Diterjemahkan dari bahasa Inggris ke bahasa Indonesia - www.onlinedoctranslator.com

Sistem Pakar Dengan Aplikasi 164 (2021) 113981

Daftar isi tersedia diSains Langsung

Sistem Pakar Dengan Aplikasi

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

Metode k-NN untuk prognosis kanker paru-paru dengan penggunaan algoritma genetika untuk pemilihan fitur

Negar MalekiA, Yasser ZeinaliB, Seyed Taghi Akhavan NiakiB,*,1

AJurusan Teknik Industri, Fakultas Teknik, Universitas Teheran, Iran
BDepartemen Teknik Industri, Universitas Teknologi Sharif, Teheran, Iran

INFO PASAL

Kata kunci:
Kanker paru-paru
Penambangan data diagnosis
stadium kanker
Algoritma genetika
Pemilihan fitur
teknik k-NN

ABSTRAK

Kanker paru-paru adalah salah satu penyakit paling umum yang diderita manusia di seluruh dunia. Identifikasi dini penyakit ini adalah pendekatan utama yang mungkin dilakukan untuk meningkatkan kemungkinan kelangsungan hidup pasien. Dalam makalah ini, teknik k-Nearest-Neighbours, yang mana algoritma genetika diterapkan untuk pemilihan fitur yang efisien guna mengurangi dimensi kumpulan data dan meningkatkan kecepatan pengklasifikasian, digunakan untuk mendiagnosis stadium penyakit pasien. Untuk meningkatkan akurasi algoritma yang diusulkan, nilai k terbaik ditentukan menggunakan prosedur eksperimental. Penerapan pendekatan yang diusulkan pada database kanker paru-paru menunjukkan akurasi 100%. Ini menyiratkan bahwa seseorang dapat menggunakan algoritme untuk menemukan korelasi antara informasi klinis dan teknik pengumpulan data untuk mendukung diagnosis stadium kanker paru-paru secara efisien.

1. Perkenalan

Mendiagnosis suatu penyakit adalah tugas yang sangat kompleks dan banyak tes biasanya diperlukan pada pasien untuk mencapai kesimpulan yang tepat. Hal ini dapat mengarahkan kita untuk menggunakan perangkat analitik, yang direncanakan untuk membantu dokter dalam mengambil keputusan. Penentuan dini akan mempersingkat waktu pengobatan dan dapat menyelamatkan nyawa. Salah satu penyakit tersebut adalah pertumbuhan ganas paru-paru, yang terjadi ketika sel-sel di jaringan paru-paru berkembang secara tidak terkendali. Pertumbuhan ini dapat menyebar ke luar paru-paru melalui proses metastasis ke jaringan terdekat atau bagian tubuh lainnya. Sebagian besar (85%) kasus kanker paru disebabkan oleh kebiasaan merokok dalam jangka panjang dan sekitar 10–15% kasus terjadi pada orang yang tidak pernah merokok (Thun dkk., 2008). Kasus-kasus ini sering kali disebabkan oleh kombinasi faktor genetik dan paparan gas radon, asbes, perokok pasif, atau bentuk polusi udara lainnya. Kanker paru-paru dapat dilihat pada rontgen dada dan pemindaian tomografi komputer (CT). Diagnosis dipastikan dengan biopsi yang biasanya dilakukan dengan bronkoskopi atau panduan CT. Keganasan paru merupakan salah satu penyakit yang menyebabkan 1,61 juta kematian di dunia setiap tahunnya (Li dkk., 2018). Kanker paru-paru menempati urutan kedua pada laki-laki dan kesepuluh pada perempuan (Naresh & Shettar, 2014). Tingkat kelangsungan hidup biasanya lebih tinggi jika keganasan dianalisis pada tahap awal. Itulah sebabnya pengungkapan dini pertumbuhan keganasan paru-paru menjadi sangat penting di mana sekitar 80% pasien berhasil dianalisis hanya pada tahap dalam atau masa penyakitnya (Wutsqa & Mandadara, 2017).

Pembelajaran mesin menggunakan algoritme ilmiah untuk membedakan pola dalam kumpulan data yang luas dan secara berulang meningkatkan penerapan bukti yang dapat dikenali ini dengan informasi tambahan. Algoritme ini umumnya digunakan di berbagai ruang dan aplikasi berbeda, misalnya, komersial, perlindungan, pendanaan, kehidupan berbasis internet, dan penemuan representasi keliru, mendapatkan berbagai jenis informasi yang dikumpulkan secara terus-menerus dan melalui berbagai sumber. Karena informasi pasien seringkali tidak dapat diakses untuk penyelidikan terbuka, memanfaatkan strategi ini untuk menilai hasil penyakit dapat menjadi tugas yang menantang (Lynch dkk., 2017).

Dalam makalah ini, metode pembelajaran mesin diterapkan untuk menyelidiki informasi mengenai keganasan paru-paru, untuk menilai intensitas sistem ini saat ini. Untuk tujuan ini, algoritma k-Nearest-Neighbors (k-NN) pertama kali dikembangkan untuk memprediksi kanker paru-paru pada tahap awal. Karena algoritma pemilihan fitur dapat mempengaruhi kinerja model kNN, algoritma genetika (GA) digunakan untuk mengoptimalkan model yang digunakan untuk memprediksi. Hal ini memungkinkan model mencapai akurasi yang lebih baik dalam tahap prediksi dan prognosis. Selain itu, nilai parameter k pada algoritma kNN ditentukan secara eksperimental dengan menggunakan pendekatan iteratif. Pada akhirnya, kinerja algoritma yang diusulkan dinilai ketika diterapkan pada database kanker paru-paru.

* Penulis yang sesuai.
*Alamat email:*maleki.negar@ut.ac.ir (N.Maleki),yasser.zeinali@gmail.com (Y.Zeinali),Niaki@sharif.edu (STA Niaki).
1PO Box 11155-9414 Azadi Ave., Teheran 1458889694, Iran.

https://doi.org/10.1016/j.eswa.2020.113981
Diterima 6 Juli 2019; Diterima dalam bentuk revisi 4 September 2020; Diterima 7 September 2020
Tersedia online 11 September 2020
0957-4174/© 2020 Elsevier Ltd. Semua hak dilindungi undang-undang.

Sisa makalah ini disusun sebagai berikut.Seksi 2memberikan gambaran umum tentang apa yang telah dilakukan dalam literatur tentang kanker paru-paru dan algoritma apa yang telah digunakan untuk diagnosis kanker.Bagian 3dengan cermat merinci teknik yang diusulkan, dan masukBagian 4 dan 5kinerja algoritma yang diusulkan dianalisis. Akhirnya,Bagian 6 menyimpulkan pekerjaan ini dan merekomendasikan pekerjaan di masa depan.

2. Tinjauan Pustaka

Pembelajaran mesin melibatkan beberapa algoritma seperti k-Nearest Neighbors (kNN), support vector machine (SVM), Naive Bayes (NBs), pohon klasifikasi (C4.5), mesin penguat gradien (GBM), dll. memproses data secara berbeda, di bagian ini, beberapa kandidat pembelajaran mesin yang baru-baru ini diusulkan di bidang temuan pertumbuhan ganas ditinjau secara kronologis.

Chen dkk. (2013)menyajikan sistem fuzzy menggunakan kNN (FkNN) untuk diagnosis penyakit Parkinson (PD). Selain itu, mereka menggunakan analisis komponen utama untuk menemukan fitur yang paling diskriminatif yang menjadi dasar dibangunnya model FkNN yang optimal. Mereka membandingkan sistem mereka dengan algoritma SVM dan menemukan bahwa metode yang mereka usulkan memiliki kinerja yang lebih baik. Akurasi klasifikasi terbaik FkNN mereka mencapai 96,07%.

Odajima & Pawlovsky (2014)menyatakan bahwa ketepatan metode kNN berubah seiring dengan jumlah tetangga dan tingkat informasi yang digunakan untuk klasifikasi. Sementara itu, mereka menunjukkan rincian variasi nilai akurasi maksimum dan minimum dengan ukuran himpunan klasifikasi dan jumlah tetangga.

Lynch dkk. (2017)menerapkan beberapa teknik klasifikasi pembelajaran yang diawasi seperti regresi linier, pohon keputusan, GBM, SVM, dan ansambel khusus ke database SIER untuk mengurutkan pasien kanker paru-paru mengenai kelangsungan hidup. Hasil penelitian menunjukkan bahwa di antara lima model individu yang digunakan, yang paling tepat adalah GBM dengan nilai root mean square error (RMSE) sebesar 15,32.Septiani dkk. (2017)membandingkan kinerja algoritma klasifikasi C4.5, NBs, dan kNN untuk mendeteksi diagnosis kanker payudara pada 670 data, masing-masing dengan 9 atribut. Mereka menunjukkan bahwa meskipun NB dan kNN memiliki akurasi yang sama yaitu 98,51%, C4.5 adalah yang terburuk dengan akurasi sebesar 91,79%.Hashi dkk. (2017) menggunakan algoritma pohon keputusan dan kNN untuk mendiagnosis penyakit diabetes dari Pima Indians Dataset yang mencakup 768 data, masing-masing dengan 8 atribut dan mencapai akurasi masing-masing 90,43% dan 76,96%. Hal ini menyiratkan bahwa pohon keputusan adalah metode dengan pengawasan yang lebih baik dalam hal akurasi klasifikasi dalam kasus ini. Kumpulan data ini juga telah digunakan Iyer dkk. (2015), Hayashi dan Yukita (2016), Sa'di dkk. (2015) dan Huang dkk. (2015)dimana mereka menerapkan metode pohon keputusan dan memperoleh akurasi masing-masing sebesar 76,96%, 83,83%, 76,52%, dan 62,17%. Khatieb dan Usman (2017)menggunakan teknik klasifikasi NB, kNN, J48, dan bagging classifier/ML pada dataset penyakit jantung yang terdiri dari 303 instance, masing-masing dengan 14 fitur. Mereka membagi hasil eksperimennya menjadi 6 kasus dan menemukan akurasi tertinggi sebesar 79,20% oleh pengklasifikasi kNN yang menggunakan seluruh 14 atribut. Lebih-lebih lagi,Tayeb dkk. (2017) menerapkan kNN juga pada kumpulan data yang dikumpulkan oleh Universitas California untuk menganalisis dua kondisi (gagal ginjal kronis dan penyakit jantung) dengan akurasi sekitar 90%.

Pradeep dan Naveen (2018)menggunakan teknik SVM, NBs, dan C4.5 pada kumpulan data kanker paru-paru North Central Cancer Treatment Group (NCCTG) untuk membantu spesialis mendapatkan kesimpulan yang lebih baik mengenai tingkat kelangsungan hidup kanker. Hasilnya menunjukkan bahwa C4.5 mempunyai kinerja yang lebih baik dalam memperkirakan keganasan paru-paru dengan peningkatan pada kumpulan data pelatihan.Al Harbi (2018)menggunakan algoritma gabungan genetik-fuzzy untuk mendiagnosis kanker paru-paru. Dia menerapkan algoritma tersebut pada 32 pasien dengan 56 atribut tanpa pengurangan dimensi apa pun dan mencapai akurasi 97,5% dengan kepercayaan 93%.Cherif (2018)mengembangkan solusi baru untuk mempercepat algoritma kNN yang bergantung pada pengelompokan dan pemisahan atribut pada database kanker payudara. Dia membandingkan algoritma yang diusulkannya dengan teknik klasifikasi lain seperti SVM, Artificial Neural Network (ANN), NBs, dan kNN. Dataset diisolasi menjadi 5 subset dari 113 kejadian, berdasarkan F-Measure setiap teknik dihitung lima kali untuk mencapai rata-rata F-Measure untuk masing-masing. Hasilnya menunjukkan bahwa meskipun ANN memiliki performa terbaik, waktu eksekusinya 2,2 kali lebih lama dibandingkan algoritma yang diusulkan.Joshi dan Mehta (2018)menggunakan algoritme pembelajaran mesin (kNN) yang terkenal untuk memeriksa pelaksanaannya pada kumpulan data diagnostik kanker payudara Wisconsin. Dataset tersebut melibatkan 569 instance dengan 32 atribut dan 2 kelas. Mereka menggunakan dua strategi reduksi dimensi penting (analisis komponen utama (PCA) dan analisis diskriminan linier (LDA) dan menunjukkan bahwa kNN dengan teknik LDA bekerja lebih baik daripada kNN dan kNN dengan PCA dengan akurasi masing-masing 97,06%, 95,29%, dan 95,88%. .Akben (2018) memanfaatkan kNN, SVM, dan NB untuk memproses data sebelumnya untuk mendeteksi penyakit ginjal kronis (CKD). Mereka pertama kali menggunakan metode pada data mentah dan menemukan bahwa klasifikasi tersebut tidak cukup akurat untuk menyemangati praktisi pengobatan. Oleh karena itu, mereka menggunakan metode tersebut setelah data diproses sebelumnya dengan pendekatan pengelompokan k-means. Hasilnya menunjukkan bahwa akurasi meningkat secara signifikan, terutama untuk pengklasifikasi kNN yang mencapai 96%.

Lakshmanaprabu dkk. (2019)mengembangkan algoritma hibrida yang melibatkan jaringan saraf dalam yang optimal (ODNN) dan analisis diskriminasi linier (LDA) untuk mengklasifikasikan nodul paru-paru sebagai ganas atau jinak. Dalam pekerjaannya, ODNN pertama kali digunakan untuk mengekstraksi fitur-fitur penting dari gambar paru-paru yang dihitung dengan tomografi (CT). Kemudian LDA diterapkan untuk mereduksi dimensi fitur. Terakhir, algoritma pencarian gravitasi yang dimodifikasi digunakan untuk mengoptimalkan ODNN. Sensitivitas, spesifisitas, dan akurasi algoritmanya masing-masing ditunjukkan sebesar 96,2%, 94,2%, dan 94,56%. Baru-baru ini,Alirezaei dkk. (2019) mengerahkan empat bi-tujuan*meta*-algoritma heuristik (multiobjective firefly (MOFA), multi-objective imperialist kompetitif algoritma (MOICA), algoritma genetika penyortiran non-dominated (NSGA-II), dan multi-objective Particle Swarm Optimization (MOPSO)) untuk menentukan jumlah atribut yang paling sedikit dengan tingkat akurasi klasifikasi tertinggi. Karena pentingnya kualitas data, pertama-tama mereka menggunakan beberapa metode pra-pemrosesan. Kemudian SVM digunakan sebagai pengklasifikasi. Di antara yang di atas*meta* -heuristik, MOFA adalah yang terbaik dengan akurasi 95,12%.

Sebagai pengklasifikasi yang diawasi, K-Nearest Neighbor digunakan sekali lagi dalam makalah ini pada kumpulan data yang tersedia untuk memprediksi kanker paru-paru pada tahap awal. SebagaiOdajima dan Pawlovsky (2014) menunjukkan bahwa nilai parameter k yang berbeda dalam algoritma kNN mempengaruhi hasil secara signifikan, pendekatan baru dalam lingkungan Python dikembangkan untuk menemukan nilai k terbaik. Selanjutnya, algoritma genetika (GA) digunakan untuk mencari fitur terbaik. Pendekatan yang diusulkan dijelaskan secara rinci.

3. Pendekatan yang diusulkan

Metodologi yang diusulkan merupakan penyempurnaan dari metode kNN. Bagian ini secara singkat memberikan latar belakang metode kNN. Kami kemudian menunjukkan bagaimana GA dapat meningkatkan akurasi metode kNN.

3.1. k-Pengklasifikasi Tetangga Terdekat

Pengklasifikasi kNN telah banyak digunakan di bidang pengenalan pola. Pengklasifikasi tetangga terdekat bergantung pada pembelajaran melalui hubungan, yaitu dengan mengontraskan tupel pengujian tertentu dan menyiapkan tupel yang serupa dengannya. Tupel persiapan digambarkan oleh$N$ sifat-sifat. Setiap tupel mengacu pada suatu titik di aN-ruang dimensi; karenanya, semua tupel persiapan disimpan di aN-ruang contoh dimensi. Ketika diberikan tupel yang tidak jelas, pengklasifikasi k-tetangga terdekat melihat ruang contoh untuk tupel yang menyiapkan K yang paling dekat dengan tupel yang tidak jelas. Tupel persiapan-k ini adalah k “tetangga terdekat” dari tupel yang tidak jelas.

Kedekatan pada algoritma kNN ditandai dengan metrik pemisahan, misalnya jarak Euclidean. Jarak Minkowski antara dua tupel, katakanlah, $X_1 = (X_{11}, X_{12}, \cdots, X_{1N})DanX_2 = (X_{21}, X_{22}, \cdots, X_{2N})$, adalah:

$$dist(X_1,X_2) = \left(\sum_{Saya=1}^{N} |X_{Saya} - X_j|^{p}\right)^{1/P} \quad (1)$$

Untuk setiap karakteristik numerik suatu titik data, perbedaan antara estimasi yang berkaitan dengan karakteristik tersebut dalam tupel X_1 Dan X_2 pertama-tama diwujudkan dengan mengkuadratkan jarak, lalu menjumlahkannya untuk semua karakteristik. Akar kuadrat selanjutnya diambil dari penghitungan pemisahan agregat. Biasanya estimasi setiap karakteristik dinormalisasi sebelum menggunakan Persamaan.(1). Ini mungkin akan meningkatkan tingkat akurasi algoritma.

Nilai k yang sesuai dapat diperoleh secara eksperimental. Dimulai dengan k = 1, set pengujian digunakan dengan Python untuk mengevaluasi tingkat kesalahan pengklasifikasi. Prosedur ini dapat diulang setiap kali dengan menambah k untuk memasukkan satu tetangga lagi. Nilai k yang memberikan tingkat kesalahan dasar terbaik dipilih (Han dkk., 2011).

3.2. Implementasi algoritma genetika

GA adalah metode pencarian heuristik. Hal ini dapat dimanfaatkan untuk mencari solusi optimal pada ruangan yang terlalu luas untuk dilihat secara komprehensif. Algoritma ini adalah metode untuk memecahkan masalah optimasi terbatas dan tidak terbatas yang didasarkan pada seleksi alam, proses yang mendorong evolusi biologis. Ini memiliki banyak penerapan antara lain dalam ilmu alam, matematika, ilmu komputer, keuangan dan ekonomi, industri, manajemen, dan teknik. Hal ini dapat mencerminkan prosedur penentuan karakteristik dalam algoritma kNN. Ada lima fase dalam algoritma genetika:

1. Populasi awal
2. Fungsi kebugaran
3. Seleksi
4. Persilangan
5. Mutasi

Teknik GA adalah metode berulang yang mencakup komunikasi populasi ke ruang pandang untuk menemukan jawaban atas suatu masalah melalui serangkaian gambar terbatas, yang disebut genom, yang dikumpulkan dalam sebuah kromosom (solusi). Prinsip dasar GA terus berlanjut: populasi kromosom yang mendasarinya diproduksi tanpa pandang bulu atau secara heuristik. Dalam setiap kemajuan perkembangan (generasi), kromosom dalam populasi didekodekan dan dinilai dengan fungsi kebugaran yang menggambarkan masalah perampingan dalam ruang pencarian. Untuk membentuk populasi lain (generasi berikutnya), kromosom dipilih berdasarkan kebugarannya. Di sini, banyak pilihan yang tersedia, salah satu yang paling rumit adalah pilihan kebugaran proporsional, di mana kromosom dipilih dengan kemungkinan yang sesuai dengan kebugaran relatifnya. Hal ini menjamin berapa kali normal individu terpilih berada di sekitar sesuai dengan kinerja relatifnya dalam populasi. Oleh karena itu, kromosom dengan kebugaran tinggi memiliki peluang lebih besar untuk menciptakan kembali dan menyampaikan individu baru ke dalam populasi, sedangkan kromosom dengan kebugaran rendah tidak.

Kromosom baru dibawa ke dalam populasi melalui operasi keturunan yang disebut persilangan dan mutasi. Operasi persilangan dilakukan dengan kemungkinan antara dua individu terpilih (induk) yang memperdagangkan bagian genomnya untuk membentuk dua kromosom baru (keturunan). Sementara itu, operasi mutasi mencegah penyatuan sebelum waktunya ke titik optima terdekat dengan memeriksa fokus baru secara acak di ruang perburuan; ini dilakukan dengan membalik bit secara acak, dengan kemungkinan rendah. GA merupakan proses iteratif stokastik yang tidak menjamin menemukan titik optimum. Selain itu, kondisi penghentian dapat diindikasikan sebagai jumlah generasi maksimal atau nilai kebugaran yang diinginkan.

3.3. Kriteria kinerja

Akurasi merupakan salah satu kriteria kinerja yang mempunyai beberapa arti daerah yang berbeda. Namun, dalam metode klasifikasi, akurasi didefinisikan sebagai ukuran statistik seberapa baik pengujian klasifikasi biner mengidentifikasi atau mengecualikan suatu kondisi dengan benar. Artinya, akurasi adalah proporsi hasil yang sebenarnya (baik positif maupun negatif sebenarnya) di antara jumlah total kasus yang diperiksa dalam eksperimen. Persamaan.(2)digunakan untuk mengukur akurasi biner:

$$Ketepatan = (dll + TN) / (dll + TN + FP + FN) \quad (2)$$

Di mana, dll=Benar-benar positif, FP=Positif palsu, TN=Benar-benar negatif, FN =Negatif palsu.

Semua besaran di atas dapat diekstraksi menggunakan matriks konfusi; tabel yang sering digunakan untuk menggambarkan kinerja model klasifikasi pada sekumpulan data uji yang diketahui nilai sebenarnya (Melamed dkk., 2003).

Kriteria kinerja lainnya adalah “sensitivitas” dan “spesifisitas”, yang juga dikenal dalam statistik sebagai fungsi klasifikasi, yang banyak digunakan dalam studi kedokteran dan bioinformatika. Sensitivitas atau penarikan kembali mengukur proporsi positif sebenarnya yang diidentifikasi dengan benar dan spesifisitas juga mengukur proporsi negatif sebenarnya dalam eksperimen. Persamaan.(3) dan (4)mendefinisikan langkah-langkah ini.

$$Kepekaan = TP / (dll + FN) \quad (3)$$

$$Kekhususan = TN / (TN + FP) \quad (4)$$

3.4. Pemilihan fitur

Tujuan pertama dalam metode pemilihan fitur yang diusulkan adalah untuk mencapai setidaknya tingkat akurasi yang sama dengan keseluruhan fitur. Tujuan kedua adalah meningkatkan tingkat akurasi. Di sini, pengumpulan informasi ekstensif mengenai fitur-fitur tidak hanya menghabiskan banyak waktu dan uang, tetapi juga informasi tambahan mengakibatkan pemborosan waktu dalam pengklasifikasian dan diagnosis. Oleh karena itu, lebih baik mengurangi dimensi jumlah fitur untuk mendapatkan respons yang lebih baik dan menemukan korelasi yang lebih baik antara fitur dan hasil.

Algoritma genetika merupakan suatu teknik untuk menyeleksi fitur-fitur terbaik. Dalam teknik ini, vektor acak biner $vektor S$terdiri dari fitur-fitur yang pertama kali dihasilkan menggunakan Persamaan.(5)(Pawlovsky & Hiroki, 2017):

$$Vektor(s_i): \quad S_j = Y_{Saya}; Y_{Saya} = \begin{cases} 1 ; jika\ Vektor_{Sj} berisi\ fitur\ i \\ 0 ; \quad jika\ tidak \end{cases} \quad (5)$$

Kemudian, fungsi tujuan berdasarkan kriteria kinerja kesalahan klasifikasi ditentukan untuk setiap kombinasi fitur yang dipilih. Fungsi tujuan ini berfungsi sebagai fungsi penalti yang harus diminimalkan untuk menemukan kombinasi fitur terbaik. Di sini, tingkat kesalahan klasifikasi (mcr)adalah secara sederhana mcr=1 - $Tingkat\ akurasi$dan diperoleh dengan menggunakan Persamaan.(6), Di mana Madalah jumlah target klasifikasi dan $A_{aku\,j}$adalah jumlah kasus yang menjadi target $Saya$diklasifikasikan sebagai sasaran $j$menggunakan metode klasifikasi. Itu $A_{aku\,j}$elemen membangun matriks di(7)disebut matriks konfusi yang bergantung pada masalah serta kumpulan datanya (Pawlovsky & Hiroki, 2017):

$$mcr = \frac{\sum_{A_{aku\,j},\ j\neq} \left[\sum_{A_{aku\,j},\ j,\ (Saya=j)} A_{aku\,j,\ j}\right]}{\sum_{A_{aku\,j}} A_{aku\,j}}; aku\ j = 1,2,\cdots,M \quad (6)$$

$$\begin{pmatrix} A_{11} & \cdots & A_{1M} \\ \vdots & \ddots & \vdots \\ A_{M1} & \cdots & A_{mm} \end{pmatrix}_{M\times M} \quad (7)$$

Sekarang, fungsi tujuan yang harus diminimalkan adalah penjumlahan tertimbang dari mcrDan N_F(jumlah fitur yang dipilih) didefinisikan sebagai

$$MinZ = w_1 * mcr + w_2 * N_F \quad (8)$$

Membagi ruas kanan Persamaan.(8)oleh w_1, kita punya:

$$MinZ = mcr + w_2/w_1 * N_F \quad (9)$$

Tabel 1
Atribut (fitur) yang terlibat dalam kumpulan data.

Usia (1)	Jenis Kelamin (2)	Polusi Udara (3)	Penggunaan alkohol (4)	Alergi Debu (5)	Bahaya Pekerjaan (6)
Risiko Genetik (7)	Batuk Darah (8) Penyakit	Kelelahan (9)	Penurunan Berat Badan (10)	Merokok (11)	Mengi (12)
Sakit dada (13)	Paru-Paru Kronis (14) Kuku	Pola Makan Seimbang (15)	Obesitas (16)	Sesak Nafas (17)	Perokok Pasif (18)
Kesulitan Menelan (19)	Jari Tabuh (20)	Sering Pilek (21)	Batuk Kering (22)	Mendengkur (23)	

Asumsi $w_2/w_1 = W$, fungsi tujuan menjadi:

$$MinZ = mcr + W * N_F \quad (10)$$

Sekarang, W dapat didefinisikan sebagai:

$$\begin{array}{l} W \propto mcr \rightarrow \\ W = \beta * mcr \rightarrow \\ MinZ = mcr + \beta * mcr * N_F \end{array} \quad (11)$$

Hal ini mengarah pada:

$$MinZ = mcr(1 + \beta * N_F) \quad (12)$$

Di mana β dapat didefinisikan sebagai penalti karena memiliki fitur tambahan ($0 \le \beta \le 1$).

Dengan menggunakan fungsi tujuan ini, GA mencoba menemukan kombinasi fitur terbaik dengan jumlah fitur minimum yang meminimalkan biaya dan tingkat kesalahan klasifikasi. Di sini, kriteria penghentian untuk mengakhiri iterasi di GA dipilih menjadi jumlah iterasi yang telah ditentukan sebelumnya.

3.5. Deskripsi data

Kotak yang ditunjukkan pada gambar ini adalah "Dataset". Pentingnya dataset menjadi bagian penelitian yang tidak dapat disangkal karena mempengaruhi hasil akhir. Kumpulan data kanker paru-paru yang dipertimbangkan diperoleh dari situs Data world (https://data.world/cancerdatahp/lung-cancer-data) berisi 1000 sampel, masing-masing dengan 23 fitur ditampilkanTabel 1. Sasaran dalam dataset ini adalah tingkat risiko penderitaan kanker paru-paru yang diklasifikasikan dalam 3 tingkatan yaitu Rendah, Sedang, dan Tinggi (lihatMeja 2).

4. Kerangka usulan prosedur diagnosis kanker paru

Struktur umum dari prosedur diagnosis yang diusulkan digambarkan dalam Gambar 1. Memiliki dataset, kotak berikutnya masukGambar 2adalah untuk memeriksa apakah diperlukan pra-pemrosesan untuk menghapus nilai yang hilang atau menggantinya dengan data yang sesuai. Barisan kumpulan data yang tidak sempurna bisa saja dihapus, tetapi kami memutuskan untuk secara otomatis mengisi nilai yang hilang menggunakan fungsi perangkat lunak untuk memanfaatkan rata-rata nilai lainnya. Kemudian, di kotak berikutnya, GA menerapkan kumpulan data bersih untuk menemukan kombinasi fitur terbaik yang memberikan korelasi tertinggi antara fitur dan target. Untuk tujuan ini, vektor *vektorS* diperoleh diGambar 3.

Sedangkan jumlah iterasi maksimum ditetapkan 10, setelah iterasi keempat nilai fungsi biaya konvergen menjadi 0,53266 seperti yang ditunjukkan pada Gambar 4. Di sini ukuran populasi adalah 20 dan probabilitas operator persilangan dipilih 0,7, probabilitas mutasi ditetapkan 0,02, berdasarkan jumlah keturunan yang dihasilkan adalah 14 dan jumlah mutan adalah 6. Selain itu, roda roulette metode memilih orang tua dalam semua operasi.

$$Jumlah \quad keturunan = 2 * bulat\left(\frac{Persimpangan * Pop}{2}\right) \quad (13)$$

$$Jumlah \quad mutan = bulat(Persentase\ mutasi * Pop) \quad (14)$$

Setelah menerapkan GA, pengklasifikasi kNN diterapkan pada pelatihan kumpulan data untuk mempelajari cara mengenali target. Pengklasifikasi kNN membutuhkan data untuk k-Nearest Neighbors untuk mengklasifikasikannya guna mendeteksi target dengan benar. Karena parameter k mempengaruhi kinerja klasifikasi secara signifikan, pendekatan berulang dalam Python digunakan untuk menemukan nilai k yang sesuai. Nilai k yang sesuai dapat diperoleh secara eksperimental seperti yang telah kami katakan sebelumnya. Dimulai dengan k = 1, set pengujian digunakan dengan Python untuk mengevaluasi pengklasifikasi. Prosedur ini dapat diulang setiap kali dengan menambah k untuk memasukkan satu tetangga lagi. Akhirnya, k terbaik akan dipilih untuk digunakan dalam model. Hasilnya ditunjukkan pada Gambar. 5 dan 6.

Hal ini jelas dariGambar. 5 dan 6bahwa GA tidak hanya meningkatkan tingkat akurasi algoritma kNN, tetapi juga k = 7 dan k = 6 masing-masing memberikan akurasi maksimum ketika GA digunakan dan tidak digunakan.

Di kotak terakhir digambarkan dalamGambar 2, akurasi kNN yang dilatih ketika diterapkan pada kumpulan data pengujian dibandingkan satu sama lain untuk mengevaluasi kinerja metodologi yang diusulkan.

5. Analisis perbandingan

Pada bagian ini, kinerja metodologi yang diusulkan dalam hal akurasi dibandingkan dengan tiga pendekatan berbeda termasuk pohon keputusan, kNN tanpa metode pemilihan fitur (GA) yang diusulkan dan tanpa mengonfigurasi parameter k, dan kNN dengan pendekatan pemilihan fitur yang diusulkan yang melibatkan konfigurasi k-parameter. Untuk mencapai tujuan ini, 500 pasien dipilih secara acak dari kumpulan data kanker paru-paru untuk digunakan dalam semua metode.

Matriks konfusi dari pendekatan pohon keputusan ditunjukkan pada Tabel 3. Terlihat pada tabel ini, akurasi metode pohon keputusan diperoleh sebesar 95,2%.

Metode kedua adalah kNN tanpa menggunakan metode pemilihan fitur maupun pendekatan konfigurasi k-parameter. Matriks konfusi dari metode ini dimasukkanTabel 4.

Untuk menganalisis domain penerapan eksperimen, kita perlu mempartisi kumpulan data menjadi dua bagian berbeda (pelatihan dan pengujian), yang menjadi dasar analisis perbedaan antara tingkat akurasi. Kami mendedikasikan 80% kumpulan data untuk pelatihan dan 20% untuk kumpulan pengujian. Eksperimen ini menghasilkan tingkat akurasi sebesar 100 persen untuk set pelatihan dan tingkat akurasi sebesar 96,2 persen untuk set pengujian. Hasil tersebut diperoleh dengan mengimplementasikan kNN dengan K sama dengan 10. Jelas bahwa perbedaan antara kedua tingkat akurasi ini tidak signifikan; oleh karena itu, model tersebut dapat diterapkan.

Hasil diTabel 4juga menunjukkan bahwa akurasi pendekatan kNN tanpa menggunakan algoritma GA untuk memilih kombinasi fitur terbaik adalah 96,2 persen ketika k-parameter diatur ke 10. Meskipun akurasi ini lebih baik dibandingkan dengan yang diperoleh dengan menggunakan metode pohon keputusan (95,2% masukTabel 3), selanjutnya dinaikkan menjadi 99,8%, ketika K diubah dari 10 menjadi 6. Matriks konfusi dari pendekatan ini ditunjukkan pada Tabel 5.

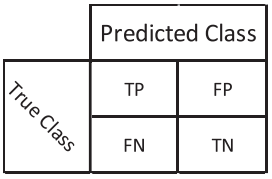
Terlebih lagi, ketika GA diterapkan, akurasinya menjadi lebih baik. Tabel 6menunjukkan kesimpulan ini. Dengan kata lain, tingkat akurasi Metode kNN dengan k = 6 tetangga yang menggunakan algoritma pemilihan fitur merupakan yang tertinggi.

Kami juga menerapkan validasi silang 10 kali lipat untuk set pelatihan untuk mendapatkan skor sebagai berikut.

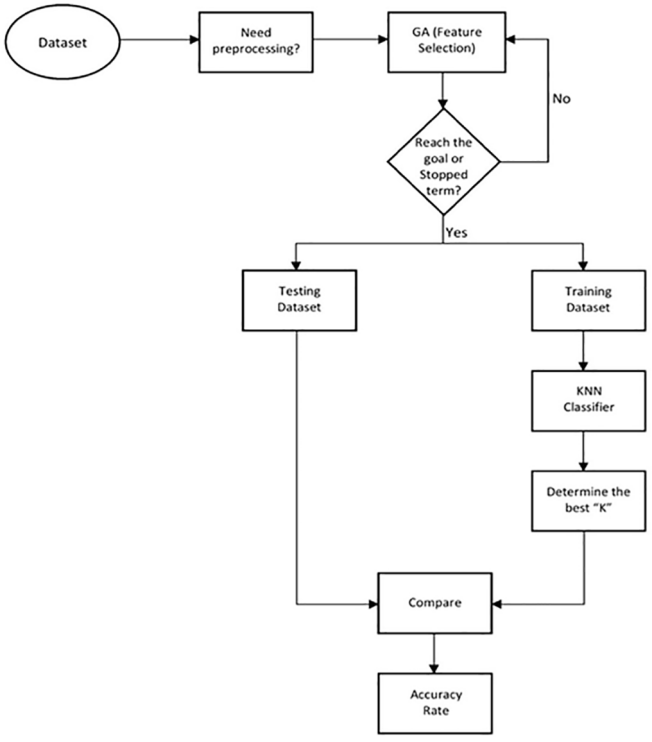
1	1	1	1	0,96078431	1	0,95918367	0,97959184	1	0,97916667

Meja 2
Estimasi parameter GA.

Iterasi Maks	Pop.	% Persilangan	% Mutasi	Waktu)	Fungsi biaya	Vektor yang Dipilih
10	20	0,7	0,3	912.5702	0,51265	[2,7,10,16,17,19]
10	20	0,8	0,3	425.12316	0,54523	[10,11,15,16,17,19,20]
10	20	0,9	0,3	400.12805	0,54523	[10,11,15,16,17,19,20]
10	20	0,7	0,4	512.17854	0,56425	[2,10,13,14,15,16,17,19]
10	20	0,8	0,4	607.71827	0,56418	[2,10,13,14,15,16,17,18,19]
10	20	0,9	0,4	894.39541	0,56425	[2,10,13,14,15,16,17 ,19]
10	20	0,7	0,5	413.03148	0,56425	[2,10,13,14,15,16,17,19]
10	20	0,8	0,5	397.124976	0,56425	[2,10,13,14,15,16,17,19]
10	20	0,9	0,5	801.900148	0,56425	[2,10,13,14,15,16, 17,19]
10	50	0,7	0,3	759.214019	0,5257	[2,10,15,16,17,19]
10	50	0,8	0,3	989.107872	0,56418	[2,10,13,14,15,16,17,18,19]
10	50	0,9	0,3	1251.439019	0,53421	[10,11,15,16,17,18,19]
10	50	0,7	0,4	1624.20197	0,56418	[2,10,13,14,15,16,17,18,19]
10	50	0,8	0,4	1724.219054	0,56418	[2,10,13,14,15,16,17,18,19]
10	50	0,9	0,4	2078.10536	0,55425	[2,10,13,14,15,16 ,17,19]
10	50	0,7	0,5	1954.028514	0,56418	[2,10,13,14,15,16,17,18,19]
10	50	0,8	0,5	1207.714546	0,56418	[2,10,13,14,15,16,17,18,19]
10	50	0,9	0,5	2007.167903	0,55425	[2,10,13, 14,15,16,17,19]
10	80	0,7	0,3	1627.21883	0,56418	[2,10,13,14,15,16,17,18,19]
10	80	0,8	0,3	1405.15904	0,56418	[2,10,13,14,15,16,17,18,19] [2
10	80	0,9	0,3	2104.01791	0,56418	,10,13,14,15,16,17,18,19]
10	80	0,7	0,4	1721.8028	0,56418	[2,10,13,14,15,16,17,18,19]
10	80	0,8	0,4	2845.677454	0,56418	[2,10,13,14,15,16,17 ,18,19]
10	80	0,9	0,4	2157.01385	0,56418	[2,10,13,14,15,16,17,18,19]
10	80	0,7	0,5	2278.02138	0,56418	[2,10,13,14,15,16,17,18,19]
10	80	0,8	0,5	1984.98026	0,56418	[2,10,13, 14,15,16,17,18,19]
10	80	0,9	0,5	2310.14806	0,56418	[2,10,13,14,15,16,17,18,19]



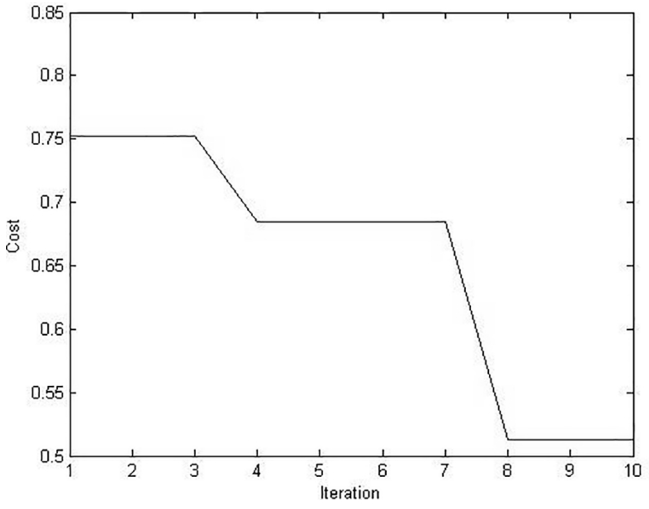
Gambar 1. Matriks Kebingungan.



Gambar 2.saudara itu kerja keras dari prosedur diagnosis kanker paru-paru yang diusulkan.

$$Vector\ S = [2, 7, 10, 16, 17, 19]; n_f = 6$$

Gambar 3. Pilihan fitur GA.



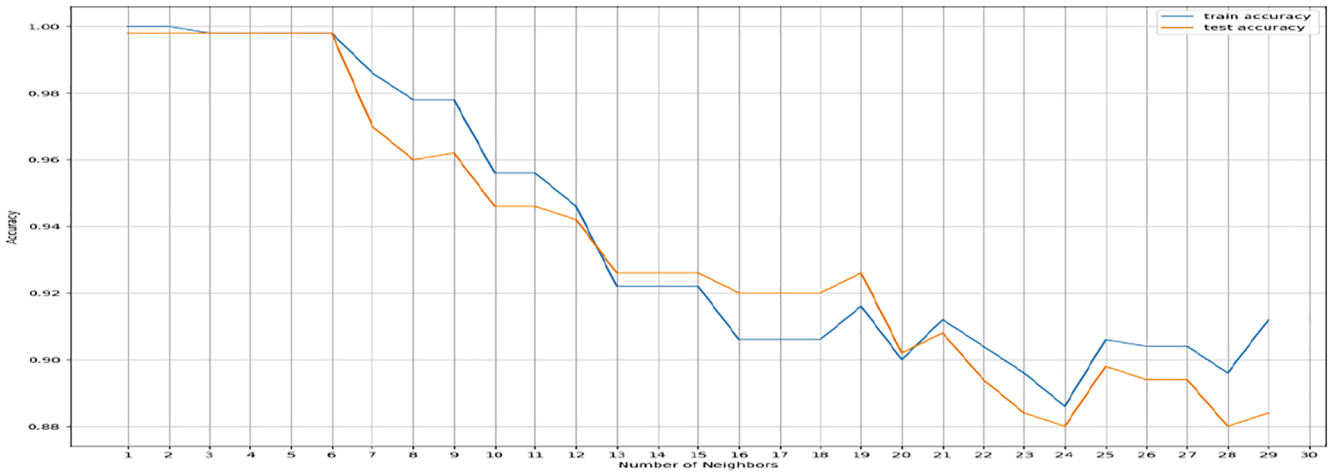
Gambar 4. Nilai fungsi biaya terbaik (0,51265).

Seperti yang terlihat jelas pada gambar di atas, skornya adalah 1 atau sangat mendekati 1.

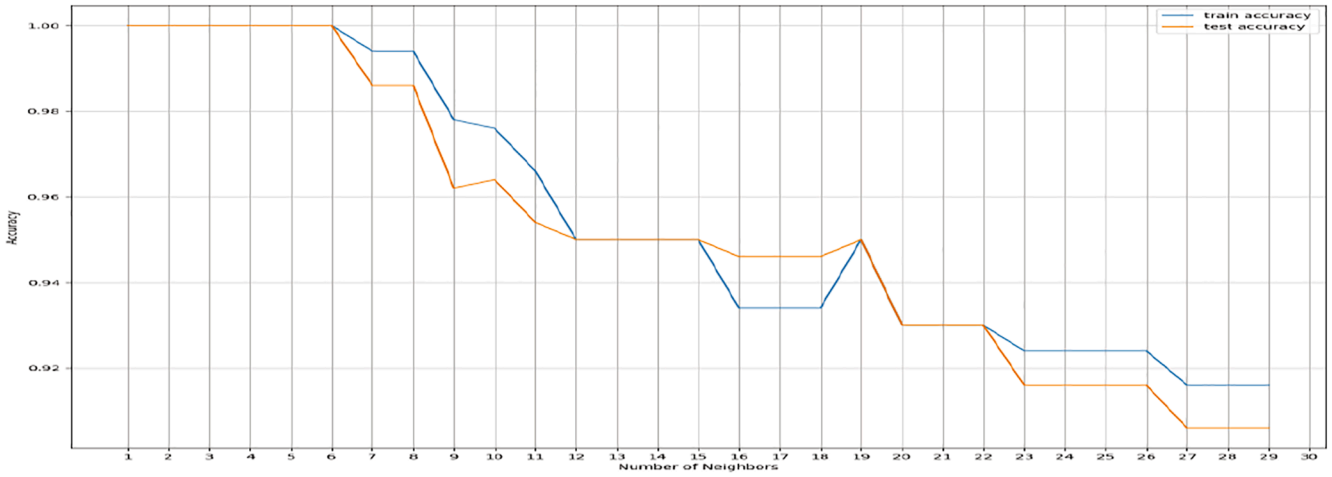
Sebagai perbandingan terakhir, di sini kami memiliki waktu CPU yang berbeda untuk metodologinyaTabel 7. Hasil pada tabel ini menunjukkan bahwa GA tidak hanya mempengaruhi akurasi k-NN, tetapi juga menurunkan waktu CPU secara signifikan.

6. Kesimpulan dan pekerjaan masa depan

Meskipun ada banyak metode pembelajaran mesin yang tersedia dalam literatur yang kinerjanya bergantung pada berbagai aspek termasuk kumpulan data tempat metode tersebut diterapkan, dalam makalah ini, metode pembelajaran mesin yang disebut kNN digabungkan dengan algoritma genetika pemilihan fitur untuk



Gambar 5.tingkat akurasi kNN sebelum menerapkan GA menggunakan "K" yang berbeda.



Gambar 6.tingkat akurasi kNN setelah GA menggunakan "K" yang berbeda.

Tabel 3
Matriks konfusi dari metode pohon keputusan.

Dari untuk	Tinggi	Rendah	Sedang	Total	%Ketepatan
Tinggi	167	6	0	173	96,53%
Rendah	0	167	6	173	96,53%
Sedang	0	12	142	154	92,21%
Total	167	185	148	500	95,20%

Tabel 4
Matriks konfusi kNN tanpa GA ("K" disetel ke 10).

Dari untuk	Tinggi	Rendah	Sedang	Total	%Ketepatan
Tinggi	150	0	7	157	95,54%
Rendah	0	151	5	156	96,79%
Sedang	0	7	180	187	96,25%
Total	150	158	192	500	**96,20%**

Tabel 5
Matriks konfusi kNN tanpa GA ("K" disetel ke 6).

Dari untuk	Tinggi	Rendah	Sedang	Total	%Ketepatan
Tinggi	150	1	0	151	99,35
Rendah	0	166	0	166	100,00%
Sedang	0	0	183	183	100,00%
Total	150	167	183	500	**99,80%**

Tabel 6
Matriks konfusi kNN dengan GA ("K" disetel ke 6).

Dari untuk	Tinggi	Rendah	Sedang	Total	%Ketepatan
Tinggi	151	0	0	151	100,00%
Rendah	0	166	0	166	100,00%
Sedang	0	0	183	183	100,00%
Total	151	166	183	500	100,00%

mengklasifikasikan risiko pasien kanker paru dalam tiga tingkatan yaitu rendah, sedang, dan tinggi. Tujuan penggunaan GA adalah untuk menentukan kombinasi fitur terbaik yang meminimalkan kesalahan perhitungan metode kNN secara keseluruhan. Selain itu, nilai terbaik untuk jumlah tetangga pada algoritma kNN ditentukan menggunakan algoritma yang dikodekan dengan Python. Terlihat bahwa ketika metode kNN digabungkan dengan algoritma pemilihan fitur, akurasi klasifikasi meningkat secara signifikan. Seperti disebutkan sebelumnya, 6 fitur telah dipilih melalui algoritma GA, yaitu [2, 7, 10, 16, 17, 19], nilai fungsi biaya dikonvergensi pada iterasi keempat dan membutuhkan waktu sekitar 912,5702 detik untuk menjalankan program dan juga nilai k terbaik adalah 6. Semua komputasi dilakukan pada laptop dengan CPU 2,20 GHz dan RAM 2 GB. Eksperimen dan hasil ini dianalisis secara cermat oleh seorang spesialis kanker paru-paru. Selain itu, spesialis menganalisis hasil menggunakan data klinis beberapa pasien dan membandingkan kondisi sebenarnya dengan kelas yang dikhususkan model untuk pasien tersebut.

Pekerjaan di masa depan mungkin melibatkan penggunaan algoritma klasifikasi pembelajaran mesin lainnya atau penggunaan fitur berbasis populasi lainnya

Tabel 7
Perbandingan model dan hasilnya.

Algoritma	Ketepatan	Waktu (Hanya Algoritma ML)
Pohon Keputusan	95,40%	0,015996
k-NN (k = 10) Jarak Euclidian k-NN	96,40%	0,0312
(k = 6)	99,80%	0,0313
GA dulu, lalu k-NN	100%	0,0156

pilihan *meta*-heuristik dan membandingkan kinerjanya dengan kinerja yang diperoleh dari pendekatan yang diusulkan.

Pernyataan kontribusi kepenulisan CReditT

Nagar Maleki: Metodologi, Perangkat Lunak, Validasi, Analisis formal, Investigasi, Kurasi data, Penulisan - draf asli, Visualisasi. **Yasser Zeinali:** Metodologi, Perangkat Lunak, Validasi, Analisis formal, Investigasi, Kurasi data, Penulisan - draf asli, Visualisasi. **Seyed Taghi Akhavan Niaki:** Konseptualisasi, Sumber Daya, Penulisan - review & editing, Pengawasan, Administrasi proyek.

Deklarasi Kepentingan Bersaing

Para penulis menyatakan bahwa mereka tidak mempunyai kepentingan finansial atau hubungan pribadi yang saling bersaing yang dapat mempengaruhi pekerjaan yang dilaporkan dalam makalah ini.

Referensi

Akben, SB (2018). Diagnosis penyakit ginjal kronis tahap awal dengan menerapkan data metode penambangan hingga urinalisis, analisis darah dan riwayat penyakit. *IRBM, 39*(5), 353–358.

Alharbi, A. (2018). Sebuah sistem komputer otomatis berdasarkan algoritma genetika dan fuzzy sistem untuk diagnosis kanker paru-paru. *Jurnal Internasional Ilmu Nonlinier dan Simulasi Numerik, 19*(6), 583–594.

Alirezaei, M., Niaki, STA, & Akhavan Niaki, SA (2019). Hibrida bi-objektif algoritma optimasi untuk mengurangi noise dan dimensi data dalam diagnosis diabetes menggunakan mesin vektor pendukung. *Sistem Pakar dengan Aplikasi, 127*, 47–57. Chen, H.-L., Huang, C.-C., Yu, X.-G., Xu, X., Sun, X., Wang, G., & Wang, S.-J. (2013). Sebuah sistem diagnosis yang efisien untuk mendeteksi penyakit Parkinson menggunakan pendekatan fuzzy k-nearest neighbour. *Sistem pakar dengan aplikasi, 40*(1), 263–271. Cherif, W. (2018). Optimalisasi algoritma K-NN dengan clustering dan reliabilitas koefisien: Aplikasi untuk diagnosis kanker payudara. *Ilmu Komputer Procedia, 127*, 293–299.

Han, J., Pei, J., & Kamber, M. (2011). *Penambangan data: Konsep dan teknik.* Elsevier.

Hashi, EK, Zaman, SU, Hasan, R. (2017). Sistem pendukung keputusan klinis ahli untuk memprediksi penyakit menggunakan teknik klasifikasi. Dalam Prosiding konferensi internasional 2017 tentang teknik kelistrikan, komputer dan komunikasi (ECCE), Cox's Bazar, Bangladesh.

Hayashi, Y., & Yukita, S. (2016). Ekstraksi aturan menggunakan ekstraksi aturan rekursif algoritma dengan graft J48 dikombinasikan dengan teknik pemilihan sampling untuk diagnosis diabetes melitus tipe 2 pada dataset Pima Indian. *Informatika dalam Kedokteran Tidak Terkunci, 2*, 92–104.

Huang, G.-M., Huang, K.-Y., Lee, T.-Yi, Weng, J. (2015). Berbasis aturan yang dapat ditafsirkan klasifikasi diagnostik nefropati diabetik pada pasien diabetes tipe 2. Dalam artikel pilihan dari konferensi bioinformatika Asia Pasifik ketigabelas (APBC 2015) (Vol. 16, Supplement 1, p. S5).

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis diabetes menggunakan klasifikasi teknik penambangan. *Jurnal Internasional Proses Data Mining & Knowledge Management (IJDKP), 5*(1), 1–14.

Joshi, A., & Mehta, A. (2018). Analisis teknik K-nearest neighbour untuk kanker payudara klasifikasi penyakit. *Jurnal Internasional Penelitian Ilmiah Terkini, 9*(4), 26126–26130.

Khatieb, N., Usman, M. (2017). Sistem prediksi penyakit jantung yang efisien menggunakan K-nearest teknik klasifikasi tetangga. Dalam Prosiding BDIOT2017 prosiding konferensi internasional tentang big data dan internet of thing, London, Inggris (hlm. 21–26).

Lakshmanaprabu, SK, Mohanty, SN, Shankar, K., Arunkumar, N., & Ramirez, G. (2019). Model pembelajaran mendalam yang optimal untuk klasifikasi kanker paru-paru pada gambar CT. *Sistem Komputer Generasi Masa Depan, 92*, 374–382.

Li, J., Usevich, K., & Comon, P. (2018). Algoritme tipe Jacobi yang konvergen secara global untuk diagonalisasi tensor simetris ortogonal simultan. *Jurnal SIAM Aplikasi Analisis Matriks, 39*(1), 1–22.

Lynch, CM, Abdollahi, B., Fuqua, JD, de Carlo, AR, Bartholomai, JA, Balgemann, RN,...Frieboes, HB (2017). Prediksi kelangsungan hidup pasien kanker paru-paru melalui teknik klasifikasi pembelajaran mesin yang diawasi. *Jurnal Internasional Informatika Medis, 108*, 1–8.

Melamed, ID, Green, R., & Turian, J. (2003). Presisi dan penarikan kembali mesin terjemahan. *Makalah disajikan pada volume pendamping prosiding makalah pendek HLT-NAACL 2003*.

Naresh, P., & Shettar, R. (2014). Teknik pengolahan dan klasifikasi citra awal deteksi kanker paru-paru untuk perawatan kesehatan preventif: Sebuah survei. *Jurnal Internasional tentang Tren Terkini dalam Teknologi Rekayasa, 11*(1), 595–601.

Odajima, K., Pawlovsky, AP (2014). Penjelasan rinci tentang penggunaan KNN metode diagnosis kanker payudara. Dalam Prosiding konferensi internasional ke-7 tentang teknik biomedis dan informatika (BMEI), Dalian, China. Pawlovsky, AP, & Hiroki, M. (2017). Sebuah metode kNN untuk prognosis kanker payudara itu menggunakan algoritma genetika untuk pemilihan komponen. *Metode, 13*, 181–186.

Pradeep, KR, & Naveen, NC (2018). Prediksi kelangsungan hidup kanker paru berdasarkan kinerja menggunakan teknik klasifikasi mesin vektor dukungan, C4. 5 dan algoritma Bayes yang naif untuk analisis layanan kesehatan. *Ilmu Komputer Procedia, 132*, 412–420.

Sa'di, S., Maleki, A., Hashemi, R., Panbechi, Z., Chalabi, K. (2015). Perbandingan data algoritma penambangan dalam diagnosis diabetes tipe II. Jurnal Internasional tentang Aplikasi Ilmu Komputasi 5(5), 1–12.

Septiani, NWP, Wulan, R., & Lestari, M. (2017). Deteksi kanker payudara menggunakan data metode klasifikasi pertambangan. *Jurnal Matematika, 1*(1), 185–191. Tayeb, S., Pirouz, M., Sun, J., Hall, K., Chang, A., Li, J.,...Sager, T. (2017). Ke arah memprediksi kondisi medis menggunakan k-tetangga terdekat. *Prosiding Konferensi Internasional IEEE 2017 tentang Big Data (Big Data), Boston, MA, USA*.

Thun, MJ, Hannan, LM, Adams-Campbell, LL, Boffetta, P., Buring, JE, Feskanich, D.,...Samet, JM (2008). Kejadian kanker paru-paru pada mereka yang tidak pernah merokok: Analisis terhadap 13 kohort dan 22 studi registrasi kanker. *Kedokteran PLoS, 5*(9), 1357–1371.

Wutsqa, DU, Mandadara, HLR (2017). Klasifikasi kanker paru menggunakan dasar radial model jaringan saraf fungsi dengan operasi titik. Dalam Makalah yang dipresentasikan pada pengolahan gambar dan sinyal, Teknik BioMedis dan Informatika (CISP-BMEI), kongres internasional ke-10 tahun 2017.