharusnya menggunakan algoritme ML yang lebih akurat dengan kumpulan data kecil, seperti model Bayesian atau RF [48,70,71] untuk meningkatkan kinerja model mereka. Dalam penelitian selanjutnya, peneliti harus menggunakan model yang sesuai dengan ukuran dataset yang digunakan. Selain itu, model berbasis jaringan saraf sensitif terhadap nilai yang hilang. Model Bayesian kuat terhadap variabel-variabel yang tidak relevan (yang mempunyai pengaruh signifikan kecil terhadap data). Selain itu, ini efisien untuk nilai yang hilang dan data berdimensi tinggi tetapi sensitif terhadap data yang berkorelasi [68]. Dalam studi dengan kumpulan data kecil dan tanpa memasukkan atau menangani data yang hilang, salah satu model berbasis Bayesian direkomendasikan.

Kinerja SVM dalam memprediksi kelangsungan hidup antara enam bulan hingga dua tahun masih rendah [11,19]. Meskipun SVM berkinerja baik dengan kumpulan data yang seimbang, kinerjanya buruk jika kumpulan data tidak seimbang.

Hal ini karena hyperplane pemisah SVM bias terhadap kelas mayoritas, yang mungkin berdampak negatif terhadap akurasi klasifikasi minoritas [72]. Studi [11, 19] yang gagal mencapai kinerja yang baik dengan penerapan SVM tanpa menangani ketidakseimbangan kumpulan data. Sejumlah besar alur kerja harus diselesaikan sebelum ML dapat diterapkan untuk memprediksi kelangsungan hidup kanker paru-paru. Misalnya, kumpulan data yang tidak seimbang harus diatasi sebelum mengembangkan algoritma ML. Sebagai [11] gagal mengatasi ketidakseimbangan data, sehingga keakuratan modelnya tidak dapat diterima. Model klasifikasi cenderung bias terhadap kelas besar dan mungkin mengabaikan kelas kecil sama sekali. Sebagian besar model klasifikasi, termasuk NN dan DT, berkinerja lebih baik ketika variabel respons didistribusikan secara merata dalam kumpulan data [73]. Oleh karena itu, disarankan agar penelitian di masa depan menangani kumpulan data yang tidak seimbang menggunakan pendekatan yang sesuai seperti teknik pengambilan sampel minoritas sintetik sebelum menerapkan algoritma ML [74]. Selain itu, kami menemukan bahwa algoritme ML yang dilaporkan dan proses pengembangan di semua penelitian tidak cukup detail untuk direplikasi pada kumpulan data lain atau validasi eksternal. Misalnya, tidak ada penelitian yang mengidentifikasi fitur-fitur data yang memiliki data yang hilang, begitu pula jumlah maupun persentase data yang hilang.

Selain itu, tiga penelitian [11,14,24] menunjukkan model mereka berperforma lebih baik dalam memprediksi waktu kelangsungan hidup yang lebih lama. Ada kemungkinan bahwa hasil tersebut disebabkan oleh kumpulan data yang hanya terdiri dari sejumlah kecil pasien yang bertahan dalam jangka waktu singkat. Oleh karena itu, tidak ada penelitian yang ditinjau yang menyertakan algoritme ML yang secara bersamaan dapat memprediksi kelangsungan hidup kanker paru-paru untuk jangka pendek, menengah, dan panjang (satu, dua, tiga, dan lima tahun). Dalam penelitian selanjutnya, peneliti harus mengembangkan algoritma ML yang mampu memprediksi kelangsungan hidup secara akurat baik untuk jangka pendek maupun jangka panjang.

Rekomendasi kami adalah menemukan karakteristik kumpulan data sebelum mengembangkan model pembelajaran mesin. Untuk mencapai hal ini, mungkin perlu menemukan statistik deskriptif untuk kumpulan data dan memvisualisasikannya untuk mengidentifikasi ukuran, distribusi, dan gangguan data. Identifikasi karakteristik ini akan membantu proses pembersihan data, pengambilan keputusan antara normalisasi dan standardisasi, serta pemilihan model pembelajaran mesin yang sesuai dengan ukuran kumpulan data.

4.5. Evaluasi model dan metrik kinerja

Validasi model sangat penting karena memungkinkan penerapannya pada populasi yang lebih besar daripada populasi sampel. Validasi silang K-fold memungkinkan pelatihan dan pengujian berurutan setiap sampel dalam kumpulan data selama validasi. Oleh karena itu, jumlah data pelatihan meningkat, sehingga menghasilkan bias pesimistis yang lebih rendah. Menggunakan bagian yang berbeda dari dataset pelatihan akan menghasilkan varians, yang merupakan variasi dalam algoritma ML [66]. Namun, varians tersebut dapat diatasi dengan menggunakan model bagging dan RF karena model ini rata-rata melakukan serangkaian observasi [66]. Selain itu, lipatan pengujian tidak tumpang tindih dalam validasi silang k-fold [75]; oleh karena itu, cocok untuk ukuran sampel kecil [76]. Namun, sembilan penelitian [13,19,23, 24,29,30,32,34,35] dengan kumpulan data kecil menggunakan pemisahan pelatihan/pengujian alih-alih validasi silang k-fold. Selain itu, metode pelatihan/pengujian untuk memvalidasi model dapat menghasilkan perkiraan tingkat kesalahan yang menyesatkan jika pemisahan yang dipilih tidak tepat [77]. Tidak ada penelitian yang melaporkan alasan pemilihan metode validasi masing-masing. Di masa depan, metode validasi harus dipilih sesuai dengan ukuran dataset dan metode ML yang digunakan. Disarankan agar peneliti melaporkan alasan pemilihan metode validasi, yang memungkinkan peneliti lain mengikuti proses yang sama jika mereka memiliki data dengan karakteristik serupa. Selain itu, hanya tiga penelitian yang menggunakan validasi eksternal, dan satu penelitian tidak melaporkan metode validasi yang digunakan. Pentingnya validasi eksternal dapat dikaitkan dengan fakta bahwa setiap populasi memiliki ciri khas lokasi yang berbeda, tidak dapat diukur, dan spesifik lokasi yang mungkin berdampak penting pada hasil akhir pasien dan komplikasinya. Oleh karena itu, melatih model tanpa validasi eksternal dapat menghasilkan performa yang berbeda secara signifikan untuk populasi yang berbeda [78].

Secara keseluruhan, penelitian ini berfokus pada penggabungan pembelajaran mesin dalam memprediksi waktu kelangsungan hidup kanker paru-paru. Studi ini mengidentifikasi berbagai hal

tantangan yang perlu diatasi, termasuk data yang hilang, outlier, ketidakseimbangan kelas, dan prediksi waktu kelangsungan hidup jangka panjang. Kumpulan data yang berisi informasi medis seringkali menghadirkan tantangan-tantangan ini sehingga memerlukan penyelesaian yang tepat. Kualitas studi yang dimasukkan dalam tinjauan ini ditemukan sedang. Namun, penelitian lebih lanjut yang dirancang dengan baik diperlukan untuk memvalidasi prediksi yang dibuat oleh pembelajaran mesin. Bidang-bidang utama yang memerlukan perbaikan mencakup teknik pra-pemrosesan, khususnya dalam hal ukuran sampel dan penanganan data yang hilang, yang tidak jelas atau dianggap berisiko tinggi di sebagian besar penelitian.

Beberapa konsekuensi negatif dapat timbul ketika menangani data yang hilang, menggunakan metode transformasi fitur yang tidak tepat, atau salah menangani fitur kategoris dalam konteks memprediksi kelangsungan hidup kanker paru-paru menggunakan pembelajaran mesin. Model ML yang dilatih berdasarkan data yang tidak lengkap atau diubah secara tidak tepat dapat memberikan prediksi yang tidak akurat, sehingga menyebabkan penilaian yang salah terhadap kelangsungan hidup kanker paru-paru. Hal ini dapat mengakibatkan keputusan yang salah bagi pasien dan berpotensi membahayakan pengobatan mereka. Kesalahan penanganan data yang hilang atau fitur kategoris juga dapat menimbulkan bias pada model ML. Selain itu, model ML yang dikembangkan tanpa mengatasi data yang hilang secara memadai atau menggunakan metode transformasi fitur yang sesuai mungkin akan kesulitan untuk melakukan generalisasi dengan baik pada data baru yang belum terlihat.

Dimasukkannya fitur-fitur yang tidak relevan atau non-prediktif ke dalam model karena pemilihan fitur yang tidak tepat dapat menurunkan akurasi prediksi. Fitur-fitur yang tidak terkait ini tidak hanya meningkatkan beban komputasi tetapi juga menimbulkan gangguan, sehingga mempersulit algoritme ML untuk mengidentifikasi pola bermakna terkait kelangsungan hidup kanker paru-paru. Pemilihan algoritme ML sangatlah penting, karena memilih algoritme yang tidak tepat untuk kumpulan data tertentu dapat menyebabkan performa kurang optimal dan prediksi tidak akurat. Setiap algoritma memiliki kekuatan dan kelemahannya masing-masing, dan memahami karakteristik ini sangat penting untuk memastikan pemilihan model yang optimal.

Selain itu, permasalahan ketidakseimbangan data, dimana satu kelas (misalnya kelangsungan hidup jangka pendek) kurang terwakili secara signifikan dibandingkan kelas lainnya, perlu diatasi. Kegagalan dalam memperhitungkan ketidakseimbangan ini dapat membuat model ML menjadi bias terhadap kelas mayoritas, sehingga menghasilkan akurasi prediksi yang buruk untuk kelas minoritas, yang mungkin penting untuk mengidentifikasi individu berisiko tinggi atau mereka yang memerlukan intervensi spesifik. Terakhir, kelalaian dalam memahami karakteristik kumpulan data, seperti ukuran, distribusi, dan noise, dapat menghambat pemrosesan awal data dan pengembangan model yang efektif. Pembersihan, normalisasi, atau standarisasi data yang tidak memadai dapat menimbulkan bias atau mendistorsi distribusi data, yang pada akhirnya menghasilkan prediksi yang tidak dapat diandalkan.

5. Keterbatasan

Tinjauan sistematis ini memiliki beberapa keterbatasan yang harus diperhatikan. Ulasan ini bertujuan untuk mengidentifikasi semua penelitian yang dilakukan untuk memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis menggunakan ML. Selain itu, penelitian yang termasuk dalam tinjauan ini memiliki perbedaan yang signifikan, sehingga membatasi kemampuan kami untuk membandingkannya. Lebih lanjut, karena beberapa penelitian gagal memberikan analisis kuantitatif kinerja model, seperti spesifisitas, sensitivitas, tingkat prediksi negatif, atau tingkat prediksi positif, tinjauan ini

Lampiran.

Lampiran A

Tabel A1

Model terbaik di setiap penelitian berdasarkan akurasi RMSE

Model	Ketepatan	Periode waktu kelangsungan hidup yang diprediksi	Ukuran publikasi	Referensi
Federasi Rusia	10.52	6 bulan	10.442	[12]
Ansambel Kustom	15.30	5 tahun	10.442	[16]
Peta Pemesanan Sendiri	15.59	5 tahun	10.442	[17]
GRNN	0,60	5 tahun	683	[20]

(bersambung di halaman berikutnya)

tidak mencakup analisis kuantitatif. Terlepas dari keterbatasan ini, sebagian besar penelitian yang ditinjau menggunakan kriteria serupa untuk memprediksi kelangsungan hidup kanker paru-paru, seperti data klinis dan algoritma ML, yang berkontribusi pada kerangka kerja yang kuat untuk membangun algoritma ML untuk memprediksi kelangsungan hidup kanker paru-paru.

6. Kesimpulan

Untuk memilih pengobatan yang paling sesuai untuk pasien tertentu dengan kanker paru-paru, penyedia layanan kesehatan harus dapat memprediksi waktu kelangsungan hidup mereka. Model ML dapat digunakan untuk memprediksi kelangsungan hidup, seperti yang saat ini digunakan di banyak bidang lain untuk memprediksi hasil. Berdasarkan temuan tinjauan ini, dapat disimpulkan bahwa minat untuk menggunakan algoritma ML untuk memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis telah meningkat sejak tahun 2012. Analisis terhadap berbagai sumber data menunjukkan bahwa database SEER adalah yang paling banyak digunakan. Meskipun kelompok pasien target bervariasi antar penelitian, analisis rinci terhadap kelompok pasien lain, seperti pasien dengan kanker paru-paru sel kecil, diperlukan. Hasil analisis metode pra-pemrosesan menunjukkan bahwa meskipun banyak penelitian menggunakan metode pemilihan fitur untuk mengidentifikasi fitur-fitur penting, sebagian besar penelitian yang ditinjau tidak memperhitungkan data yang hilang, normalisasi, penskalaan, dan data standar. Oleh karena itu, ada kemungkinan bahwa algoritma ML yang digunakan dalam penelitian ini mungkin bias. Oleh karena itu, masih diperlukan studi fitur untuk memprediksi kelangsungan hidup kanker paru-paru berdasarkan data klinis dengan lebih memperhatikan metode preprocessing dan memberikan lebih detail mengenai metode tersebut.

Kontribusi penulis

Altuhaifa, Dr. Win, dan Dr. Su mengonsep dan merancang penelitian ini. Altuhaifa, Dr. Win, dan Dr. Su mengumpulkan artikel-artikel yang telah direview dan menyaringnya, dan menyusun naskahnya. Fatimah Altuhaifa melakukan analisis terhadap artikel-artikel yang diulas.

Pendanaan

Tak dapat diterapkan.

Deklarasi kepentingan bersaing

Para penulis menyatakan tidak ada konflik kepentingan.

Ucapan Terima Kasih

Kami berterima kasih atas beasiswa yang diberikan Kementerian Pendidikan Tinggi Arab Saudi kepada Fatimah Altuhaifa. Kami juga berterima kasih kepada Universitas Wollongong Australia yang telah menyediakan akses ke sebagian besar perpustakaan digital dan mesin pencari ilmiah yang digunakan dalam penelitian ini. Kami sangat berterima kasih kepada Rachel Lawson, Pustakawan Ilmu Pengetahuan, Kedokteran, dan Layanan Penghubung Kesehatan di Perpustakaan Gedung 16, atas bantuannya yang sangat berharga dalam menemukan artikel yang ditinjau. Keahlian dan dedikasinya sangat penting bagi penelitian kami.

Tabel A1(lanjutan)

Model	Ketepatan	Periode waktu kelangsungan hidup yang diprediksi	Ukuran publikasi	Referensi
Regresi Punggung Bukit	2.70	3 tahun	291	[27]
ANN untuk pria	2.32	5 tahun	38.262	[31]
DNN	0,314	5 tahun	10.001	[36]

Tabel A2

Model terbaik di setiap penelitian berdasarkan akurasi AUC

Model	Ketepatan	Periode kelangsungan hidup yang diprediksi	Ukuran publikasi	Referensi
Pohon IKLAN	93.8	5 tahun	57.254	[11]
MNN	78.30	2 tahun	239	[13]
JST	92.0	5 tahun	Tidak dilaporkan	[14]
CNN	92.0	5 tahun	Tidak dilaporkan	[14]
Federasi Rusia	90.1	3 tahun	50.687	[10]
XGBoost	78.6	1 tahun	5973	[15]
Federasi Rusia	80.2	5 tahun	5123	[18]
IRF	98.0	2 tahun	509	[19]
NB basis-K Gaussian	88.1	5 tahun	321	[21]
AdaBoost	71.3	2 tahun	809	[23]
Regresi logistik	76.0	5 tahun	585	[24]
DL	74.4	5 tahun	16.613	[25]
SVM	71.0	3 tahun	371	[26]
catatan	81.0	3 tahun	291	[27]
SORG ML	71.4	1 tahun	150	[28]
LASO	75.3	1 tahun	1563	[29]
DL	81.3	5 tahun	704	[30]
Federasi Rusia	95.1	5 tahun	17.484	[33]
Federasi Rusia	68.0	5 tahun	739	[35]
Regresi Cox	72.2	5 tahun	2166	[38]
XGBoost	Perempuan 93.0	1 tahun	28.458	[39]

Tabel A3

Model terbaik berdasarkan akurasi C-statistik

Model	Ketepatan	Periode waktu kelangsungan hidup yang diprediksi	Ukuran publikasi	Referensi
DL	73.9	5 tahun	17.322	[22]
DL	63.6	5 tahun	1137	[32]
Federasi Rusia	67.2	5 tahun	506	[34]
DL	83.4	5 tahun	4617	[37]

Lampiran B

Tabel B1

Model pada setiap penelitian berdasarkan akurasi RMSE untuk survival time 6 bulan

Algoritma ML	Akurasi untuk setiap referensi
Federasi Rusia	10.52 [12]
LR	10.63 [12]
GBT	10.65 [12]
Ansambel Kustom	10.84 [12]

Tabel B2

Model pada setiap penelitian berdasarkan akurasi RMSE untuk survival time 2 tahun

Algoritma ML	Akurasi untuk setiap referensi
Federasi Rusia	15.70 [12]
LR	15,77 [12]
GBT	15,65 [12]
Ansambel Kustom	16.26 [12]

Tabel B3

Model pada setiap penelitian berdasarkan akurasi RMSE untuk survival time 3 tahun

Algoritma ML	Akurasi untuk setiap referensi
LR	3.01 [27]
Regresi Punggung Bukit	2,70 [27]
Jalur SVR	2,78 [27]
Poli SVM	2.81 [27]

Tabel B4

Model pada setiap penelitian berdasarkan akurasi RMSE untuk survival time 5 tahun

Algoritma ML	Akurasi untuk setiap referensi
Federasi Rusia	20.51 [12], 15,63 [16]
LR	21.37 [12], 15.38 [16]
GBT	21.14 [12], 15.32 [16]
Ansambel Kustom	21.18 [12], 15.30 [16]
SVM	15.82 [16]
DT	15.81 [16]
HC	16.202 [17]
MBC	16.250 [17]
Pengelompokan K-Means	16.193 [17]
SOM	15.591 [17]
Faktorisasi Matriks Non-negatif	16.589 [17]
GRNN	0,60 [20]
JST	Laki-laki 2.32 [31], perempuan 2,52 [31]
Bayes Bersih	0,315 [36]
DNN	0,314 [36]
Regresi logistik	0,314 [36]
Pengklasifikasi Malas LWL	0,318 [36]
J48 DT	0,32 [36]
Pengklasifikasi Meta (ASC)	0,316 [36]
Pembelajar Aturan (OneR)	0,339 [36]

Tabel B5

Model pada setiap penelitian berdasarkan akurasi AUC untuk survival time 6 bulan

Algoritma ML	Akurasi untuk setiap referensi
SVM	60,5 [11], 79,0 [14]
JST	74,0 [11], 82,0 [14]
J48 DT	78.1 [11]
Federasi Rusia	78.1 [11], 80,0 [14]
Peningkatan Logit	80,5 [11]
Subruang Acak	80,8 [11]
Pohon IKLAN	80,8 [11]
CNN	83,0 [14]
RNN	81,0 [14]
catatan	77,0 [14]

Tabel B6

Model pada setiap penelitian berdasarkan akurasi AUC untuk survival time 1 tahun

Algoritma ML	Akurasi untuk setiap referensi
SVM	63,4 [11], 73,0 [15], laki-laki 86,4 [39], 91,6 [39]
JST	80,2 [11]
J48 DT	78,6 [11]
Federasi Rusia	81.1 [11], 85,2 [10], 73,6 [15], laki-laki 86,0 [39], perempuan 91,2 [39]
MTLR	82.1 [10]
Peningkatan Logit	82.1 [11]
Subruang Acak	82,4 [11]
Pohon IKLAN	83,0 [11]
Algoritma Logistik	71,0 [15], laki-laki 85,5 [39], perempuan 92,1 [39]
DT	laki 84,3 [39], perempuan 90,4 [39]
XGBoost	78,6 [15], laki-laki 87,8 [39], perempuan 93,0 [39]
SORGA	71,4 [28]
LASO	75,3 [29]
KNN	Laki-laki 83,6 [39], perempuan 89,8 [39]

Tabel B7

Model pada setiap penelitian berdasarkan akurasi AUC untuk survival time 2 tahun

Algoritma ML	Akurasi untuk setiap referensi
SVM	57,0 [11], 64,0 [14], 51,6 [19]
JST	81,6 [11], 86,0 [14]
J48 DT	81,8 [11]
<small>Federasi Rusia</small>	86,7 [11], 66,0 [14], 96,8 [19], 55,0 [24]
<small>Peningkatan Logit</small>	87,7 [11]
Subruang Acak	87,8 [11]
<small>Pohon IKLAN</small>	88,4 [11]
CNN	86,0 [14]
RNN	86,0 [14]
<small>catatan</small>	63,0 [14], 54,0 [19]
Regresi logistik	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]
GBT ringan	54,0 [24]
mendukung pengelompokan vektor SVC	55,0 [24]

Tabel B8

Model pada setiap penelitian berdasarkan akurasi AUC untuk survival time 3 tahun

Algoritma ML	Akurasi untuk setiap referensi
SVM	71,0 [26], 73,0 [27], laki-laki 70,6 [39], perempuan 81,1 [39]
<small>Federasi Rusia</small>	90,1 [10], 76,0 [27], laki-laki 70,2 [39], perempuan 80,7 [39]
<small>catatan</small>	81,0 [27]
MTLR	86,4 [10]
LR	74,0 [27]
LASO	73,6 [29]
Regresi logistik	Laki-laki 69,9 [39], perempuan 81,8 [39]
DT	Pria [39], perempuan 80,1 [39] Laki-laki
XGBoost	72,4 [39], perempuan 82,9 [39] Laki-laki
KNN	67,2 [39], perempuan 79,5 [39]

Tabel B9

Model pada setiap penelitian berdasarkan akurasi AUC untuk survival time 5 tahun

Algoritma ML	Akurasi untuk setiap referensi
SVM	56,4 [11], 84,0 [14], laki-laki 72,0 [39], perempuan 82,3 [39]
JST	92,3 [11], 92,0 [14]
J48 DT	84,7 [11], 94,4 [33]
<small>Federasi Rusia</small>	90,5 [11], 86,0 [14], 89,9 [10], 69,2 [23], 68,8 [25], 80,259 [18], 95,1 [33], 68,0 [35], laki-laki 71,6 [39], perempuan 82,0 [39]
<small>Peningkatan Logit</small>	93,0 [11]
Subruang Acak	93,2 [11], 56,0 [24]
MTLR	87,0 [10]
<small>Pohon IKLAN</small>	93,7 [11]
CNN	92,0 [14]
RNN	91,0 [14]
<small>catatan</small>	83,0 [14], 59,7 [21]
Regresi logistik	63,2 [23], 76,0 [24], 92,7 [33], laki-laki 71,0 [39], perempuan 83,0 [39]
SPNN	71,0 [24]
DT	69,2 [23], 78,166 [18], laki-laki 69,4 [39], perempuan 81,4 [39]
MLP	71,0 [24]
XGBoost	54,0 [24], laki-laki 73,5 [39], perempuan 84,2 [39]
Gaussian NB	73,0 [24]
GBT ringan	67,0 [24]
MTLR	87,0 [10]
LR	62,2 [21]
Basis Gaussian K NB	88,1 [21]
Bagging	70,6 [23]
AdaBoost	71,3 [23]
mendukung pengelompokan vektor	67,0 [24]
SVC DL	74,4 [25], 81,6 [30]
LASO	65,6 [29]
Bayes Bersih	94,2 [33]
Regresi Cox	72,2 [38]
KNN	Laki-laki 68,6 [39], perempuan 80,7 [39]

Tabel B10
Model berdasarkan akurasi statistik-C untuk waktu kelangsungan hidup 5 tahun

model ml	Akurasi untuk setiap referensi
DL	73,9 [22], 63,6 [32], 83,4 [37]
Regresi Cox	56.1 [32], kelompok GSE72094 63,0 [34], kelompok TCGA 64,5 [34], 64,0 [37]
Cox NN	56.2 [32]
Federasi Rusia	GSE72094 kelompok 67,2 [34], kelompok TCGA 65,6 [34], 67,8 [37]

Lampiran C

Tabel C1
Validasi

	Validasi silang K-fold	Pemisahan data	Tidak dilaporkan	Validasi eksternal
[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]	✓			

Lampiran D

Tabel D1
penilaian PROBAST

Referensi	Peserta studi	Prediktor	Hasil	Ukuran sampel dan data yang hilang	Analisis statistik
[11]	Rendah	Tidak jelas	Rendah	Tidak jelas	Tidak jelas
[12]	Rendah	Rendah	Rendah	Tidak jelas	Rendah
[13]	Rendah	Rendah	Rendah	Tidak jelas	Rendah
[14]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[10]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[15]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[16]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[17]	Rendah	Tidak jelas	Rendah	Tidak jelas	Tidak jelas
[18]	Rendah	Rendah	Rendah	Rendah	Tidak jelas
[19]	Rendah	Rendah	Rendah	Tidak jelas	Rendah
[20]	Rendah	Tidak jelas	Rendah	Berisiko tinggi	Tidak jelas
[21]	Rendah	Tidak jelas	Rendah	Berisiko tinggi	Rendah
[22]	Rendah	Tidak jelas	Rendah	Rendah	Tidak jelas
[23]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[24]	Rendah	Rendah	Rendah	Berisiko tinggi	Rendah
[25]	Rendah	Rendah	Rendah	Rendah	Rendah
[26]	Rendah	Tidak jelas	Rendah	Berisiko tinggi	Tidak jelas
[27]	Rendah	Tidak jelas	Rendah	Tidak jelas	Rendah
[28]	Rendah	Tidak jelas	Rendah	Tidak jelas	Tidak jelas
[29]	Rendah	Tidak jelas	Rendah	Tidak jelas	Tidak jelas
[30]	Rendah	Tidak jelas	Rendah	Berisiko tinggi	Tidak jelas

(bersambung di halaman berikutnya)

Tabel D1(lanjutan)

Referensi	Peserta studi	Prediktor	Hasil	Ukuran sampel dan data yang hilang	Analisis statistik
[31]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[32]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[33]	Rendah	Rendah	Rendah	Rendah	Tidak jelas
[34]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[35]	Rendah	Rendah	Rendah	Berisiko tinggi	Tidak jelas
[36]	Rendah	Rendah	Rendah	Tidak jelas	Rendah
[37]	Rendah	Rendah	Rendah	Rendah	Tidak jelas
[38]	Rendah	Rendah	Rendah	Tidak jelas	Tidak jelas
[39]	Rendah	Rendah	Rendah	Rendah	Tidak jelas

Referensi

- [1] WHO, "Kanker [Online]. Tersedia: <https://www.who.int/news-room/fact-sheets/detail/cancer>, 2022.
- [2] S. Tomassini, N. Falconelli, P. Sernani, L. Burattini, AF Dragoni, Diagnosis nodul paru-paru dan klasifikasi histologi kanker dari data tomografi komputer oleh jaringan saraf konvolusional: survei, Comput. biologi. medis. 146 (Juli 2022), 105691, <https://doi.org/10.1016/j.combiomed.2022.105691>.
- [3] LA Vale-Silva, K.Rohr, Prediksi kelangsungan hidup kanker jangka panjang menggunakan pembelajaran mendalam multimodal, dalam bahasa Inggris, Sci Rep-Uk 11 (1) (29 Juni 2021). ARTN 1350510.1038/ s41598-021-92799-4.
- [4] Y. Yang, L. Xu, L. Sun, P. Zhang, SS Farid, Aplikasi pembelajaran mesin dalam prediksi kekambuhan dan kelangsungan hidup kanker paru-paru yang dipersonalisasi, Comput. Struktur. Bioteknologi. J.20 (2022) 1811–1820, <https://doi.org/10.1016/j.csbj.2022.03.035>.
- [5] EM Nwanosike, BR Conway, HA Merchant, SS Hasan, Potensi aplikasi dan kinerja teknik dan algoritma pembelajaran mesin dalam praktik klinis: tinjauan sistematis, Int. J.Med. Inf. 159 (Maret 2022), 104679, <https://doi.org/10.1016/j.jmedinf.2021.104679>.
- [6] I. Kaur, MN Doja, T. Ahmad, Penambahan data dan pembelajaran mesin dalam penelitian kelangsungan hidup kanker: gambaran umum dan rekomendasi masa depan, J. Biomed. Inf. 128 (April 2022), 104026, <https://doi.org/10.1016/j.jbi.2022.104026>.
- [7] A. Liberati, dkk., Pernyataan PRISMA untuk melaporkan tinjauan sistematis dan meta-analisis penelitian yang mengevaluasi intervensi layanan kesehatan: penjelasan dan elaborasi, BMJ 339 (21 Juli 2009) b2700, <https://doi.org/10.1136/bmj.b2700>.
- [8] Perangkat lunak tinjauan sistematis Covidence." Veritas Health Innovation. <https://www.covidence.org/>(diakses).
- [9] D. Bzdok, N. Altman, M. Krzywinski, Statistik versus pembelajaran mesin, Nat. Metode 15 (4) (April 2018) 233–234, <https://doi.org/10.1038/nmeth.4642>.
- [10] T. He, J. Li, P. Wang, Z. Zhang, Sistem prediksi kecerdasan buatan tingkat kelangsungan hidup individu untuk adenokarsinoma paru, Comput. Struktur. Bioteknologi. 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, Prediksi kelangsungan hidup kanker paru-paru menggunakan data ensemble penambahan pada data SIER, Sci. Program. 20 (1) (2012) 29–42, <https://doi.org/10.1155/2012/920245>, dalam bahasa Inggris.
- [12] JA Bartholomaj, HB Frieboes, Prediksi kelangsungan hidup kanker paru-paru melalui regresi pembelajaran mesin, klasifikasi, dan teknik statistik, 2018, dalam: Proc IEEE Int Symp Signal Proc Inf Tech, Des 2018, hlm. 632–637, <https://doi.org/10.1109/ISSPIT.2018.8642753>.
- [13] Y. Dagli, S. Choksi, S. Roy, Prediksi kelangsungan hidup dua tahun di antara pasien kanker paru-paru bukan sel kecil, LN Comput. Vis. Bioma. 31 (2019) 169–177, https://doi.org/10.1007/978-3-030-04061-1_17, dalam bahasa Inggris.
- [14] S. Doppalapudi, RG Qiu, Y. Badr, Prediksi dan pemahaman periode kelangsungan hidup kanker paru-paru: pendekatan pembelajaran mendalam, Int. J.Med. Inf. 148 (April 2021), 104371, <https://doi.org/10.1016/j.jmedinf.2020.104371>.
- [15] Z. Huang, C. Hu, C. Chi, Z. Jiang, Y. Tong, C. Zhao, Model kecerdasan buatan untuk memprediksi kelangsungan hidup 1 tahun metastasis tulang pada pasien kanker paru-paru non-sel kecil berdasarkan pada algoritma XGBoost, 2020, BioMed Res. Int. (2020), 3462363, <https://doi.org/10.1155/2020/3462363>.
- [16] CM Lynch, dkk., Prediksi kelangsungan hidup pasien kanker paru-paru melalui teknik klasifikasi pembelajaran mesin yang diawasi, Int. J.Med. Inf. 108 (Des 2017) 1–8, <https://doi.org/10.1016/j.jmedinf.2017.09.013>.
- [17] CM Lynch, VH van Berkel, HB Frieboes, Penerapan teknik analisis tanpa pengawasan pada data pasien kanker paru-paru, PLoS One 12 (9) (2017), e0184370, <https://doi.org/10.1371/journal.pone.0184370>.
- [18] XY Mei, Memprediksi kelangsungan hidup lima tahun secara keseluruhan pada pasien dengan kanker paru-paru non-sel kecil dengan algoritma Relief dan hutan acak, dalam: 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (Iaeac), 2017, hal. 2527–2530, <https://doi.org/10.1109/IAEAC.2017.8054479>, dalam bahasa Inggris.
- [19] P. Nanda, N. Duraipandian, Prediksi tingkat kelangsungan hidup dari kanker paru-paru non-sel kecil menggunakan hutan acak yang ditingkatkan, dalam: Prosiding Konferensi Internasional ke-5 tentang Teknologi Komputasi Inventif (Icict-2020), 2020, hal. 93–97, <https://doi.org/10.1109/IcICT48043.2020.9112558>, dalam bahasa Inggris.
- [20] K. Qaddoum, Prediksi Kelangsungan Hidup Pasien Kanker Paru Menggunakan GRNN-CP, 2020.
- [21] K. R, GR R, Akurasi meningkatkan prognosis kanker paru-paru untuk meningkatkan kemampuan bertahan hidup pasien menggunakan sistem pengklasifikasi Gaussian yang diuluskan, J. Med. sistem. 43 (7) (24 Mei 2019), <https://doi.org/10.1007/s10916-019-1297-2>, 201.
- [22] Y. She, dkk., Pengembangan dan validasi model pembelajaran mendalam untuk kelangsungan hidup kanker paru-paru non-sel kecil, JAMA Netw. Buka 3 (6) (1 Juni 2020), e205842, <https://doi.org/10.1001/jamanetworkopen.2020.5842>.
- [23] JA Sim, YH Yun, Memprediksi kelangsungan hidup kanker paru-paru bebas penyakit menggunakan pengukuran hasil yang dilaporkan pasien (PRO) dengan perbandingan lima teknik pembelajaran mesin (MLT), Stud. Teknologi Kesehatan. Inf. 264 (21 Agustus 2019) 1588–1589, <https://doi.org/10.3233/SHT190548>.
- [24] M. Yakar, D. Etiz, S. Yilmaz, O. Celik, AK Guntulu, M. Metintas, Prediksi kelangsungan hidup dan kelangsungan hidup bebas perkembangan menggunakan pembelajaran mesin pada kanker paru-paru stadium III: studi percontohan, Turk. Onkol. Derg. 36 (4) (2021) 446–458, <https://doi.org/10.5505/tjo.2021.2788>, dalam bahasa Inggris.
- [25] L. Yang, dkk., Sistem prognostik pembelajaran mendalam yang baru meningkatkan prediksi kelangsungan hidup untuk kanker paru-paru non-sel kecil stadium III, Cancer Med. (2 Mei 2022), <https://doi.org/10.1002/cam4.4782>.
- [26] J. Yu, dkk., LUADpp: model prediksi efektif pada prognosis adenokarsinoma paru berdasarkan fitur mutasi somatik, Kanker BMC 19 (1) (22 Maret 2019) 263, <https://doi.org/10.1186/s12885-019-5433-7>.
- [27] Y. Liu, dkk., Mengembangkan panel gen prognostik waktu bertahan hidup pada pasien adenokarsinoma paru menggunakan pembelajaran mesin, Transl. Res Kanker. 9 (6) (Jun 2020) 3860–3869, <https://doi.org/10.21037/tcr-19-2739>.
- [28] G. Zhong, dkk., Validasi eksternal dari algoritme pembelajaran mesin SORG untuk memprediksi kelangsungan hidup 90 hari dan 1 tahun pasien dengan metastasis tulang belakang akibat kanker paru-paru: kohort bi-center baru-baru ini dari Tiongkok, Spine J. 23 (5) (Mei 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, Sebuah studi kohort besar yang mengidentifikasi model prediksi prognosis baru untuk adenokarsinoma paru melalui strategi pembelajaran mesin, BMC Cancer 19 (1) (5 September 2019) 886, <https://doi.org/10.1186/s12885-019-6101-7>.
- [30] YH Lai, WN Chen, TC Hsu, C. Lin, Y. Tsao, S. Wu, Prediksi kelangsungan hidup kanker paru-paru non-sel kecil secara keseluruhan dengan mengintegrasikan microarray dan data klinis dengan pembelajaran mendalam, Sci. Rep.10 (1) (13 Maret 2020) 4679, <https://doi.org/10.1038/s41598-020-61588-w>.
- [31] HRCPTsokos, Model jaringan saraf tiruan untuk memprediksi kelangsungan hidup kanker paru-paru, J. Data Anal. Inf. Proses. (2017), <https://doi.org/10.4236/jdaip.2017.51003>.
- [32] J. Wang, N. Chen, J. Guo, X. Xu, L. Liu, Z. Yi, SurvNet: jaringan saraf dalam baru untuk analisis kelangsungan hidup kanker paru-paru dengan nilai yang hilang, Front. Onkol. 10 (2020), 588990, <https://doi.org/10.3389/fonc.2020.588990>.
- [33] A. Safiyari, R. Javidan, Memprediksi kelangsungan hidup kanker paru-paru menggunakan metode pembelajaran ansambel, dalam bahasa Inggris, dalam: Proceedings of the 2017 Intelligent Systems Conference (Intellisys), 2017, hlm. <Pergi ke ISI-/WOS: 000456827800090.
- [34] B. Ma, Y. Geng, F. Meng, G. Yan, F. Song, Identifikasi biomarker prognostik enam belas gen untuk adenokarsinoma paru menggunakan metode pembelajaran mesin, 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, Identifikasi penanda tujuh gen untuk memprediksi kelangsungan hidup pasien adenokarsinoma paru menggunakan analisis data multi-omics terintegrasi, J. Clin. Laboratorium. Dubur. 36 (2) (Februari 2022), e24190, <https://doi.org/10.1002/jcla.24190>.
- [36] S. Huang, I. Arpacı, M. Al-Emran, S. Kılıçarslan, MA Al-Sharafi, Analisis komparatif pembelajaran mesin klasik dan teknik pembelajaran mendalam untuk memprediksi kelangsungan hidup kanker paru-paru, Multimed. Alat. Aplikasi. (2023), <https://doi.org/10.1007/s11042-023-16349-y>.
- [37] L. Jin, Q. Zhao, S. Fu, F. Cao, B. Hou, J. Ma, Pengembangan dan validasi model pembelajaran mesin untuk memprediksi kelangsungan hidup pasien dengan NSCLC stadium-III yang direseksi, Front. Onkol. 13 (2023), 1092478, <https://doi.org/10.3389/fonc.2023.1092478>.
- [38] H. Ma, L. Tong, Q. Zhang, W. Chang, F. Li, Identifikasi 5 tanda tangan gen dalam prediksi kelangsungan hidup pasien karsinoma sel skuamosa paru berdasarkan analisis data multiomik terintegrasi, 2020, BioMed Res. Int. (2020), 6427483, <https://doi.org/10.1155/2020/6427483>.
- [39] Y. Wang, dkk., Investigasi berbasis pembelajaran mesin tentang prognosis kanker paru-paru spesifik gender, Medicina (Kaunas) 57 (2) (22 Jan 2021), <https://doi.org/10.3390/medicina57020099>.
- [40] A. Moskowitz, K. Chen, Mendefinisikan kelompok pasien, dalam: Analisis Sekunder Catatan Kesehatan Elektronik, CH), Cham, 2016, hlm. 93–100, bab. 93–100.

- [41] DSK Karunasingha, Root mean square error atau mean error absolut? Gunakan juga rasionya, *Inf. Sains*. 585 (2022) 609–629, <https://doi.org/10.1016/j.pada.tahun.2021.11.036>.
- [42] F. Melo, Area di bawah kurva ROC, dalam: *Encyclopedia of Systems Biology*, 2013. Springer Nature.
- [43] D. Westreich, SR Cole, MJ Funk, MA Brookhart, T. Sturmer, Peran statistik dalam pemilihan variabel untuk model skor kecenderungan, *Pharmacoepidemiol. Obat Aman*. 20 (3) (Maret 2011) 317–320, <https://doi.org/10.1002/pds.2074>.
- [44] RF Wolff, dkk., PROBAST: alat untuk menilai risiko bias dan penerapan studi model prediksi, *Ann. Magang. medis*. 170 (1) (1 Januari 2019) 51–58, <https://doi.org/10.7326/M18-1376>.
- [45] Y. Kumar, S. Gupta, R. Singla, YC Hu, Tinjauan sistematis teknik kecerdasan buatan dalam prediksi dan diagnosis kanker, *Arch. Hitung. Metode Eng*. 29 (4) (2022) 2043–2070, <https://doi.org/10.1007/s11831-021-09648-w>.
- [46] EW Steyerberg, MJ Eijkemans, FE Harrell Jr., JD Habbema, Pemodelan prognostik dengan analisis regresi logistik: perbandingan metode seleksi dan estimasi dalam kumpulan data kecil, *Stat. medis*. 19 (8) (30 April 2000) 1059–1079, [https://doi.org/10.1002/\(sici\)1097-0258\(20000430\)19:8<1059::bantuansim412>3.0.co;2-o](https://doi.org/10.1002/(sici)1097-0258(20000430)19:8<1059::bantuansim412>3.0.co;2-o).
- [47] T. van der Ploeg, PC Austin, EW Steyerberg, Teknik pemodelan modern haus data: studi simulasi untuk memprediksi titik akhir dikotomis, *BMC Med. Res. Metodologi*. 14 (22 Desember 2014) 137, <https://doi.org/10.1186/1471-2288-14-137>.
- [48] V. Khadse, PN Mahalle, SV Biraris, Perbandingan empiris algoritme pembelajaran mesin yang diawasi untuk data internet of things, dalam bahasa Inggris, di: *Konferensi Internasional Keempat tentang Kontrol dan Otomasi Komunikasi Komputasi (Iccubea)* tahun 2018, 2018. <Pergi ke ISI>/WOS:000493801500117.
- [49] DM Rice, Penalaran kausal, dalam: *Kalkulus Pemikiran*, 2014, hlm.95–123.
- [50] Pengembang Scikit-learn (Lisensi BSD), 6.3. Memproses data sebelumnya, <https://scikit-belajar.org/stable/modules/preprocessing.html>.
- [51] Masyarakat Kanker Amerika. "Statistik Utama untuk Kanker Paru-paru." Masyarakat Kanker Amerika. <https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html> (diakses).
- [52] K. Maharana, S. Mondal, B. Nemade, Tinjauan: teknik pra-pemrosesan data dan augmentasi data, *Global Transit. Proses*. 3 (1) (2022) 91–99, <https://doi.org/10.1016/j.gltip.2022.04.020>.
- [53] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, Survei tentang data yang hilang dalam pembelajaran mesin, *J. Big Data* 8 (1) (2021) 140, <https://doi.org/10.1186/s40537-021-00516-9>.
- [54] AR Donders, GJ van der Heijden, T. Stijnen, KG Moons, Ulasan: pengenalan lembut terhadap imputasi nilai-nilai yang hilang, *J. Clin. Epidemiol.* 59 (10) (Oktober 2006) 1087–1091, <https://doi.org/10.1016/j.jclinepi.2006.01.014>.
- [55] DB Rubin, Imputasi ganda setelah 18+ tahun, *J. Am. Statistik. Asosiasi*. 91 (434) (Juni 1996) 473–489, <https://doi.org/10.1080/01621459.1996.10476908>, dalam bahasa Inggris.
- [56] T. Kose, S. Ozgur, E. Cosgun, A. Keskinoglu, P. Keskinoglu, Pengaruh imputasi data yang hilang pada kinerja prediksi pembelajaran mendalam untuk refluks vesikoureteral dan studi klinis infeksi saluran kemih berulang, 2020, *BioMed Res. Int.* (2020), 1895076, <https://doi.org/10.1155/2020/1895076>.
- [57] N. S, EA C, MC M, HD Lee, ReliefF untuk Pemilihan Fitur Multi-Label, 2013, <https://doi.org/10.1109/BRACIS.2013.10> [On line]. Tersedia: <https://ieeexplore.ieee.org/document/6726418>.
- [58] NG Gyori, M. Palombo, CA Clark, H. Zhang, DC Alexander, Distribusi data pelatihan berdampak signifikan pada estimasi struktur mikro jaringan dengan pembelajaran mesin, *Magn. Alasan. medis*. 87 (2) (Februari 2022) 932–947, <https://doi.org/10.1002/mrm.29014>.
- [59] MS Chen, JW Han, PS Yu, Data mining: gambaran umum dari perspektif database, dalam bahasa Inggris, *IEEE Trans. Tahu. Data Bahasa Inggris*. 8 (6) (Des 1996) 866–883, <https://doi.org/10.1109/69.553155>.
- [60] MM Siraj, NA Rahmat, MM Din, Survei tentang pendekatan dan teknik penambahan data yang menjaga privasi, dalam bahasa Inggris, dalam: 2019 8th International Conference on Software and Computer Applications (Icsa 2019), 2019, hlm. 65–69, <https://doi.org/10.1145/3316615.3316632>.
- [61] M. Gal, DL Rubinfeld, Standardisasi data, *SSRN Electron. J.* (2018), <https://doi.org/10.2139/ssrn.3326377>.
- [62] C. Seger, Investigasi teknik pengkodean variabel kategori dalam pembelajaran mesin: biner versus one-hot dan hashing fitur, SWEDEN, di: KTH ROYAL INSTITUTE OF TECHNOLOGY, 2018 [Online]. Tersedia: <https://www.diva-port.al.org/smash/get/diva2:1259073/FULLTEXT01.pdf>.
- [63] PA Noseworthy, dkk., Menilai dan mengurangi bias dalam kecerdasan buatan medis: pengaruh ras dan etnis pada model pembelajaran mendalam untuk analisis EKG, *Circ. Elektrofisiol Aritmia*. 13 (3) (Maret 2020), e007988, <https://doi.org/10.1161/CIRCEP.119.007988>.
- [64] E. Scornet, G. Biau, J.-P. Vert, Konsistensi hutan acak, *Ann. Statistik*. 43 (4) (2015), <https://doi.org/10.1214/15-aos1321>.
- [65] Y. Ao, H. Li, L. Zhu, S. Ali, Z. Yang, Algoritme hutan acak linier dan kelebihanannya dalam pemodelan regresi logging berbantuan pembelajaran mesin, *J. Petrol. Sains. bahasa Inggris* 174 (2019) 776–789, <https://doi.org/10.1016/j.petrol.2018.11.067>.
- [66] T. Hastie. "Metode Berbasis Pohon." Universitas Stanford. <https://hastie.su.domains/MOOC/Slides/trees.pdf> (diakses).
- [67] M. Schweinberger. "Model Berbasis Pohon di R." Lang. Teknologi. Data Anal. Laboratorium. (LADAL). <https://ladal.edu.au/tree.html#References> (diakses).
- [68] Layanan EE, Ilmu Data dan Analisis Big Data: Menemukan, Menganalisis, 2015. Memvisualisasikan dan Menyajikan Data.
- [69] JV Tu, Keuntungan dan kerugian menggunakan jaringan saraf tiruan versus regresi logistik untuk memprediksi hasil medis, *J. Clin. Epidemiol.* 49 (11) (1996) 1225–1231, [https://doi.org/10.1016/s0895-4356\(96\)00002-9](https://doi.org/10.1016/s0895-4356(96)00002-9).
- [70] R. van de Schoot, JJ Broere, KH Perryck, M. Zondervan-Zwijnenburg, NE van Loey, Menganalisis kumpulan data kecil menggunakan estimasi Bayesian: kasus gejala stres pasca trauma setelah ventilasi mekanis pada korban luka bakar, dalam bahasa Inggris, *Eur. J. Psiko-traumatol.* 6 (2015). ARTN 2521610.3402/ejpt.v6.25216.
- [71] D. McNeish, Tentang penggunaan metode bayesian untuk mengatasi masalah sampel kecil, *Struct. Persamaan. Model: Seorang Multidisiplin*. J.23 (5) (2016) 750–773, <https://doi.org/10.1080/10705511.2016.1186549>.
- [72] FQ Han, M. Lei, WJ Zhao, JX Yang, Mesin vektor Dukungan baru untuk klasifikasi data ketidakseimbangan, *Intell. Otomatis. Komputasi Lembut*. 18 (6) (2012) 679–686, <https://doi.org/10.1080/10798587.2012.10643277>, dalam bahasa Inggris.
- [73] P. Kumar, R. Bhatnagar, K. Gaur, A. Bhatnagar, Klasifikasi data yang tidak seimbang: tinjauan metode dan aplikasi, *IOP Conf. Ser. Materi. Sains. bahasa Inggris* 1099 (1) (2021), <https://doi.org/10.1088/1757-899x/1099/1/012077>.
- [74] NV Chawla, KW Bowyer, LO Hall, WP Kegelmeyer, SMOTE: teknik pengambilan sampel berlebih minoritas sintetik, *J. Artif. Intel. Res.* 16 (2002) 321–357, <https://doi.org/10.1613/jair.953>, dalam bahasa Inggris.
- [75] S. Raschka, STAT 479: Catatan Kuliah Pembelajaran Mesin, Universitas Wisconsin-Madison, 2018.
- [76] S. Yadav, S. Shukla, dalam bahasa Inggris, dalam: *Analysis of K-fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification*, 2016, hlm. 78–83, <https://doi.org/10.1109/lacc.2016.25>. Int Conf Adv Compu.
- [77] R. Gutierrez-Osuna, *Pengantar Analisis Pola*, Texas A&M University, 2005.
- [78] TC Canturk, dkk., Tinjauan pelingkupan model prediksi komplikasi dalam bedah tulang belakang: analisis pengembangan model, validasi dan dampak, *North Am. Sosial Tulang Belakang*. J.11 (Sep 2022), 100142, <https://doi.org/10.1016/j.xnsj.2022.100142>.