Jurnal Pra-bukti

EKNN: Pengklasifikasi Ensemble yang Menggabungkan Konektivitas dan Kepadatan ke dalam ***k***NN dengan Aplikasi untuk Diagnosis Kanker<!–<ForCover>Mahfouz MA, Shoukry A, Ismail MA, EKNN: Pengklasifikasi Ensemble yang Menggabungkan Konektivitas dan Kepadatan ke dalam***k***NN dengan Aplikasi untuk Diagnosis Kanker, ***Kecerdasan Buatan Dalam Kedokteran***, doi: 10.1016/j.artmed.2020.101985</ForCover>–>

Mohamed A. Mahfouz, Amin Shoukry, Mohamed A. Ismail

PII: S0933-3657(20)31250-1

DOI: https://doi.org/10.1016/j.artmed.2020.101985

Referensi: SENI 101985

Untuk tampil di: ***Kecerdasan Buatan Dalam Kedokteran***

Tanggal diterima: 14 Desember 2019

Tanggal Revisi: 2 November 2020

Tanggal Diterima: 2 November 2020

Silakan kutip artikel ini sebagai:¿doi:https://doi.org/

EKNN: Pengklasifikasi Ensemble yang Menggabungkan Konektivitas dan Kepadatan ke dalam *k*NN dengan Aplikasi untuk Diagnosis Kanker

Muhammad A. Mahfouz*[1], Amin Shoukry[1, 2], Muhammad A.Ismail[1]

[1]*Departemen Teknik Komputer dan Sistem, Fakultas Teknik, Universitas Alexandria, Mesir*

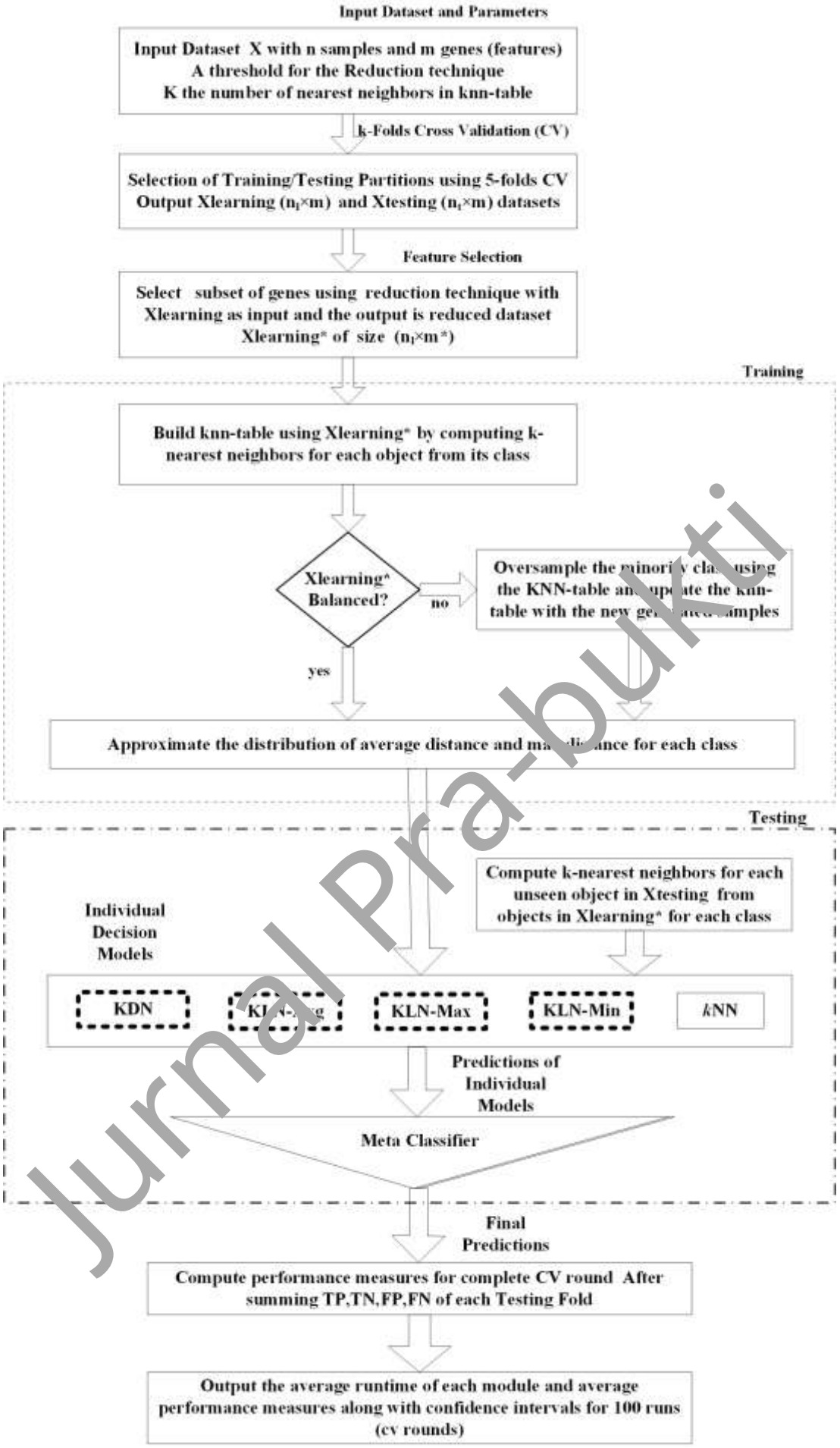
[2]*Departemen Ilmu dan Teknik Komputer, Mesir Jepang, Universitas Sains dan Teknologi, Alexandria, Mesir*

[1]*Email: mamahfouz@gmail.com ,drmaismail@gmail.com*

[2]*Email: amin.shukry@ejust.edu.eg*

Abstrak grafis

1- Dalam makalah ini parameter K dan *k* memainkan peran yang berbeda (lihat Tabel 1).



Highlight

- Studi penelitian ini mengusulkan pengklasifikasi ansambel yang mencakup empat pengklasifikasi dasar relatif terhadap pengklasifikasi tradisional *k*NN.
- Model keputusan yang diusulkan didasarkan pada distribusi kepadatan, keterkaitan tunggal, keterkaitan rata-rata, dan keterkaitan lengkap dari kumpulan data yang diteliti.
- Estimasi distribusi rata-rata keterkaitan dan keterkaitan lengkap masing-masing kelas terbukti lebih tersebar dibandingkan pendekatan terkait lainnya dalam hal koefisien Bhattacharyya.
- Eksperimen kami menunjukkan bahwa model keputusan yang diusulkan memiliki opini yang tidak berkorelasi. Semakin kecil kumpulan data, semakin tinggi peningkatan kinerjanya *k*NN.
- Penggunaan tabel K terdekat mengurangi waktu klasifikasi yang dibutuhkan oleh model keputusan berbasis kepadatan
- EKNN telah diuji pada lima dataset kanker dan tujuh dataset standar lainnya. Peningkatan kinerja telah dicapai dibandingkan dengan beberapa hasil yang dilaporkan dari teknik deteksi kanker berbasis gen.
- Ansambel yang diusulkan menargetkan kumpulan data kecil namun dapat sedikit dimodifikasi agar berfungsi pada kumpulan data besar atau aliran data.
- Ansambel yang diusulkan dapat berfungsi sebagai pemilih gen dengan algoritma genetika *k*NN dalam beberapa karya terkait.

Abstrak

Dalam pendekatan berbasis microarray untuk diagnosis kanker otomatis, penerapannya tradisional *k*-tetangga terdekat *k* Algoritma NN dapat mengalami beberapa kesulitan seperti jumlah gen yang besar (dimensi ruang fitur yang tinggi) dengan banyak gen yang tidak relevan (noise) dibandingkan dengan sedikitnya jumlah sampel yang tersedia dan ketidakseimbangan ukuran sampel kelas target. Penelitian ini memberikan pengklasifikasi ansambel berdasarkan model keputusan yang diturunkan *k*NN yang dapat diterapkan pada permasalahan yang ditandai dengan kumpulan data berukuran kecil yang tidak seimbang. Metode klasifikasi yang diusulkan merupakan gabungan dari metode tradisional *k*Algoritma NN dan empat model klasifikasi baru yang diturunkan darinya. Model yang diusulkan memanfaatkan peningkatan kepadatan dan konektivitas menggunakan K_1-tabel tetangga terdekat (tabel KNN) dibuat selama fase pelatihan. Dalam model kepadatan, sampel tidak terlihat *kamu* diklasifikasikan sebagai milik kelas t jika mencapai peningkatan kepadatan tertinggi ketika sampel ini ditambahkan ke dalamnya, yaitu sampel yang tidak terlihat dapat menggantikan lebih banyak tetangga di tabel KNN untuk sampel kelas t dibandingkan kelas lainnya. Dalam tiga model konektivitas lainnya, mean dan deviasi standar distribusi rata-rata, jarak minimum dan maksimum ke K tetangga anggota masing-masing kelas dihitung dalam fase pelatihan. Kelas t ke mana *kamu* mencapai kemungkinan kepemilikan tertinggi pada distribusinya yang dipilih, yaitu penambahan *kamu* terhadap sampel kelas ini menghasilkan perubahan paling kecil terhadap distribusi model keputusan yang sesuai untuk kelas t. Menggabungkan hasil prediksi dari empat model individu dengan model tradisional *k*NN membuat ruang pengambilan keputusan menjadi lebih diskriminatif.

Dengan bantuan tabel KNN yang dapat diperbarui secara online pada tahap pelatihan, peningkatan kinerja telah dicapai dibandingkan dengan tradisional *k*Algoritma NN dengan sedikit peningkatan waktu klasifikasi. Metode ansambel yang diusulkan mencapai peningkatan akurasi yang signifikan dibandingkan dengan akurasi yang dicapai menggunakan pengklasifikasi dasarnya pada kumpulan data Kentridge, GDS3257, Notterman, Leukemia, dan CNS. Metode tersebut juga dibandingkan dengan beberapa metode ansambel yang ada dan teknik mutakhir dengan menggunakan teknik reduksi dimensi yang berbeda pada beberapa dataset standar. Hasilnya membuktikan keunggulan EKNN dibandingkan beberapa pengklasifikasi individu dan ansambel terlepas dari pilihan strategi seleksi gen.

Kata kunci: Diagnosis Kanker, Klasifikasi Ensemble, Analisis Ekspresi Gen, Tetangga Terdekat.

1. Perkenalan

Tingginya biaya pengumpulan data pasien dan kompleksitas eksperimen seringkali membuat studi penelitian yang berasal dari satu pusat kesehatan memiliki ukuran sampel yang terbatas (dataset kecil) seperti pada diagnosis kanker berbasis microarray. Kumpulan data kecil juga ada di banyak penelitian lain dalam pembelajaran mesin seperti pemrosesan citra medis, diagnosis dan prognosis penyakit, serta prediksi hasil. Kumpulan data kecil seperti itu biasanya tidak seimbang dan berdimensi tinggi [1]. Sebagai pengklasifikasi nonparametrik, *k*NN sangat dipengaruhi oleh outlier dalam kumpulan data kecil. Fokus dari pekerjaan ini adalah pada pengembangan pengklasifikasi ansambel berdasarkan tradisional *k*NN dan model keputusan individu terkait dengan *k*NN yang dapat diterapkan pada masalah yang ditandai dengan kumpulan data kecil yang tidak seimbang dan berisik tanpa menambah waktu klasifikasi.

Itu *k*-tetangga terdekat (*k*Algoritma NN) adalah algoritma pembelajaran berbasis instance yang sederhana. Diberikan sampel yang tidak terlihat *kamu*, algoritme menemukan *k* sampel terdekat dengan *kamu* menurut metrik jarak, lalu *kamu* ditugaskan ke kelas yang memiliki jumlah sampel maksimum di *k* lingkungan terdekat dari *kamu* (pemungutan suara pluralitas). Dari definisi tersebut jelas bahwa *k*NN mudah diimplementasikan, dapat menangani kumpulan data yang memiliki lebih dari dua kelas dan tidak menerapkan asumsi pada distribusi data masukan. Namun, tradisional *k*NN tidak memiliki skalabilitas untuk mengelola kumpulan data yang sangat besar. Juga *k*Aturan NN perlu menemukan identifikasi *k*-lingkungan dan ketika diterapkan pada data yang tidak seimbang, ada kecenderungan untuk menetapkan sampel yang tidak dikenal ke label kelas mayoritas. Selain itu, diketahui bahwa aturan pluralitas menjadi kurang optimal jika jumlah label banyak dan jumlah contoh sedikit. Banyak penelitian yang mencoba mengatasi permasalahan tradisional yang dihadapi di atas *k*Pengklasifikasi NN [2]-[3, 4].

Dalam makalah ini, sebuah ansambel dari *k*-Model keputusan berbasis tetangga terdekat (EKNN) diusulkan dan diterapkan pada diagnosis kanker menggunakan data ekspresi gen. Empat model keputusan baru yaitu KDN, KLN-Avg, KLN-Min dan KLN-Max diusulkan selain model tradisional *k*NN. Dalam fase pelatihan, tabel KNN yang berisi daftar K tetangga terdekat, untuk setiap sampel, dipelihara dan statistik berdasarkan tabel tersebut dihitung untuk setiap kelas. Pada tahap pengujian, statistik yang dihitung dan informasi yang disimpan dalam tabel KNN digunakan untuk menghitung keputusan model yang diusulkan.

Studi eksperimental kami menunjukkan bahwa ansambel EKNN yang diusulkan mengungguli ansambel tradisional *k*pengklasifikasi NN. Ia mampu membagi sampel secara akurat ke dalam kelasnya masing-masing dengan mempertimbangkan gen yang dipilih. Sedangkan skema yang diusulkan meningkatkan kompleksitas fase pelatihan dibandingkan dengan skema tradisional *k*NN, kompleksitas tahap pengujian sebanding dengan tradisional *k*NN karena penggunaan tabel KNN. Waktu pelatihan dan klasifikasi model keputusan yang diusulkan dan tradisional *k*NN dapat dikurangi secara signifikan menggunakan struktur data pohon kd. Tabel KNN juga dapat membantu menangani data yang tidak seimbang melalui up-sampling. Hasil penelitian menunjukkan bahwa keputusan gabungan dari berbagai model ternyata lebih baik dibandingkan dengan model keputusan individual. Kami telah memvalidasi secara eksperimental bahwa EKNN kurang sensitif terhadap parameter input dan menangani noise lebih baik dibandingkan pengklasifikasi dasar tunggal yang sangat baik seperti yang diklaim dalam [5]. Semakin tinggi keragaman di antara pengklasifikasi dasar, semakin baik kinerjanya [6]. Keragaman antara model keputusan yang diusulkan tampaknya bergantung pada data, namun percobaan kami menunjukkan bahwa model keputusan yang diusulkan memiliki pendapat yang tidak berkorelasi terutama ketika kinerja *k*NN terdegradasi jika kumpulan datanya kecil. Jarak Bhattacharyya [7] digunakan dalam penelitian ini untuk mendapatkan gambaran jarak antara asumsi distribusi normal. Pedoman teori pembelajaran ansambel masih terbatas karena kompleksitasnya [8].

Dalam penelitian ini, kinerja algoritma yang diusulkan pada beberapa dataset standar dan kanker dianalisis. Kinerja EKNN pada lima dataset kanker dibandingkan dengan jaringan saraf [9], pohon keputusan [10], naïve Bayes [11], hutan acak, bagging dan adaBoost dengan *k*NN dan pohon keputusan (DT) sebagai pengklasifikasi dasar, selain tradisional *k*NN [12]. Selain itu, EKNN dibandingkan dengan dua pendekatan terbaru lainnya untuk diagnosis kanker TC-VGC [13] dan RBG-CD [14].

Sisa dari makalah ini disusun sebagai berikut. Bagian 2 membahas pekerjaan terkait. Bagian 3 menjelaskan notasi yang digunakan dalam makalah serta beberapa materi latar belakang. Bagian 4 membahas skema yang diusulkan secara rinci. Di bagian 5, hasil eksperimen dibahas. Terakhir, Bagian 6 menyimpulkan makalah ini dan menyoroti arah penelitian di masa depan.

2. Pekerjaan Terkait

2.1. Kerabat KNN

Itu*k*Aturan klasifikasi NN, yang merupakan suara mayoritas sederhana, mengabaikan perbedaan antara kualitas sampel pelatihan dan memperlakukan semua tetangga terdekat secara setara. Untuk mengatasi masalah ini, beberapa metode pemungutan suara jarak jauh telah diusulkan. Dalam [2], algoritma yang disebut sebagai (WKNN) telah dikembangkan, dimana bobot yang lebih besar diberikan kepada tetangga terdekat yang lebih dekat. Tidak jelas*k*Algoritma NN [15] mengeksploitasi ketidakpastian fuzzy untuk meningkatkan kinerja tradisional*k*Algoritma NN, namun ia mengumpulkan buktinya dari*k*tetangga terdekat yang melakukan penyetelan*k*membosankan. Algoritma aturan keputusan fuzzy (FRNN) [16] menggabungkan gagasan informasi lokal dan pendekatan kemungkinan untuk mencapai kinerja yang lebih baik dalam hal ketahanan dan akurasi. FRNN menghindari masalah pemilihan nilai optimal*k*dengan mempertimbangkan semua pola pelatihan sebagai tetangga dengan derajat berbeda yang berasal dari kekhasan fuzzy dari data pelatihan.

Selain itu, sejumlah aturan klasifikasi tetangga terdekat metrik jarak adaptif dikembangkan dengan menyatakan sampel uji sebagai kombinasi linier dari sampel pelatihan yang dinormalisasi sehingga memberikan sampel yang lebih jauh lebih penting [17]. Dalam [17], Jumlah tertimbang dari*k*tetangga terdekat yang diperoleh dengan menggunakan metode yang disebut (CFKNN) ini memiliki potensi untuk mendekati sampel uji. Selain itu, pengklasifikasi populer (PNN) [18] adalah pengklasifikasi berbasis rata-rata lokal yang berupaya menemukan tetangga terdekat semu dengan menggabungkan berbagai jarak antara sampel pengujian dan sampelnya.*k*tetangga terdekat di setiap kelas. Dalam [19], sebuah algoritma yang disebut sebagai (MKNN) diusulkan yang mempertimbangkan informasi lingkungan dari sampel pelatihan untuk menilai kedekatannya dengan sampel pengujian dan untuk membantu membuat keputusan klasifikasi yang benar. Juga di [20], algoritma klasifikasi tetangga terdekat yang ditingkatkan diusulkan melalui penggabungan informasi lingkungan untuk menunjukkan bahwa tetangga terdekat tradisional bersifat sepihak dan mungkin tidak meyakinkan. Baru-baru ini, di [21],*k*-tetangga terdekat terlebih dahulu dianggap sebagai informasi lingkungan masing-masing sampel, kemudian*k*-tetangga terdekat umum (*k*GNN) dari sampel pengujian ditemukan mengambil keputusan klasifikasi melalui suara terbanyak. Di dalam*k*GNN, informasi lingkungan bersama dari sampel pengujian dan pelatihan dipertimbangkan dan wilayah yang tumpang tindih digunakan untuk memutuskan apakah sampel pelatihan merupakan a*k*GNN dari sampel pengujian. Beberapa metrik jarak telah diadopsi dalam algoritma di atas, dan kinerjanya telah dievaluasi pada [22].

Selain itu,*k*Aturan NN biasanya mengalami outlier [23], [16] dan [4], terutama dalam kasus data pelatihan berukuran kecil. Salah satu solusi untuk masalah ini adalah berbasis rata-rata lokal*k*-pengklasifikasi tetangga terdekat (LMKNN), diusulkan pada [23]. LMKNN menghitung vektor rata-rata lokal dari*k*tetangga terdekat dari sampel pengujian di setiap kelas untuk mencapai ketahanan terhadap outlier dan meningkatkan akurasi klasifikasi. Karya terkait lainnya adalah [24] yang mengusulkan penggantian sistem pemungutan suara terbanyak di*k*NN dengan menggabungkan mean lokal berdasarkan tetangga terdekat (LMKNN) [23] dan bobot jarak*k*-tetangga terdekat (DWKNN) [3] metode. Penggabungan kedua metode ini terbukti mampu meningkatkan akurasi proses klasifikasi. Pada tahap pengujian, LMKNN menggunakan jarak terdekat terhadap setiap vektor rata-rata lokal (pusat tetangga terdekat) dari setiap kelas data, yang dianggap efektif dalam mengatasi pengaruh negatif outlier [21].

Studi yang relevan dengan kontribusi yang diusulkan dalam makalah ini adalah [25]. Dalam [25], secara sederhana*k*Aturan NN disebut MinKL yang memperhitungkan label semua tetangganya, bukan hanya label yang paling umum. MinKL menunjukkan perbaikan *k*NN dalam beberapa kasus terutama ketika jumlah label banyak dan jumlah sampel sedikit. Dalam ansambel yang diusulkan, penggunaan beberapa aturan seperti MinKL membuatnya lebih kuat dan akurat dibandingkan MinKL. Perlu diketahui, serupa dengan usulan EKNN, nilai K pada MinKL, LMKNN, dan DWKNN berbeda dengan nilai K semula.*k*NN, dimana nilai*k*adalah jumlah tetangga terdekat dari seluruh data pelatihan, sedangkan pada pendekatan ini nilai K adalah jumlah tetangga terdekat dari setiap kelas pada data pelatihan [21].

2.2. Algoritma pemilihan fitur dan klasifikasi untuk data microarray

Algoritme yang efisien dan efektif diperlukan untuk mengekstrak dan memilih fitur dari data microarray yang meningkatkan kinerja pengklasifikasi. Selain metode statistik untuk mereduksi dimensi seperti analisis komponen utama (PCA) yang didasarkan pada proyeksi ortogonal [26], ada dua kategori utama teknik yang umum digunakan untuk mereduksi dimensi tinggi data microarray: filter [27] dan wrapper [28]. Metode filter, seperti *T*–test dan signal-noise-rate (SNR), memprioritaskan gen berdasarkan satu atau lebih metrik yang telah ditentukan dan memilih gen dengan peringkat teratas. Metode wrapper juga telah banyak digunakan untuk memilih gen fitur dari data microarray. Mereka menghasilkan subset gen fitur alternatif dan memilih subset dengan akurasi klasifikasi tertinggi. Algoritma genetika (GA) digabungkan dengan*k*NN di [29] untuk mengidentifikasi gen diskriminatif.

Pendekatan lain untuk diagnosis kanker didasarkan pada komputasi pasangan gen-gen yang membedakan menggunakan koefisien korelasi seperti pada algoritma TC-VGC [13] atau dengan menghitung bicluster yang membedakan seperti pada algoritma RBG-CD [14]. Baik TC-VGC maupun RBG-CD sensitif terhadap parameter masukannya dan memerlukan penyetelan parameter ekstensif namun tidak memerlukan langkah reduksi dimensi. Kesimpulan umumnya adalah tidak ada pengklasifikasi tunggal yang unggul dalam semua jenis kumpulan data dan langkah reduksi dimensi memainkan peran penting dalam diagnosis kanker.

Jaringan saraf konvolusional (CNN) adalah arsitektur NN yang populer untuk pembelajaran mendalam dan sering digunakan untuk ekstraksi fitur dan klasifikasi [30], [31], [32], [33], [34]. Kemampuan CNN untuk mempelajari karakteristik data masukan mengurangi pemrosesan yang diperlukan dibandingkan dengan algoritma klasifikasi lainnya. Namun, CNN menderita biaya komputasi yang tinggi

dan membutuhkan banyak data pelatihan [32]. Ketika jumlah kelas besar dan ukuran data pelatihan terlalu kecil, CNN sulit dilatih dengan presisi tinggi.

Dalam model deteksi kebocoran yang diusulkan pada [32], clustering digunakan untuk mengurangi jumlah kelas dengan mengelompokkan pipa berdasarkan tekanan dan aliran. Nomor cluster saluran pipa digunakan sebagai label kategori.

2.3. Metode Ensemble Terkait

Beberapa Ensemble dalam literatur dibuat dan disesuaikan untuk menangani data microarray. Dalam [35], beberapa kernel Support Vector Machines (MK-SVM) digunakan untuk mengubah masalah pemilihan fitur menjadi masalah pembelajaran beberapa parameter kemudian algoritma seperti pohon digunakan untuk mengekstraksi aturan klasifikasi dari vektor dukungan yang diperoleh. Dalam [36], pengklasifikasi ansambel dibangun melalui pembelajaran kernel SVM yang berbeda, seperti linier, polinomial, fungsi basis radial (RBF), dan sigmoid. Hasil prediksi dari masing-masing model digabungkan melalui pemungutan suara mayoritas. Dalam [37], ansambel pengklasifikasi pohon keputusan dibangun pada subset data yang berbeda di mana transformasi KPCA komponen utama kernel diterapkan pada masing-masing subset ini. Penggunaan KPCA yang telah disesuaikan (dengan RBF sebagai kernel) memungkinkan ansambel yang dihasilkan menangani data yang tidak dapat dipisahkan secara linier dan mengungguli pengklasifikasi Random Forest (RF) yang canggih dan beberapa metode lain yang ada pada data eksperimennya.

Pendekatan ansambel terbukti berguna dalam menangani kumpulan data kecil seperti pada [1], di mana sejumlah besar jaringan saraf identik dalam hal topologi dan fungsi neuron digunakan untuk prediksi. Berbeda dengan pendekatan ansambel kami, contoh desain jaringan saraf optimal dengan kinerja tertinggi di [1] dipilih sebagai model kerja. Dibandingkan dengan ansambel di atas, EKNN sederhana.

2.4. Metode untuk mengatasi input data yang tidak seimbang

Kinerja dari *k*NN dipengaruhi oleh adanya ketimpangan data masukan karena kelas mayoritas cenderung memiliki akurasi klasifikasi yang tinggi, sedangkan kelas minoritas cenderung memiliki akurasi klasifikasi yang rendah. Beberapa metode yang ada mengandalkan penyesuaian skor sampel yang tidak terlihat pada tahap pengujian dengan memberikan bobot yang lebih tinggi untuk kelas minoritas [38]. Namun, pendekatan ini sangat sensitif terhadap parameter masukannya. Pendekatan lainnya adalah dengan melakukan resize pada input dataset, resize dilakukan dengan cara melakukan under-sampling pada kelas mayoritas atau over-sampling pada kelas minoritas atau dengan melakukan keduanya secara bersamaan (hybrid) [39] hingga datanya seimbang. Strategi yang mencapai kinerja tertinggi dalam [40] adalah dengan mengambil sampel sampel kelas negatif (mayoritas) yang mendekati semua sampel positif. Perlu dicatat bahwa pengklasifikasi ansambel pohon seperti hutan acak dapat bekerja dengan baik pada data yang tidak seimbang tanpa pengambilan sampel naik atau turun.

Pendekatan lain untuk menangani data yang tidak seimbang adalah penggunaan metrik kinerja yang sesuai. TPR-TNR tertimbang [41] adalah metrik bernilai tunggal untuk menilai kinerja pengklasifikasi yang memperhitungkan kumpulan data yang tidak seimbang. Hasil eksperimen yang diberikan dalam [41] menunjukkan bahwa TPR-TNR mampu mengklasifikasikan pengklasifikasi lebih akurat dibandingkan metode yang merumuskan masalah pemeringkatan sebagai masalah pengambilan keputusan yang harus diselesaikan dengan menggunakan metode MCDM mana pun di mana pengklasifikasi adalah alternatifnya dan metrik yang tersedia adalah metrik yang tersedia. kriteria [42]. Penggunaan metrik tertimbang terbukti dapat diandalkan dalam kasus data yang tidak seimbang atau ketika kesalahan klasifikasi terkait biaya berbeda. Selain itu, algoritme pembelajaran yang diberi sanksi (atau pelatihan yang sensitif terhadap biaya) yang meningkatkan biaya kesalahan klasifikasi pada kelas minoritas [37] mampu memberikan hasil yang lebih andal.

3. PENDAHULUAN

Bagian ini memberikan materi pendukung untuk membantu pembaca lebih memahami bagian selanjutnya. Singkatan dan simbol yang umum digunakan tercantum pada Tabel 1.

Tabel 1 Singkatan dan simbol yang digunakan dalam teks

Singkatan	Keterangan
*k*NN	Algoritma k-tetangga terdekat tradisional
KDN	model keputusan yang diusulkan berdasarkan Kepadatan
KLN-Rata-rata	model keputusan yang diusulkan berdasarkan keterkaitan rata-rata
KLN-Maks	model keputusan yang diusulkan berdasarkan hubungan lengkap
KLN-Min	model keputusan yang diusulkan berdasarkan hubungan tunggal
EKNN	Skema Klasifikasi ansambel yang diusulkan
X	kumpulan data yang terdiri dari*n sampel*×*m gen*
T	vektor label target t=(T_1, T_2,..., T_N) sesuai dengan*N* sampel di**X**;
X_T	kumpulan data yang terdiri dari sampel kelas t di**X**
N_T	jumlah sampel masuk$\mathbf{X}_T$
N	jumlah sampel seluruhnya
C	jumlah total kelas (2 untuk kelas biner)
$Aku\ g_{Saya,\ Gj}$	korelasi Pearson absolut antara vektor fitur yang mewakili ekspresi genG_{Saya}DanG_j
$D(S_{Saya,\ Sj})$	Koefisien Korelasi Pearson Jarak antar sampel S_{Saya}DanS_j
$x_{aku,j}$	j^{th}nilai fitur dari*Saya*thsampel dalam kumpulan data X
Tabel KNN	Tabel ukuran*n*×*k*sel, berisi indeks K tetangga terdekat dan jaraknya untuk setiap sampel
k	jumlah tetangga terdekat yang digunakan*kNN*
K	jumlah tetangga terdekat dari setiap kelas yang disimpan dalam tabel KNN untuk EKNN
$N_T(X)$	K tetangga terdekat sampel*X*milik kelas*T*
α_{Saya}	Parameter KLN-Avg untuk menghitung skor kelas*Saya*
β_{Saya}	Parameter KLN-Min untuk menghitung skor kelas*Saya*
η_{Saya}	Parameter KLN-Max untuk menghitung skor kelas*Saya*
δ	ambang batas yang digunakan dalam model KDN untuk memungkinkan penambahan skor untuk sampel terdekat tetapi tidak terdekat

3.1. Kesamaan antara dua distribusi probabilit

Jarak Bhattacharyya [7] digunakan dalam penelitian ini untuk mengukur jarak antara dua distribusi probabilitas yang berbeda. Jarak antara dua distribusi normal $N_P(\mu_P, \sigma_P)$ dan $N_Q(\mu_Q, \sigma_Q)$ dihitung sebagai berikut:

$$= \frac{1}{4}\left(\frac{2+}{2} \quad \frac{2}{2} + 2\right) + \quad \frac{1}{4}\left(\frac{(\quad - \quad)}{2+\ 2}\right)^2 \qquad (1)$$

Dimana μ_P, $\sigma_P$daμ_Q, dadalah mean dan deviasi standar dari*P*-th dan*Q*-distribusi normal ke-th, masing-masing. Jarak Bhattacharyya diadopsi untuk mengukur kesamaan antara estimasi distribusi setiap kelas menggunakan KLN-Avg, KLN-M dan KLN-Max untuk pilihan K yang berbeda. Jarak yang diukur dibandingkan dengan jarak antara distribusi probabilitas diskrit setiap kelas oleh algoritma MinKL [25].

Jarak Bhattacharyya antara dua distribusi probabilitas diskrit*P*Dan*Q*pada domain yang sama X dihitung sebagai berikut:

$$= - \quad \left(\sum_{\in} \sqrt{(\ )\ (\ )}\right) \qquad (2)$$

Jarak d terletak pada [0,∞] untuk kasus diskrit dan kontinu.

3.2. Kumpulan Data Ekspresi Gen

Ekspresi gen diukur menggunakan teknologi DNA microarray dengan membandingkan ekspresi gen suatu sel yang dihambat pada kondisi tertentu (sampel c) dengan ekspresi gen sel lain yang dipertahankan pada kondisi normal (sampel n) [10]. Dalam diagnosis kanker, setiap sampel dalam kumpulan data masukan diberi label sebagai ganas atau normal dalam klasifikasi biner atau diberi tingkatan kanker. Satu ekspresi gen berarti satu nilai fitur dalam vektor fitur. Demikian pula, satu kolom fitur memiliki banyak ekspresi gen dari satu gen pada semua sampel. Entri matriks ekspresi gen adalah bilangan real yang mewakili ekspresi gen setiap sampel.

Tabel 2 Kumpulan Data Ekspresi Gen

Himpunan data Kanker	Ref.	jenis	normal/ ganas	total sampel	TIDAK. fitur
Kentridge (Usus besar)	[43]	tidak seimbang	22/40	62	2000
GDS3257 (Paru-paru)	[44]	tidak seimbang	58/49	107	2517
Notterman (Usus besar)	[45]	seimbang	18/18	36	7457
Leukemia (Darah)	[46]	tidak seimbang	25/47	72	7129
SSP (Grogi)	[47]	tidak seimbang	21/39	60	7110

Kesulitan utama dalam analisis jenis data ini adalah banyaknya jumlah gen (dimensi ruang fitur yang tinggi) dengan banyak gen yang tidak relevan (noise) dibandingkan dengan jumlah sampel yang sangat sedikit. Selain itu, kumpulan datanya biasanya tidak seimbang, yaitu kelas mayoritas cenderung memiliki sampel yang jauh lebih banyak. Contoh data tersebut adalah kumpulan data yang tercantum pada Tabel 2 yang digunakan dalam penelitian eksperimental kami. Untuk kumpulan data yang sangat kecil ini, skema validasi tradisional di mana data dibagi menjadi kumpulan pelatihan dan pengujian tetap tidak praktis karena tidak ada cukup data untuk dipartisi. Bagian 3.4 membahas validasi silang sebagai solusi untuk masalah ini. Selain itu, kumpulan data Kentridge, Leukemia, dan SSP sedikit tidak seimbang. Seperti yang ditunjukkan pada Tabel 2, semua kumpulan data ekspresi gen yang terdaftar memiliki dimensi tinggi yang berkisar antara 2000 hingga 7457 untuk kumpulan data Kentridge dan Notterman. Dataset asli GDS3257 (Lung) telah diproses menggunakan *Babelomics* alat [ ] dan setelah langkah penyaringan, terdiri dari 2517 gen sebelum menerapkan langkah reduksi dimensi.

3.3. Kesamaan antara dua sampel

Jarak Euclidean, blok kota, kosinus, dan korelasi adalah metrik yang umum digunakan dengan *k* Algoritma NN [22]. Ukuran kesamaan yang paling umum dalam bidang bioinformatika adalah yang didasarkan pada koefisien korelasi. Kebanyakan dari mereka memiliki ukuran ketidaksamaan yang sesuai. Koefisien korelasi Pearson dari dua variabel acak x dan y secara formal didefinisikan sebagai berikut:

$$s(X,kamu) = \frac{1}{N-1}\sum_{Saya=1}^{N}\left(\frac{X_{Saya}-\bar{X}}{\sigma_X}\right)\left(\frac{kamu_{Saya}-\bar{kamu}}{\sigma_{kamu}}\right) \quad (3)$$

Di mana $X, kamu$ adalah maksud dari X_{Saya} Dan $kamu_{Saya}$, Dan σ_X, σ_{kamu} adalah standar deviasinya masing-masing. Ini adalah ukuran seberapa baik sebuah garis lurus dapat dipasangkan pada plot sebar X Dan $kamu$.

Koefisien korelasi Pearson absolut digunakan sebagai metrik kesamaan antara pasangan gen pada langkah reduksi dimensi dan jarak Pearson yang sesuai, diberikan dalam Persamaan. (4), mengukur jarak antara pasangan sampel dalam langkah pelatihan dan klasifikasi.

$$(,) = 1 - \quad ((,)) \quad (4)$$

Ketika koefisien korelasi Pearson turun menjadi [-1, 1], jarak Pearson dan korelasi absolut terletak pada [0, 1].

3.4. Validasi Silang (CV)

Saat memecahkan masalah prediksi, model prediksi yang dihasilkan harus dapat digeneralisasikan ke kumpulan data independen (sampel pengujian tidak diketahui), yaitu menggeneralisasi secara akurat dalam praktik. Model prediksi biasanya divalidasi dengan mempartisi data masukan menjadi dua partisi (misalnya 70% untuk pelatihan dan 30% untuk pengujian). Namun, dalam masalah kita, jumlah sampel sangat kecil dan tidak ada cukup sampel untuk dipartisi tanpa kehilangan kemampuan pemodelan dan pengujian yang signifikan. Dalam karya ini, skema Cross-Validation (CV) k-folds [49] diterapkan dalam menguji model yang diusulkan. Kumpulan data masukan dibagi menjadi f partisi (lipatan), partisi f -1 berpartisipasi dalam pelatihan, dan kelas instance

milik partisi yang tersisa diprediksi oleh model keputusan berdasarkan pelatihan yang dilakukan pada partisi pelatihan f-1. Proses ini diulangi sebanyak k kali untuk membentuk putaran CV lengkap setelah itu kelas setiap sampel diidentifikasi.

3.5. Metrik Evaluasi Kinerja

Efektivitas EKNN telah dievaluasi menggunakan metrik kinerja terkenal seperti akurasi, sensitivitas, dan spesifisitas [49] yang melibatkan TP dan TN, masing-masing jumlah sampel positif dan negatif yang diklasifikasikan dengan benar, serta FP dan FN, jumlah sampel positif dan negatif yang salah diklasifikasikan.

Akurasi mengukur efektivitas skema klasifikasi secara keseluruhan. Ini dihitung sebagai berikut:

$$= \frac{+}{+ \quad + \quad +} \tag{5}$$

Sensitivitas mengukur kemampuan pengklasifikasi untuk mengenali pola dari kelas positif. Itu dapat diperoleh dengan menggunakan persamaan berikut:

$$= \frac{\quad}{+} \tag{6}$$

Spesifisitas mengukur kemampuan pengklasifikasi untuk mengenali pola dari kelas negatif. Persamaan berikut digunakan untuk menghitung spesifisitas:

$$= \frac{\quad}{+} \tag{7}$$

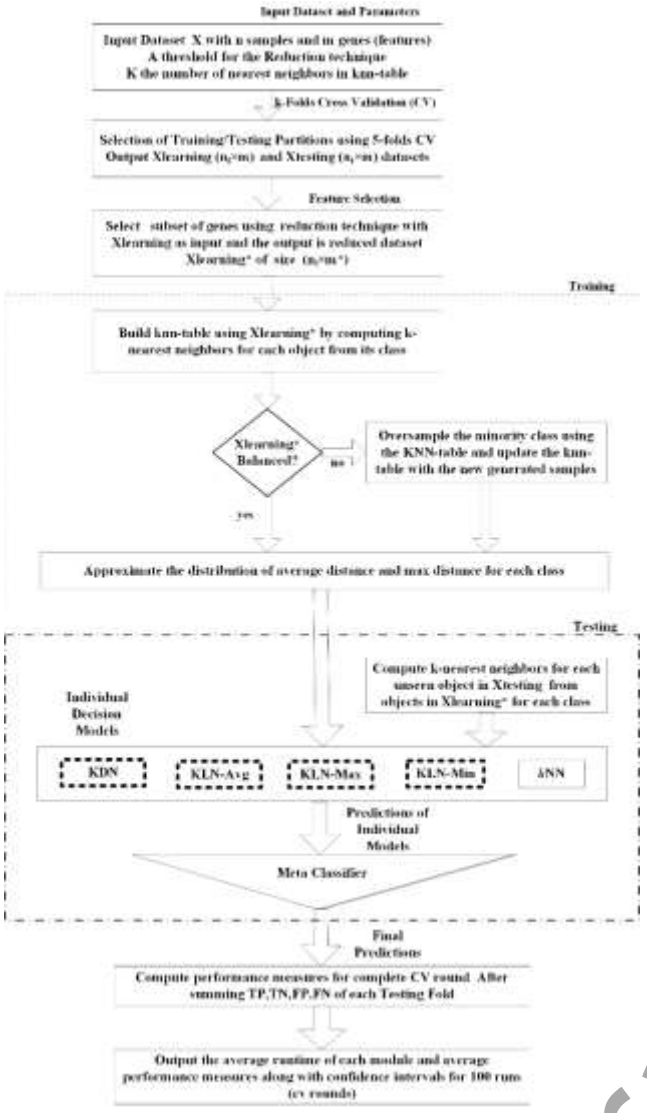
Dalam diagnosis kanker, sensitivitas lebih penting daripada spesifisitas karena sensitivitas menunjukkan sebarapa besar pengklasifikasi mampu mengidentifikasi dengan tepat pasien kanker yang mungkin dapat diobati saat ini, namun tidak dapat diobati di kemudian hari (misalnya kanker serviks). Secara umum, sensitivitas dan spesifisitas yang seimbang dapat diterima. AUC (Area di bawah kurva ROC) adalah metrik kinerja yang menggabungkan sensitivitas dan spesifisitas. (1 - Spesifisitas) vs. Sensitivitas diplot dan area di bawah kurva dibandingkan dengan area di bawah kurva ROC lainnya. AUC antara 0 dan 1, semakin tinggi AUC, semakin baik kinerjanya. (1-Spesifikasi) mengacu pada tingkat positif palsu. Meskipun akurasi dihitung berdasarkan ambang batas tertentu, AUC dapat dihitung untuk beberapa ambang batas pada waktu yang bersamaan. Pengklasifikasi yang baik harus memiliki tingkat positif palsu yang rendah dan tingkat negatif palsu yang rendah.

Dari pengertian akurasi, akurasi tinggi berarti rendah (FP+FN)/*N*jadi ketika ada perbedaan besar antara biaya kesalahan klasifikasi FP dan FN operasional, atau antara frekuensi kelas operasional dibandingkan dengan yang ada di set pelatihan, maka AUC adalah indikator kinerja yang lebih baik daripada akurasi. Selain itu, kami menggunakan skor akurasi seimbang yang tersedia di [50]. Ini didefinisikan sebagai rata-rata perolehan kembali yang diperoleh pada setiap kelas. Hal ini lebih bermakna dalam kasus kumpulan data yang tidak seimbang.

Dalam studi eksperimental kami, ukuran kinerja yang dilaporkan adalah rata-rata selama 30 kali berjalan. Beberapa ukuran kinerja dilaporkan dengan interval kepercayaan 95% untuk menunjukkan keandalan perkiraan kami. Setelah setiap putaran validasi silang 5 kali lipat, TP, TN, FP, dan FN dari semua lipatan dijumlahkan untuk mendapatkan matriks konfusi tunggal, dan metrik kinerja apa pun yang diinginkan dapat dihitung untuk putaran ini.

4. Skema Klasifikasi yang Diusulkan

Skema EKNN yang diusulkan merupakan kombinasi strategi pemilihan fitur ditambah dengan klasifikasi ansambel. Gambar 1 menunjukkan arsitektur tingkat atas EKNN, dan bagian berikut menjelaskan berbagai tahapannya secara rinci. Persegi panjang putus-putus yang tebal, pada gambar. 1, mewakili kontribusi kami.



Gambar 1. Tata Letak Tingkat Atas dari skema yang diusulkan

Berbeda dengan tradisional *k*Algoritma NN di mana tidak ada pekerjaan yang dilakukan dalam fase pelatihan, di EKNN, tabel K-neighbours terdekat (tabel KNN) dibuat selama waktu pelatihan dan statistik yang diperlukan untuk model keputusan yang diusulkan dihitung. Jika kumpulan data masukan tidak seimbang, tabel KNN dapat digunakan untuk mengidentifikasi sampel perbatasan dan melakukan pengambilan sampel berlebihan pada kelas minoritas. Membangun tabel KNN membantu mengurangi kompleksitas komputasi model keputusan yang diusulkan seperti yang dijelaskan di bagian selanjutnya. Empat model keputusan yang diusulkan merupakan kontribusi utama dari penelitian ini. Seperti yang ditunjukkan pada Gambar 1, setelah memilih set pelatihan sebagai Xlearning dan set pengujian sebagai Xtesting menggunakan validasi silang 5 kali lipat, teknik reduksi dimensi diterapkan pada Xlearning untuk menghasilkan set pelatihan tereduksi Xlearning* yang memiliki jumlah sampel yang sama dengan *Belajar tetapi dengan serangkaian fitur yang dikurangi (gen dalam masalah diagnosis kanker). Kemudian KNNtable dibuat menggunakan Xlearning*. Jika Xlearning* seimbang maka pengambilan dapat diterapkan, jika tidak, pengambilan sampel berlebih pada kelas minoritas akan dilakukan di Xlearning*. Selain itu, tabel KNN diperbarui oleh sampel yang baru dibuat. Untuk mengurangi dampak ketidakseimbangan data terhadap keakuratan *k* NN dan model keputusan lain yang diusulkan, kami memilih solusi yang diusulkan di [51], disebut sebagai borderline-SMOTE2, juga diimplementasikan di [50]. Informasi yang diperlukan oleh Borderline-SMOTE2 untuk mengambil sampel kelas minoritas secara berlebihan dapat diambil saat membuat tabel KNN. Dimulai dengan mengidentifikasi sampel perbatasan dari kelas minoritas. Setelah menghitung *k*tetangga terdekat dari setiap sampel di kelas minoritas, sampel tersebut memiliki semuanya *k*tetangga terdekat dari kelas mayoritas dianggap kebisingan. Sampel inti adalah sampel yang mempunyai kurang dari *k*/2 tetangga terdekat dari kelas mayoritas. Sampel yang tersisa dianggap sampel perbatasan. Setelah mengidentifikasi kumpulan sampel perbatasan B, setiap kali ada kebutuhan untuk menghasilkan contoh sintetik, satu sampel acak $B_j$di B dipilih dan selisihnya (*perbedaan*) antara sampel tersebut dan sampel lain yang dipilih secara acak M_{Saya}itu *M*-tetangga terdekat dihitung. Sampel baru yang dihasilkan sama dengan B_{Saya}+*r berbeda*Di mana *R*adalah bilangan acak antara 0 dan 1 jika M_{Saya}milik kelas minoritas, sebaliknya *R*terletak antara 0 dan 0,5. Dengan demikian, contoh-contoh baru yang dihasilkan lebih dekat dengan kelas minoritas. Tabel KNN diperbarui secara bertahap berdasarkan sampel yang baru dihasilkan bersama dengan statistik yang dihitung. Dalam tahap pengujian, *k*- Tetangga terdekat dari sampel yang tidak terlihat dari setiap kelas diidentifikasi. Empat model keputusan yang disebut sebagai KDN, KLN-Min, KLN-Avg dan KLN-Max diusulkan, selain model tradisional. *k* NN. Keputusan akhir dihitung dari model keputusan dasar menggunakan susun. Untuk setiap putaran CV yang lengkap, metrik kinerja dihitung.

4.1. Membangun Tabel K-Nearest-Neighbors (Tabel KNN)

Untuk setiap sampel, K sampel terdekat dari kelasnya diidentifikasi dan diurutkan, dalam urutan menurun, menurut ukuran kemiripan dari yang terdekat hingga yang terjauh. Sampel K terdekat ini dapat diperbarui secara dinamis dengan kedatangan sampel baru untuk memiliki pengklasifikasi online. Untuk setiap sampel baru, kami menyimpan K tetangga terdekatnya dari setiap kelas beserta jaraknya untuk mengurangi kompleksitas komputasi model keputusan KDN yang diusulkan, seperti yang dijelaskan di bagian berikut. Tabel KNN dihitung berdasarkan Persamaan. (4). Untuk memperbarui tabel KNN secara dinamis, ketika sampel pelatihan masuk *kamu* sampai kesamaannya dengan setiap sampel dari kelas t dihitung. Sebuah entri ditambahkan dengan daftar sampel K teratas yang serupa. Jika sampel yang masuk dapat menggantikan salah satu tetangga terdekat dari sampel sebelumnya X_{Saya} kemudian sampel yang masuk ditambahkan ke $N_T(X_{Saya})$ alih-alih sampel yang lebih jauh ini. Namun, agar pengklasifikasi EKNN yang diusulkan dapat belajar dari aliran data, kita juga harus menghitung statistik yang diperlukan secara dinamis. Dengan bantuan tabel KNN, untuk mengklasifikasikan sampel yang tidak terlihat, kita dapat mengidentifikasi tetangga dari tetangga dari sampel yang tidak terlihat untuk dua tingkat atau lebih dan menghitung statistik yang diperlukan berdasarkan beberapa tetangga tersebut. Berurusan dengan data aliran berada di luar cakupan pekerjaan ini.

4.2. Menghitung Statistik yang Diperlukan dari tabel KNN

Biarkan $N_T(X_{Saya})$ adalah himpunan K tetangga terdekat suatu sampel $X_{Saya} \in X_T$. Membiarkan *kdistmin*(X_{Saya}), *kdistmax*(X_{Saya}), *kdistavg*(X_{Saya}) menjadi jarak minimum, maksimum, dan rata-rata dari X_{Saya} kepada anggota $N_T(X_{Saya})$, masing-masing.

Membiarkan $kdistmin_T$, $kdistmax_T$ Dan $kdistavg_T$ menjadi rata-rata jarak minimum, maksimum, dan rata-rata ke K tetangga terdekat untuk setiap kelas *T* masing-masing. Mereka dihitung sebagai rata-rata *kdistmin*(X_{Saya}), *kdistmax*(X_{Saya}) Dan *kdistavg*(X_{Saya}) masing-masing. Rata-rata dari *kdistmin*(X_{Saya}) untuk kelas t (untuk semua $X_{Saya} \in X_T$) didefinisikan sebagai $kdistmin_T$ dan dihitung sebagai berikut:

$$= \frac{\sum \in ( )}{| |} = \sum \in (,)/| | \quad 8)$$

Rata-rata dari *kdistmax*(X_{Saya}) untuk kelas t (untuk semua $X_{Saya} \in X_T$) didefinisikan sebagai $kdistmax_T$ dan dihitung sebagai berikut

$$= \frac{\sum \in ( )}{| |} = \sum \in (,)/| | \quad (9)$$

Rata-rata dari *kdistavg*(X_{Saya}) untuk kelas t (untuk semua $X_{Saya} \in X_T$) didefinisikan sebagai $kdistavg_T$ dan dihitung sebagai berikut $= \sum \in ( )/| | = \sum \in (\sum \in ( ) , /)/| |$ (10)

Berasumsi bahwa *kdistmin*, *kdistmax* Dan *kdistavg* seluruh sampel X_T mengikuti distribusi normal, deviasi standar yang sesuai disebut sebagai $kminsigma_T$ $kmaxsigma_T$ Dan k masing-masing, dapat dihitung sebagai berikut:

$$= \sqrt{\frac{1}{| |-1} \sum \in (( ) -)^2} \quad (11)$$

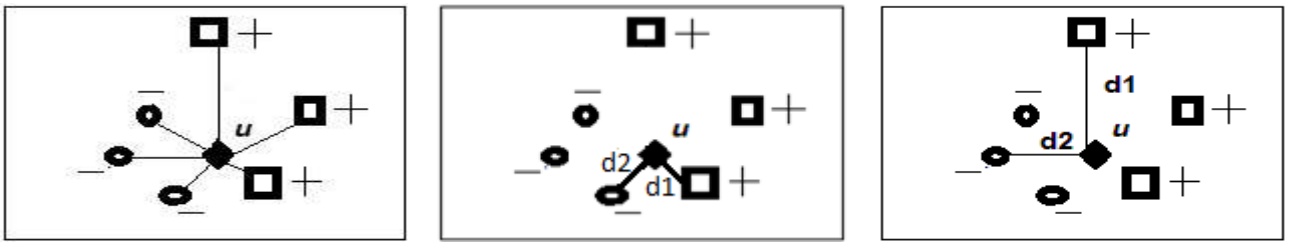
$$= \sqrt{\frac{1}{| |-1} \sum \in (( ))^2} \quad (12)$$

$$= \sqrt{\frac{1}{| |-1} \sum \in (( ) -)^2} \quad (13)$$

Jika *kdistmax*(X_{Saya}), *kdistmin*(X_{Saya}) atau *kdistavg*(X_{Saya}) untuk apa pun $X_{Saya} \in X_T$ mengambil nilai ekstrim (yaitu lebih jauh dari tiga kali standar deviasi yang bersangkutan). X_{Saya} dihapus dari perhitungan $kdistmax_T$, $kdistmin_T$ atau $kdistavg_T$ masing-masing. Proses ini dapat diulangi untuk mengurangi efek outlier pada komputasi. Sisa datanya digunakan untuk menghitung nilai baru $kavgsigma_T$, $kminsigma_T$, $kmaxsigma_T$, $kdistavg_T$, $kdistmin_T$ Dan $kdistmax_T$.

4.3. Model Keputusan

Seperti ditunjukkan pada Gambar 2 a), jarak rata-rata dari *kamu* ke K tetangga terdekatnya dari setiap kelas dipertimbangkan dalam menghitung $kdistavg_{+1}$(*kamu*) dan $kdistavg_{-1}$(*kamu*). Pada Gambar 2 b), d1 dan d2 mewakili $kdistmin_{+1}$(*kamu*), $kdistmin_{-1}$(*kamu*) masing-masing untuk kelas positif dan negatif. Sedangkan pada Gambar 2 c) d1 dan d2 mewakili $kdistmax_{+1}$(*kamu*), $kdistmax_{-1}$(*kamu*) masing-masing untuk kelas positif dan negatif. Meski tetangga terdekat *kamu* termasuk dalam kelas positif, kelas yang akan ditugaskan ke sampel tak terlihat *kamu* oleh model yang diusulkan akan bergantung pada skor yang dihitung untuk setiap kelas menggunakan Persamaan. (15), (17) dan (19), masing-masing.



a) Keterkaitan Rata-rata b) Keterkaitan Tunggal c) Tautan Lengkap

Gambar 2. Jarak yang dihitung untuk sampel u yang tidak terlihat dalam a) KLN-Rata-rata b) KLN-Min dan c) KLN-Max

4.3.1. Tetangga terdekat k tradisional (kNN)

Suatu sampel diklasifikasikan dengan mengidentifikasi sampelnya *k* tetangga, menghitung jumlah tetangga yang termasuk dalam setiap kelas, dan menugaskannya ke kelas yang mayoritas anggotanya *k* milik tetangga. *k* adalah bilangan bulat positif, biasanya kecil. Jika *k*=1, maka sampel tersebut ditugaskan ke kelas tetangga terdekatnya. Dalam masalah klasifikasi biner (dua kelas), memilih adalah hal yang bermakna *k* menjadi angka ganjil untuk menghindari suara seri.

4.3.2. Model Keputusan yang Diusulkan berdasarkan Strategi Tautan Rata-Rata (KLN-Avg)

Untuk mengklasifikasikan sampel yang tidak terlihat *kamu*, pertama, himpunan K tetangga terdekat dari *kamu* dari setiap kelas *T* dilambangkan N_T(*kamu*) dihitung bersama dengan mereka *kdistavg*$_T$*(kamu)* sebagai berikut:

$$( ) = \sum_{\in ( )} \frac{(,)}{} \quad (4)$$

Skor sampel *kamu* untuk kelas *T* dihitung berdasarkan z-score yang dimodifikasi sebagai berikut:

$$\text{Skor rata-rata KLN} = -\propto \left| \frac{\text{kdistavg ( )}-\text{kdistavg |}}{\text{kavgsigma}} \right. \quad (15)$$

Jika KLN-Avg akan digunakan sebagai pengklasifikasi tunggal maka parameter α_T untuk *T*=1 hingga c harus disetel dengan baik. Pada kasus ini *kamu* ditugaskan ke kelas *T* mempunyai skor maksimal. Pendekatan lain adalah dengan mengatur α_T=1 untuk *T*=1 hingga c dan susun pengklasifikasi tunggal ini. Namun, jika KLN-Avg digabungkan dengan model keputusan lain menggunakan penumpukan maka parameter tersebut tidak perlu disesuaikan (yaitu ditetapkan ke 1) dan skor akan dimasukkan sebagaimana adanya ke pengklasifikasi meta.

4.3.3. Usulan model keputusan berdasarkan single linkage (KLN-Min)

Dalam model ini, untuk mengklasifikasikan sampel yang tidak terlihat *kamu* [illegible] *kdistmin*$_T$*(kamu)* dihitung sebagai berikut:

$$( ) = \min_{\in ( )} (,) \quad (16)$$

Skor *kamu* untuk kelas *T* dihitung berdasarkan skor-z yang dimodifikasi diberikan sebagai berikut (menggunakan mean dan varians yang dihitung sebelumnya):

$$- = - \left| - \frac{(()-)}{} \right. \quad (17)$$

Jika KLN-Mins akan digunakan sebagai pengklasifikasi tunggal maka parameternya β_T untuk *T*=1 hingga c harus disetel dengan baik. Pada kasus ini *kamu* ditugaskan ke kelas *T* mempunyai skor maksimal. Pendekatan lain adalah dengan menetapkan β_T=1 untuk *T*=1 hingga c dan susun pengklasifikasi tunggal ini. Namun, jika KLN-Min digabungkan dengan model keputusan lain menggunakan penumpukan maka parameter tersebut tidak perlu disesuaikan (yaitu ditetapkan ke 1) dan skor akan dimasukkan ke dalam pengklasifikasi meta.

4.3.4. Model keputusan yang diusulkan berdasarkan keterkaitan lengkap (KLN-Max)

Model ini mengikuti prosedur yang serupa dengan KLN-Avg di atas, namun berdasarkan pada *kdistmax*$_T$*(kamu)* yang dihitung sebagai berikut:

$$\in ( ) = (,) \quad (18)$$

skor sampel *kamu* untuk kelas *T* dihitung berdasarkan skor-z yang dimodifikasi sebagai berikut:

$$\text{KLN-Maks}=e^{-\eta t \left| \text{dist maxt (u)- kdist maxt)/k max sigmat} \right|} \quad (19)$$

Jika KLN-Mins akan digunakan sebagai pengklasifikasi tunggal maka parameternya η_T untuk *T*=1 hingga c harus disetel dengan baik. Pada kasus ini *kamu* ditugaskan ke kelas *T* mempunyai skor maksimal. Pendekatan lain adalah dengan mengatur η_T=1 untuk *T*=1 hingga c dan susun pengklasifikasi tunggal ini. Namun, jika KLN-Min digabungkan dengan model keputusan lain menggunakan penumpukan maka parameter tersebut tidak perlu disesuaikan (yaitu ditetapkan ke 1) dan skor akan dimasukkan ke dalam pengklasifikasi meta.

4.3.5. Usulan Model Keputusan berdasarkan Kepadatan (KDN)

KDN tidak memerlukan statistik yang dihitung sebelumnya. Sebaliknya, sampel masuk *kamu* diklasifikasikan ke dalam kelas *T* jika menambahkan sampel ini ke kelas *T* menghasilkan peningkatan kepadatan kelas t yang lebih besar dibandingkan kelas lainnya. Biarkan N_T(*kamu*) adalah himpunan yang memuat K tetangga terdekatnya *kamu* dari kelas *T*. Untuk mengukur peningkatan kepadatan setiap kelas, KNN-

1- Dalam makalah ini parameter K dan *k* memainkan peran yang berbeda (lihat Tabel 1).

tabel dipindai dan apakah sampel masuk *kamu* dapat menggantikan K tetangga dari sampel lain X_i ke kelas T, skor kelas T bertambah satu. Terakhir, kelas yang memiliki nilai maksimal dipilih. Untuk mengurangi kompleksitas komputasi, hanya N_T (*kamu*) untuk setiap kelas T diperiksa. Untuk menghitung skor kelas T, awalnya disetel ke nol dan diperbarui sebagai berikut:

untuk setiap $X_i \in N_T$(*kamu*)
  Identifikasi sampel terjauh X_H di antara tetangga terdekat X_i (dari tabel KNN). Jika (d(X_{kamu},X_i) <d(X_H,X_i)) Kemudian
    tambahkan 1 pada skor kelas T Lain
  jika (d(X_H,X_i))/ D(X_{kamu},X_i)) > maka
    [(D(X_H,X_i))/ D(X_{kamu},X_i)) – δ]/K ditambahkan ke nilai kelas T
berakhir jika;
akhir untuk;

Jika nilai δ diatur ke 1 maka kemungkinan skor seri meningkat. Dalam hal ini algoritma harus melakukan prosedur yang sama pada tetangga tingkat atas (tetangga dari tetangga) yang dapat menambah waktu klasifikasi. Kami menggunakan 0,9 sebagai nilai yang tepat untuk δ dalam percobaan kami.

4.3.6. Contoh Penjelasan

Tabel 3 menunjukkan contoh penghitungan skor model yang diusulkan pada kumpulan data buatan satu dimensi X. Kumpulan data X memiliki dua kelas positif dan negatif dan memiliki 10 nilai di setiap kelas yang tercantum dalam dua kolom pertama , masing-masing. Nilai K dipilih 2 sehingga jarak ke 2 tetangga terdekat dari masing-masing kelas dihitung seperti yang ditunjukkan pada kolom 3-4 dan 10-11. Pada kolom 5-7 dan 12-14 nilai rata-rata, minimum dan maksimum pada 3-4 dan 10-11 dihitung masing-masing untuk kelas positif dan negatif. μ dan σ pada kolom 5 dan 12 dihitung menggunakan Persamaan. (10) dan Persamaan. (13), masing-masing. Demikian pula, di kolom 6 dan 13, μ dan σ dihitung menggunakan Persamaan. (8) dan Persamaan. (11), masing-masing. Di kolom 7 dan 14, μ dan σ dihitung menggunakan Persamaan. (9) dan Persamaan. (12), masing-masing. Semua perhitungan sebelumnya dilakukan selama waktu pelatihan. Untuk mengklasifikasikan sampel yang tidak terlihat X=3 jarak ke 2-NN dari kelas positif (yaitu 2 dan 4) dihitung dalam kolom 3-4 seperti yang ditunjukkan pada baris sebelum yang terakhir. Selain itu, jarak ke 2-NN dari kelas negatif (yaitu 1 dan 5) dihitung di kolom 10-11. Pada kolom 5-7 dan 12-14 nilai rata-rata, minimum dan maksimum pada kolom 3-4 dan 10-11 dihitung masing-masing untuk kelas positif dan negatif. Skor kelas positif dan negatif dihitung pada baris terakhir untuk KLN-Rata-rata, KLN-Min dan KLN-Max 5, 6 dan 7 masing-masing. Untuk model KDN, X=3 memiliki 2-NN dari kelas positif (2, 4) dengan jarak (1, 1). X=3 lebih dekat ke X=2 dari 4 dan 7 (4 dan 7 adalah 2-NN dari X=2). Demikian pula, X=3 lebih dekat ke X=4 daripada 2 dan 7. Jadi totalnya ada empat tetangga yang dapat digantikan oleh X=3 di kelas positif. Demikian pula, total tiga tetangga dapat digantikan oleh X=3 di kelas negatif (δ = 1). Jika pemungutan suara maksimal digunakan, X=3 akan diklasifikasikan sebagai negatif karena KLN-Rata-rata, KLN-Min dan KLN-Max memiliki skor yang lebih tinggi untuk kelas negatif. Namun, Tradisional kNN (dengan k=1 atau 2) akan ditugaskan X=3 ke kelas positif.

Tabel 3 Contoh untuk menunjukkan bagaimana skor dihitung untuk model yang diusulkan pada Kumpulan Data Buatan

		Dist. 2-NN		Kdist			Skor KDN untuk X=3		Dist. 2-NN		Kdist		
X_+	X_-	untuk X_+		rata-rata	menit.	maks.	+	-	dari X_-		rata-rata	menit.	maks.
2	1	2	5	3.5	2	5	2	2	4	5	4.5	4	5
4	5	2	3	2.5	2	3	2	1	4	1	2.5	1	4
7	6	3	2	2.5	2	3	-	-	1	4	2.5	1	4
9	10	2	3	2.5	2	3	-	-	4	5	4.5	4	5
12	21	3	2	2.5	2	3	-	-	11	1	6	1	11
14	22	2	2	2	2	2	-	-	1	1	1	1	1
16	23	2	3	2.5	2	3	-	-	1	1	1	1	1
19	24	3	1	2	1	3	-	-	1	1	1	1	1
20	25	1	6	3.5	1	6	-	-	1	2	1.5	1.5	2
26	27	6	7	6.5	6	7	-	-	2	3	2.5	2.5	3
mikro				2.75	2.2	3.7	-	-			2.7	1.8	3.7
Σ				1.438	1.398	1.702	-	-			1.75	1.25	3.02
untuk X=3		1	1	1	1	1	4	3	2	2	2	2	2
skor				+ . 29	- . 77	+ . 42	- . 85	+ . 20	- . 63	+ . 57	- . 43	2	2

4.4. Menggabungkan keputusan masing-masing model

Dalam penelitian ini, masing-masing model EKNN diterapkan pada kumpulan data X dan prediksinya dicatat. Skor dari *k*NN, KLN-Avg, KLN-Min, KLN-Max, dan KDN untuk dataset input X dapat direpresentasikan sebagai tabel ukuran *N* oleh 5. Pemungutan suara mayoritas tertimbang dan penumpukan dapat digunakan untuk menggabungkan keluaran masing-masing model. Dalam penelitian ini, skor dari lima model keputusan digunakan sebagai masukan untuk pengklasifikasi meta lain untuk belajar dari keluaran model tahap pertama. Dengan menggunakan pemungutan suara mayoritas tertimbang [52], bobot harus ditentukan untuk model keputusan EKNN yang berbeda berdasarkan kinerja masing-masing model. Pendekatan naive brute force dapat digunakan untuk mencari bobot yang sesuai untuk setiap model keputusan sehingga diperoleh hasil klasifikasi ansambel yang optimal. Bobot biasanya dihitung sedemikian rupa sehingga jumlahnya sama dengan satu. Dalam hasil eksperimen kami, pendekatan bertumpuk digunakan dalam menggabungkan model keputusan. Semua parameter model yang diusulkan ditetapkan ke 1 untuk semua kelas dan skor diteruskan ke pengklasifikasi meta, sedangkan jika pendekatan pemungutan suara digunakan, parameter α_T, β_T Dan η_T untuk setiap kelas T harus disetel dan sampel *kamu* ditugaskan ke kelas T mana pun yang mempunyai skor maksimal.

5. Hasil Eksperimen dan Pembahasan

Untuk memvalidasi kinerja algoritma yang diusulkan, algoritma ini telah diuji pada lima kumpulan data kanker standar. Setelah langkah pemilihan fitur, matriks ekspresi gen tereduksi dimasukkan ke algoritma sebagai masukan untuk klasifikasi ansambel. Keluaran dari setiap model klasifikasi dipertimbangkan dalam mengklasifikasikan sampel yang tidak terlihat.

5.1. Contoh Keputusan KNN Tradisional yang Salah Dibandingkan EKNN

Tabel 4 menunjukkan beberapa contoh prediksi tradisional yang salah *k*NN dibandingkan dengan empat model yang diusulkan dan pengklasifikasi ansambel EKNN. Misalnya, sampel kumpulan data Kentridge ke-18, ke-39, dan ke-50, pada Tabel 4, salah diklasifikasikan berdasarkan metode tradisional *k*pengklasifikasi NN. Selain itu, sampel GDS3257 ke-2, ke-20, dan ke-44 salah diklasifikasikan berdasarkan tradisional *k*NN. Terdapat peningkatan yang signifikan dalam kinerja EKNN secara keseluruhan dibandingkan dengan tradisional *k*NN. Terlihat dari Tabel 4 bahwa sampel sulit untuk diklasifikasikan *k*NN diklasifikasikan dengan benar oleh EKNN. Hal ini terutama disebabkan oleh kemampuan pengklasifikasi ansambel yang diusulkan EKNN untuk belajar dari prediksi berbagai model keputusan. Agar dapat menggunakan max voting pada Tabel 4, alih-alih menyempurnakan parameter model yang diusulkan, kami menumpuknya satu per satu dengan nilai satu yang diberikan pada parameter setiap kelas dan output dari pengklasifikasi meta digabungkan menggunakan max voting . Untuk kinerja yang sangat prediktif, nilai K juga harus disesuaikan secara terpisah untuk setiap pengklasifikasi individu yang diusulkan. Selain itu, tidak semua model individu harus berpartisipasi dalam pengambilan keputusan akhir. Terkadang menyusun subset dari pengklasifikasi individual yang diusulkan dapat menghasilkan kinerja yang lebih tinggi.

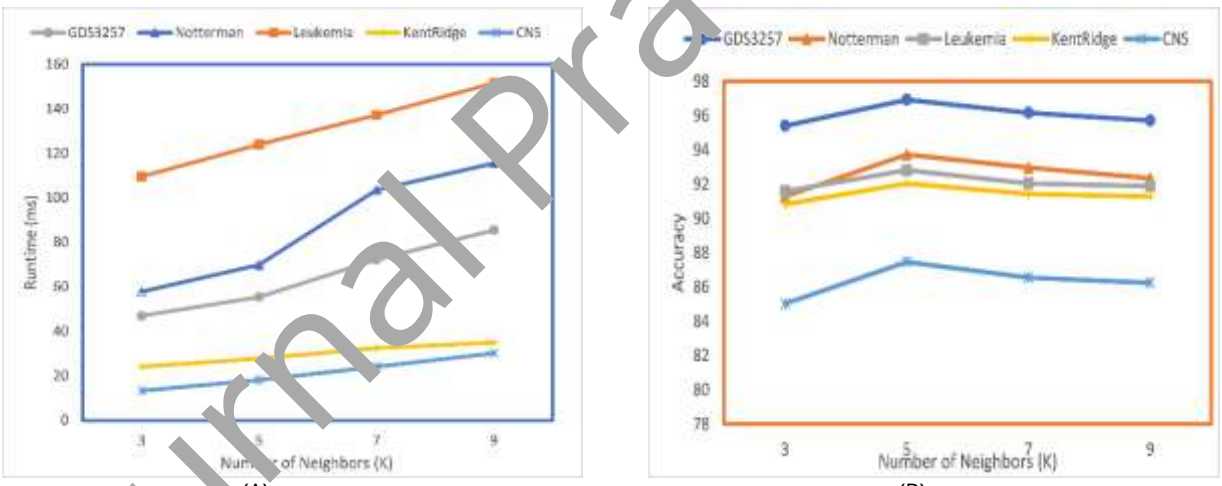
1- Dalam makalah ini parameter K dan *k* memainkan peran yang berbeda (lihat Tabel 1).

Tabel 4 Label yang Diberikan Pada Sampel Keras Dengan Berbagai Model Keputusan

Himpunan data	Sampel TIDAK.	Sebenarnya Label	Ramalan					
			*k*NN	KDN	KLN-Min	KLN-Rata-rata	KLN-Maks	EKNN
Kentridge	18	- 1	1	- 1	1	- 1	- 1	- 1
	39	- 1	1	- 1	- 1	- 1	1	- 1
	50	- 1	1	- 1	1	- 1	- 1	- 1
GDS3257	2	- 1	1	- 1	- 1	- 1	- 1	- 1
	20	- 1	1	- 1	- 1	- 1	1	- 1
	44	- 1	1	- 1	- 1	- 1	- 1	- 1
Notterman	1	1	- 1	1	- 1	1	1	1
	19	- 1	1	- 1	- 1	- 1	- 1	- 1

5.2. Menyetel parameter input K untuk EKNN

Banyaknya tetangga terdekat K mempengaruhi kinerja pengklasifikasi EKNN. Oleh karena itu, dalam penelitian ini dipelajari pengaruh K terhadap akurasi klasifikasi dan waktu klasifikasi. Hasilnya ditunjukkan pada Gambar. 3 (a), yang menunjukkan peningkatan waktu klasifikasi dengan peningkatan K. Gambar. 3 (b) mencerminkan bahwa akurasi klasifikasi meningkat hingga K = 5, dan melampaui titik ini kinerja klasifikasi juga memburuk atau tetap hampir sama. Dengan menggunakan WEKA, kami menemukan bahwa keakuratannya tradisional*k*NN meningkat hingga*k*=3 sementara Notterman mencapai akurasi terbaiknya *k*=1. Model keputusan yang diusulkan memanfaatkan informasi tetangga terdekat lebih baik daripada model tradisional*k*NN dengan bantuan tabel EKNN.



(A) (B)

Gambar 3. Jumlah tetangga terdekat (K) versus (a) Waktu klasifikasi (b) Akurasi klasifikasi EKNN

5.3. Kinerja EKNN pada dataset standar

Tabel 5 merangkum hasil klasifikasi EKNN dan membandingkan akurasinya dengan kelima model dasarnya: *k*NN, KLN-Rata-rata, KLN-Min, KLN-Max dan KDN. Selain itu, EKNN dibandingkan dengan empat metode ansambel lainnya berdasarkan pohon keputusan: Random Forest, Bagging, dan AdaBoost. Kolom terakhir memberikan hasil kumpulan tiga algoritma pembelajaran mesin*k*NN, DT dan NN menggunakan penumpukan dengan regresi logistik sebagai meta-classifier. Penumpukan dengan regresi logistik juga digunakan untuk menggabungkan model individual EKNN. Seperti yang ditunjukkan pada Tabel 5, kinerja EKNN lebih baik *k*NN serta semua pengklasifikasi dasarnya pada empat kumpulan data. Namun, sehubungan dengan pengklasifikasi ansambel lainnya, mereka

1- Dalam makalah ini parameter K dan*k*memainkan peran yang berbeda (lihat Tabel 1).

memberikan akurasi yang cukup tinggi pada beberapa dataset, classifier EKNN yang diusulkan masih mampu mengungguli semuanya pada dataset Ionosphere dan Parkinson.

Tabel 5 Hasil Kinerja EKNN Dibandingkan Model Individual dan Metode Ensemble lainnya

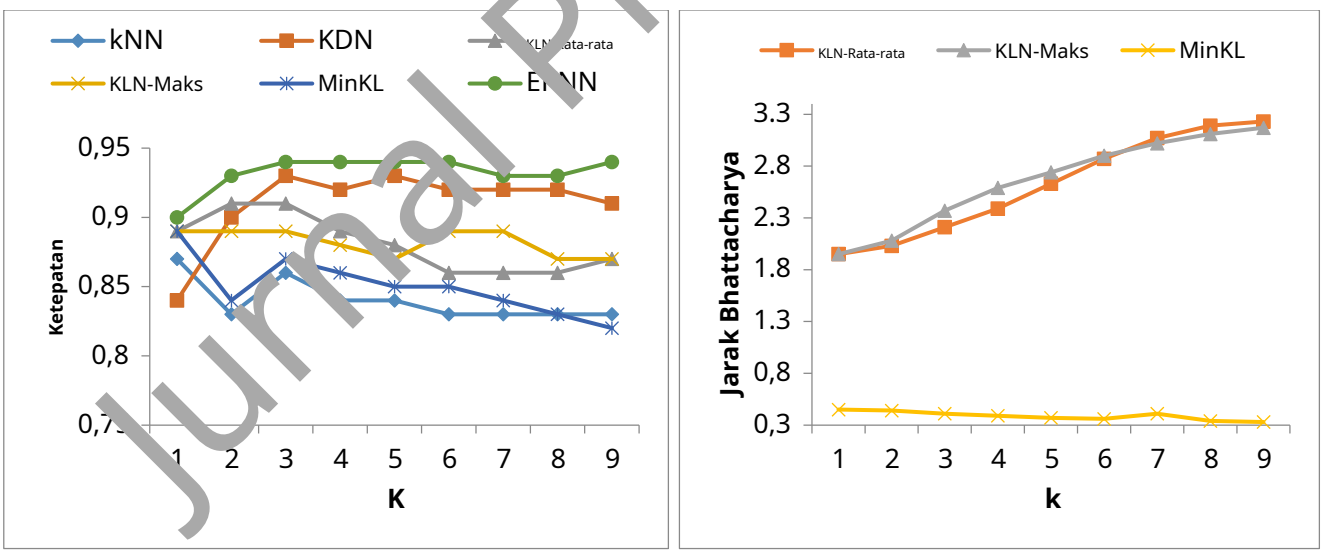
Himpunan data	Sampel/ Fitur/ Kelas	*k*NN	KLN-Minimal	KLN - Rata-rata	KLN - Maks	KDN	EKNN Menumpuk	Acak Hutan	Pohon Keputusan Mengantongi	AdaBoost	Menumpuk (NN,KNN, DT)
parkinson	195/23/2	79,82	74,67	78,31	75,62	90,31	**93,78**	92.83	82,65	85.12	92.82
Ionosfir	351/34/2	87,60	81,23	91,18	89,41	93,13	**96.48**	93.17	88.39	90,88	92,59
Transfusi	748/05/2	69,85	68,45	70,19	69,67	71,01	74,23	73.26	74.53	74.34	**77.27**
Diabetes	768/08/2	72,65	66,16	71,81	70,32	71,25	75,81	75,78	75.91	74.34	**76.17**

5.4. Pengaruh perubahan nilai K terhadap keragaman sebaran KLN-Avg, KLN-Min dan KLN-Max

Seperti terlihat pada Tabel 6, untuk dataset Ionosfer, kesenjangan kinerja antara aturan MinKL dan aturan mayoritas sangat kecil, terutama untuk K kecil. Hal ini disebabkan oleh fakta bahwa distribusi pusat untuk dataset Ionosfer sangat dekat dengan Fungsi Dirac Delta dalam hal ini MinKL direduksi menjadi kekuasaan mayoritas [25].

Selain itu, kinerja model KLN-Avg dan KLN-Max secara signifikan lebih baik daripada MinKL untuk semua nilai K. Penjelasan kami untuk peningkatan kinerja ini adalah fakta bahwa aturan KLN-Avg dan KLN-Max membuat prediksi berdasarkan seluruh kelas distribusi.

Faktor penting yang mempengaruhi kinerja model keputusan yang diusulkan KLN-Avg dan KLN-Max adalah jarak antara pusat distribusinya untuk setiap kelas. Semakin jauh jaraknya, semakin baik kinerja model keputusan tersebut. Di sisi lain, model KDN tidak mengasumsikan sebaran tertentu tetapi hanya mengukur peningkatan kepadatan. KDN diharapkan memiliki kinerja yang lebih baik dibandingkan model lainnya ketika distribusi kelas yang berbeda berdekatan satu sama lain.



(a) Akurasi

(b) keanekaragaman (jarak Bhattacharya)

Gambar 4. (a) Akurasi algoritma terkait dibandingkan dengan EKNN untuk nilai K yang berbeda (b) Jarak antara distribusi dua kelas dataset Ionosfer

Tabel 6 menunjukkan distribusi parameter link rata-rata dan maksimum untuk dua kelas dataset Ionosfer yaitu p dan q. Untuk MinKL, distribusi probabilitas bersifat diskrit dan koefisien Bhattacharyya dihitung seperti pada Persamaan(2) sedangkan untuk KLN-Avg dan KLN-Max distribusinya kontinu dan Persamaan(1) digunakan.

Seperti terlihat pada Tabel 6, sebaran dataset KLN-Avg dan KLN-Max pada dataset Ionosphere lebih tersebar dibandingkan dengan MinKL untuk dataset yang sama dengan nilai K yang berbeda. Model KLN-Min tidak memiliki K sebagai parameter. Dari definisinya nilai K selalu satu (oleh karena itu tidak tercantum pada Tabel 6). Koefisien Bhattacharyya di Ionosfer untuk KLN-Min adalah 0,00124. Jarak antara kedua distribusi MinKL sedikit menurun seiring dengan meningkatnya nilai K sedangkan untuk KLN-Avg dan KLN-Max terus meningkat seiring dengan peningkatan K. Hal ini mungkin menjelaskan mengapa kinerja EKNN terbaik ditemukan pada nilai K biasanya lebih besar dibandingkan MinKL dan *k*NN. Namun, semakin besar nilai K maka estimasi distribusinya kurang akurat.

1- Dalam makalah ini parameter K dan *k* memainkan peran yang berbeda (lihat Tabel 1).

Tabel 6 Pengaruh perubahan nilai K terhadap distribusi model usulan vs MinKL

K		KLN-Maks			KLN-Rata-rata			MinKL		
		μ_{mikro}	σ	D	μ_{mikro}	σ	D	P	Q	D
1	P	0,97	0,05	1.59	0,97	0,05	1.59	0,985	0,015	0,45
	Q	0,60	0,15		0,60	0,15		0,289	0,711	
3	P	0,97	0,05	2.21	0,96	0,06	2.37	0,992	0,008	0,41
	Q	0,55	0,13		0,51	0,14		0,351	0,649	
5	P	0,96	0,06	2.63	0,95	0,06	2.74	0,993	0,007	0,37
	Q	0,52	0,12		0,46	0,14		0,393	0,607	
7	P	0,95	0,06	3.04	0,94	0,06	3.02	0,995	0,005	0,35
	Q	0,50	0,12		0,43	0,13		0,427	0,573	
9	P	0,96	0,07	3.34	0,93	0,07	3.17	0,996	0,004	0,33
	Q	0,48	0,11		0,41	0,13		0,451	0,549	

5.5. Perbandingan kinerja EKNN dengan pengklasifikasi dasarnya dan Ensemble on Cancer Data lainnya

Tabel 7 menunjukkan kinerja EKNN menggunakan GCA [53] sebagai teknik reduksi dimensi dibandingkan dengan model dasar dan tiga pengklasifikasi ansambel, yaitu AdaBoost, Bagging dan Random Forest, pada lima dataset kanker. Pohon Keputusan (DT) dan *k*NN digunakan sebagai pengklasifikasi dasar untuk AdaBoost dan Bagging. Hasil untuk pengklasifikasi ansambel lainnya diperoleh dengan menggunakan WEKA [12]. Pada tabel 7, Random Forest mencapai akurasi tertinggi pada dua dataset sementara EKNN mencapai kinerja terbaik dibandingkan ansambel lain pada tiga dari lima dataset dalam hal akurasi.

Selain itu, seperti ditunjukkan pada Tabel 7, EKNN mengungguli EKNN tradisional. EKNN memiliki akurasi tertinggi dibandingkan semua model individualnya. Akurasi klasifikasi ensemble EKNN pada Kentridge, GDS3257, Leukemia, CNS dan Prostate-I lebih tinggi dibandingkan akurasi terbaik model individual masing-masing sebesar 4,16%, 1,60%, 1,65%, 2,50%, dan 1,30%.

Tabel 7 Hasil Kinerja EKNN Dibandingkan Model Individu dan Ensemble lainnya

Himpunan data	***k*NN**		KDN	KLN-Rata-rata	KLN-Maks	KLN-Minimal	EKNN (Mayoritas)			Akurasi Ensembel lainnya				
										Acak Hutan	Mengantongi		AdaBoost	
											*k*NN	DT	*k*NN	DT
	Acc.	*k*	Acc.	Acc.	Acc.	Acc.	Acc.	Bal. Acc.	AUC					
Kentridge	87.8	3	79.6	81.1	**89.4**	85.5	**93.12**	92.01	0,983	91.9	87.0	83.8	83.9	87.1
GDS3257	93.7	1	85.2	**97.5**	94.7	93.3	99.06	99.06	0,991	**100,0**	99.0	97.2	99.1	94.4
Leukemia	92.3	2	**97.2**	87.3	95.4	88.7	**98.81**	98.13	0,997	98.6	98.6	94.4	97.2	88.9
SSP	**76.2**	1	75.1	73.6	71.3	70.2	78.11	74.83	0,798	**88.3**	75.0	80.0	80.0	71.6
Prostat-I	93.1	1	95.1	94.1	**96.9**	95.4	**98.16**	97.31	0,976	96.3	88.2	88.2	86.8	91.9

5.6. Perbandingan kinerja EKNN dengan algoritma lain yang berasal dari kNN

Tabel 8 menunjukkan akurasi dan peringkat kinerja klasifikasi pengklasifikasi EKNN relatif terhadap kinerja delapan yang dilaporkan *k* Pengklasifikasi terkait NN pada tiga kumpulan data dunia nyata dari repositori UCI. Akurasi dari delapan pengklasifikasi lainnya pada Tabel 8 dilaporkan pada [20]. Peringkat "1" adalah yang terbaik dan "8" adalah yang terburuk untuk setiap kumpulan data. Seperti yang ditunjukkan pada Tabel 8, metode EKNN yang diusulkan mencapai akurasi tertinggi pada dua dari tiga kumpulan data dunia nyata jika dibandingkan dengan rata-rata semua pengklasifikasi lainnya.

1- Dalam makalah ini parameter K dan *k* memainkan peran yang berbeda (lihat Tabel 1).

Tabel 8 Akurasi EKNN Dibandingkan dengan algoritma relatif terhadap kNN

Data (Perwakilan UI)	sampel./ att./kelas.	LMKNN	MKNN	PNN	WKNN	CFKNN	FRNN	HBKNN	kGNN	EKNN
Dada	277/9/2	71.82	71.07	72.01	72.46	70.24	68.27(8)	71,80	**72,86(1)**	71,94(4)
Jantung	303/13/2	79,98	77,81(8)	79.17	77,95	79.12	80,96	79.11	79.31	**81.31(1)**
Hati	345/6/2	65.21	64.13	64.72	64.42	65,78	58,85(8)	64,77	65.81	**65,87(1)**

5.7. Evaluasi Kinerja EKNN dengan Teknik Reduksi Dimensi Berbeda

Tabel 9 menunjukkan kinerja EKNN dibandingkan dengan hasil yang dilaporkan pada [53] untuk beberapa teknik klasifikasi menggunakan dua teknik reduksi dimensi yang berbeda yaitu ACA [54] dan GCA [53]. EKNN memiliki akurasi yang lebih rendah dibandingkan NN tetapi memiliki akurasi tertinggi dibandingkan dengan Decision Trees, Naïve Bayes dan*k*NN. Pendekatan yang diusulkan dibandingkan dengan NN adalah pendekatan yang sederhana, inkremental, dan memiliki waktu pelatihan yang lebih sedikit.

Seperti terlihat pada tabel 10, akurasi EKNN pada dataset Kentridge adalah 93,3 yang merupakan yang terbaik di antara pengklasifikasi lainnya. Keakuratan NN, Decision Trees, Naïve Bayes dan*k*NN adalah yang dilaporkan dalam [54] menggunakan ACA. Selain itu, EKNN mencapai akurasi 99,06 pada GDS3257 yang lebih tinggi dari rata-rata akurasi yang dilaporkan TC-VGC [13] yaitu 95,16 (menggunakan parameter*kor*=0,8-09 dan hasil rata-rata pada parameter masukan lain disebut*tanda tangan*). TC-VGC sensitif terhadap dua parameter masukannya*kor*Dan*tanda tangan*dan memerlukan penyetelan parameter ekstensif untuk parameter ini tetapi tidak memerlukan langkah reduksi dimensi. Demikian pula, hasilnya membuktikan keunggulan EKNN dibandingkan model individual, apa pun pilihan strategi seleksi gennya.

Tabel 9 Kinerja Berbagai Algoritma Klasifikasi pada Kumpulan Data Leukemia

Klasifikasi Algoritma	Ketepatan/ Redup. Pengurangan Teknik	Ketepatan/ Redup. Pengurangan Teknik	Akurasi Tanpa Redup. Pengurangan
Pohon Keputusan	94.[illegible]ACA	95.3/GCA	91.2
NN	97.1/AC[illegible]	96.2/GCA	91.2
Bayes Naif	82.4/ACA	68.6/GCA	41.2
Hutan Acak	9[illegible]6/PCA	98.6/ PCA	88.9
SVM(RBF)	93.1/ PCA	93.1/ PCA	65.3
kNN	91.2/ACA	92.3/GCA	82.4
EKNN	94,7/ACA	95,91/GCA	88.4

Tabel 10 Performa Berbagai Algoritma Klasifikasi pada Dataset Kentridge

K[illegible]sifikasi [illegible]goritma	Ketepatan/ Redup. Pengurangan Teknik	Ketepatan/ Redup. Pengurangan Teknik	Akurasi Tanpa Redup. Pengurangan
Pohon Keputusan	91.9/ACA	93.1/GCA	82.3
NN	90.3/ACA	92.3/GCA	83.9
Bayes Naif	67,7/ACA	71.6/GCA	35.5
Hutan Acak	85,5/ PCA	90.3/GCA	83.9
SVM(RBF)	85.4/ PCA	87.1/GCA	80.6
kNN	83.9/ACC	87,8/GCA	79.0
EKNN	93.1/ACA	93.3/GCA	85.2

5.8. Kompleksitas komputasi EKNN

Membangun tabel KNN pada tahap pelatihan adalah O(*buku*$_2$), Di mana *N* adalah jumlah sampel dan K adalah jumlah sampel terdekat. Memori yang dibutuhkan untuk algoritma yang diusulkan hanya O(*buku*) sel. Sebelum menerapkan salah satu model individual untuk mengklasifikasikan sampel yang masuk, jarak dari sampel ini ke semua sampel yang disimpan perlu dihitung dengan cara yang serupa dengan *k*NN saja, hal ini membutuhkan komputasi yang intensif, terutama ketika ukuran set pelatihan bertambah tetapi biasanya jumlah sampel jauh lebih sedikit dibandingkan jumlah gen dalam set data kanker. Oleh karena itu, waktu klasifikasi model KLN yang diusulkan adalah O(*N*) sedangkan untuk KDN, jika kita menguji hanya tetangga dari sampel yang masuk untuk diklasifikasikan, maka waktu klasifikasinya adalah O(*k*+*N*)=O(*N*). Total kompleksitas komputasi untuk mengklasifikasikan sampel menggunakan EKNN adalah O(*N*) tidak termasuk biaya penumpukan.

5.9. Waktu berjalan EKNN

Tabel 11 menunjukkan waktu CPU yang dibutuhkan selama fase pelatihan dan pengujian EKNN dibandingkan dengan NN, SVM dan RF. Waktu pelatihan sesuai dengan waktu kecocokan rata-rata dari pemisahan CV 5 kali lipat yang dilakukan menggunakan nilai parameter optimal, dan waktu pengujian adalah waktu klasifikasi rata-rata (skor) dari satu sampel. Waktu pelatihan EKNN adalah jumlah waktu yang digunakan oleh masing-masing model, sedangkan waktu pengujian melibatkan overhead untuk penumpukan (meta classifier) selain waktu pengujian yang digunakan oleh masing-masing model. Perlu dicatat bahwa waktu untuk mereduksi dimensi dan memperkirakan jumlah tetangga yang akan disimpan dalam tabel KNN tidak disertakan di sini. Hasil pada Tabel 11 menunjukkan bahwa teknik EKNN yang diusulkan dapat dilakukan secara komputasi. Misalnya, waktu pelatihan maksimum untuk kumpulan data terbesar (GDS3257 dengan 107 sampel) hanya 36,4 ms, yang jauh lebih rendah dibandingkan waktu RF dan NN, namun jauh lebih tinggi dibandingkan SVM. Sebagian besar waktu pelatihan dan pengujian EKNN dihabiskan dalam komputasi tetangga, penggunaan kd-tree [55] selanjutnya dapat mengurangi waktu pelatihan dan pengujian EKNN menggunakan sejumlah kecil fitur setelah reduksi dimensi. Seperti terlihat pada Tabel 11, waktu pengujian *k* NN yang menggunakan kd-tree sebanding dengan SVM dan membutuhkan waktu sekitar 5% dari waktu pengujian tanpa menggunakan kd-tree. Model keputusan yang diusulkan diimplementasikan dalam Python 3.8 di windows 10 dan semua tugas pembelajaran mesin telah dilakukan menggunakan Scikit-Learn [50] dengan Intel Core i5, prosesor 2,5 GHz, dan RAM 16 GB. Kode sumber serta beberapa kumpulan data yang diproses dan contoh keluaran akan tersedia di [56].

Tabel 11 rata-rata runtime EKNN dibandingkan dengan algoritma terkait (ms)

Himpunan data	Sel. Gen	Tidak. Sampel	Waktu pelatihan rata-rata untuk satu splite CV 5 kali lipat (ms)				Waktu pengujian rata-rata untuk satu sampel (ms)					
			SVM	RF	NN	EKNN	SVM	Federasi Rusia	NN	kNN	kNN (kd-pohon)	EKNN
Kentridge	21	62	0,36	91.6	95.8	15.[illegible]	0,17	1.09	0,17	8.38	0,29	22.5
GDS3257	34	107	0,51	91.2	117	36.4	0,10	0,64	0,10	13.8	0,16	33.6
Notterman	55	36	0,38	99.[illegible]	99.6	8.6	0,14	0,88	0,13	6.95	0,43	17.4
Leukemia	50	72	0,37	[illegible].1	128.6	20.2	0,15	0,98	0,15	9.84	0,24	25.1
SSP	84	60	0,44	93.4	102.6	14.2	0,16	1.15	0,21	8.03	0,27	21.7

6. Kesimpulan

Dalam kajian penelitian ini, EKNN yang diusulkan merupakan ansambel tradisional *k*NN dan empat model klasifikasi baru yang diusulkan berasal dari *k*NN diistilahkan sebagai KDN, KLN-Avg, KLN-Min dan KLN-Max. Dari studi eksperimental kami, hal berikut dapat disimpulkan.

- Model keputusan yang diusulkan mampu mengklasifikasikan sampel dengan benar yang mungkin salah diklasifikasikan secara tradisional *k*NN.
- Ensembel yang diusulkan EKNN selalu berkinerja lebih baik daripada model individual pada semua kumpulan data yang diselidiki.
- Penggunaan tabel KNN yang dibangun pada tahap pelatihan mengurangi waktu klasifikasi EKNN dan model individu yang diusulkan.
- Tabel KNN dapat membantu mengatasi data yang tidak seimbang.
- Tabel KNN dan statistik yang dihitung dalam fase pelatihan dapat diperbarui secara bertahap dan algoritma yang diusulkan, dengan sedikit modifikasi, dapat bekerja pada data aliran namun hal ini di luar cakupan makalah ini.
- Waktu pelatihan dan klasifikasi model keputusan yang diusulkan dan tradisional *k*NN dapat direduksi secara signifikan menggunakan kd-tree seperti yang ditunjukkan pada Tabel 11. Hal ini memungkinkan EKNN digunakan untuk kumpulan data yang besar.

1- Dalam makalah ini parameter K dan *k* memainkan peran yang berbeda (lihat Tabel 1).

- Jarak Bhattacharyya memberikan wawasan tentang kompleksitas kumpulan data. Semakin besar jarak antara distribusi kelas-kelas yang berbeda, untuk suatu model keputusan tertentu, semakin rendah kompleksitasnya dan semakin tinggi kinerja yang diharapkan dari model keputusan tersebut. Selain itu, dapat membantu mengurangi ruang pencarian dalam menyetel nilai K.

Oleh karena itu, kami dapat menyimpulkan secara masuk akal bahwa usulan EKNN bersama dengan teknik reduksi dimensi dapat membantu para ahli biologi dalam memprediksi kanker di beberapa bagian tubuh secara akurat. Analisis skema yang diusulkan dalam makalah ini menyarankan beberapa arah untuk pekerjaan di masa depan:

1. Menggunakan ansambel KDN, KLN-Avg, KLN-Min atau KLN-Max sebagai ganti*k*NN sebagai pemilih gen mirip dengan [57].
2. Mencari teknik scalable untuk membangun tabel KNN untuk kumpulan data yang sangat besar.
3. Dalam pekerjaan ini kami menggabungkan kelima pengklasifikasi secara bersamaan, namun terkadang menumpuk lebih sedikit pengklasifikasi dapat menghasilkan kinerja yang lebih tinggi. Untuk menyempurnakan EKNN agar memiliki kinerja prediktif yang tinggi, semua kemungkinan kombinasi lebih dari dua pengklasifikasi (26 kombinasi) dapat dicoba dalam aplikasi kehidupan nyata.
4. Menerapkan EKNN pada permasalahan lain yang ditandai dengan jumlah kelas yang besar dan ukuran data latih yang kecil seperti model deteksi kebocoran yang diusulkan pada [32].

Telah ditemukan bahwa dengan mempartisi kumpulan data yang sangat besar, sebagian besar*k*-tetangga terdekat dapat ditemukan secara lokal di setiap partisi daripada mencari seluruh kumpulan data [58]. Dalam [58], teknik scalable untuk*k*NN, yang mengandalkan pengelompokan fuzzy, telah diperkenalkan untuk menghasilkan tabel KNN untuk kumpulan data besar. Di setiap partisi, sampel inti dan batas dapat diperkirakan menggunakan teori himpunan kasar. Strategi yang mungkin dilakukan adalah memulai dengan mempartisi Xlearning* menggunakan pengelompokan fuzzy, menghitung tabel KNN untuk inti setiap partisi dan tabel lain untuk semua sampel batas, lalu menggabungkan semua tabel KNN menjadi satu. Dalam studi eksperimental kami, kami mempertimbangkan kumpulan data kecil sehingga langkah partisi tidak diperlukan.

Konflik kepentingan

Para penulis menyatakan tidak ada konflik kepentingan.

Referensi

[1] Shaikhina, T. dan NA Khovanova,*Menangani kumpulan data terbatas dengan jaringan saraf dalam aplikasi medis: Pendekatan data kecil.*Kecerdasan buatan dalam kedokteran, 2017.**75**: P. 51-63.

[2] Dudani, SA,*Aturan k-nearest-neighbor berbobot jarak.*Transaksi IEEE pada Sistem, Manusia, dan Sibernetika, 1976(4): hal. 325-327.

[3] Batista, G. dan DF Silva.*Bagaimana parameter k-nearest neighbour mempengaruhi kinerjanya*. di dalam*Simposium Argentina tentang kecerdasan buatan*. 2009. hal.

[4] Deng, Z., dkk.,*Algoritma klasifikasi kNN yang efisien untuk data besar.*Neurokomputer, 2016.**195**: P. 143-148.

[5] Kuncheva, LI,*Menggabungkan pengklasifikasi pola: metode dan algoritma*. 2004: John Wiley & Putra.

[6] Kuncheva, LI dan JJ Rodriguez,*Ansambel pengklasifikasi dengan oracle linier acak.*Transaksi IEEE tentang Pengetahuan dan Rekayasa Data, 2007. **19**(4): hal. 500-508.

[7] Bhattacharya, A.,*Tentang ukuran perbedaan antara dua populasi statistik yang ditentukan oleh distribusi probabilitasnya.*Banteng. Matematika Kalkuta. sosial, 1943.**35**: P. 99-109.

[8] Freund, Y. dan RE Schapire.*Generalisasi teori keputusan pembelajaran online dan penerapan untuk peningkatan*. di dalam*Konferensi Eropa tentang teori pembelajaran komputasi*. 1995. Peloncat.

[9] Khan, J., dkk.,*Klasifikasi dan prediksi diagnostik kanker menggunakan profil ekspresi gen dan jaringan saraf tiruan.*Pengobatan alam, 2001.**7**(6): hal. 673-679.

[10] Wang, Y., dkk.,*Seleksi gen dari data microarray untuk klasifikasi kanker--pendekatan pembelajaran mesin.*Komputasi Biol Kimia, 2005.**29**(1): hal. 37-46 DOI: 10.1016/j.compbiolchem.2004.11.001.

[11] Wu, M.-Y., dkk.,*Identifikasi biomarker dan klasifikasi kanker berdasarkan data microarray menggunakan model laplace naif bayes dengan mean shrinkage.*Transaksi IEEE/ACM pada biologi komputasi dan bioinformatika, 2012.**9**(6): hal. 1649-1662.

[12] Witten, IH, dkk.,*Penambangan Data: Alat dan teknik pembelajaran mesin praktis*. 2016: Morgan Kaufmann.

[13] Shin, E., dkk.,*TC-VGC: sistem klasifikasi tumor menggunakan variasi korelasi gen.*Metode dan program komputer dalam biomedis, 2011. **104**(3): hal. e87-e101.

[14] Mahfouz, MA*RBG-CD: Diagnosis Kanker Genetik Berbasis Residu*. di dalam*Konferensi Internasional tentang Sistem Cerdas dan Informatika Tingkat Lanjut*. 2016. Pegas DOI: 10.1007/978-3-319-48308-5_40.

[15] Keller, JM, MR Gray, dan JA Gives,*Algoritma fuzzy k-tetangga terdekat.*Transaksi IEEE pada sistem, manusia dan sibernetika, 1985(4): hal. 580-585.

[16] Sarkar, M.,*Algoritma tetangga terdekat yang fuzzy-kasar dalam klasifikasi.*Himpunan dan Sistem Fuzzy, 2007.**158**(19): hal. 2134-2152.

[17] Xu, Y., dkk.,*Pengklasifikasi tetangga terdekat dari kasar hingga halus K.*Surat pengenalan pola, 2013.**34**(9): hal. 980-983.

[18] Zeng, Y., Y. Yang, dan L. Zhao,*Aturan tetangga terdekat semu untuk klasifikasi pola.*Sistem Pakar dengan Aplikasi, 2009.**36**(2): hal. 3587- 3595.

[19] Liu, H. dan S. Zhang,*Penghapusan data yang bising menggunakan mutual k-nearest neighbour untuk penambangan klasifikasi.*Jurnal Sistem dan Perangkat Lunak, 2012. **85**(5): hal. 1067-1074.

[20] Lin, Y., dkk.,*Pengklasifikasi tetangga terdekat baru melalui penggabungan informasi lingkungan.*Neurokomputer, 2014.**143**: P. 164-169.

[21] Pan, Z., Y. Wang, dan W. Ku,*Pengklasifikasi tetangga terdekat k-harmonik baru berdasarkan cara multi-lokal.*Sistem Pakar dengan Aplikasi, 2017.**67**: P. 115-125.

[22] Medjahed, SA, TA Saadi, dan A. Benyettou,*Diagnosis Kanker Payudara dengan menggunakan k-Nearest Neighbor dengan Jarak dan Aturan Klasifikasi yang Berbeda.*Jurnal Internasional Aplikasi Komputer, 2013.**62**(1).

[23] Mitani, Y. dan Y. Hamamoto,*Pengklasifikasi nonparametrik berbasis rata-rata lokal.*Surat Pengenalan Pola, 2006.**27**(10): hal. 1151-1159.

[24] Syaliman, K., E. Nababan, dan O. Sitompul.*Meningkatkan akurasi k-nearest neighbour menggunakan mean lokal berdasarkan dan bobot jarak*. di dalam*Jurnal Fisika: Seri Konferensi*. 2018. Penerbitan IOP.

[25] Cheamanunkul, S. dan Y. Freund.*Peningkatan Aturan kNN untuk Set Pelatihan Kecil*. di dalam*Konferensi Internasional ke-13 2014 tentang Pembelajaran Mesin dan Aplikasi*. 2014.IEEE.

[26] Dai, JJ, L. Lieu, dan D. Rocke,*Pengurangan dimensi untuk klasifikasi dengan data microarray ekspresi gen.*Aplikasi statistik dalam genetika dan biologi molekuler, 2006.**5**(1).

[27] Kohavi, R. dan GH John,*Pembungkus untuk pemilihan subset fitur.*Kecerdasan buatan, 1997.**97**(1-2): hal. 273-324.

[28] Langley, P.*Pemilihan fitur yang relevan dalam pembelajaran mesin*. di dalam*Prosiding simposium Musim Gugur AAAI tentang relevansi*. 1994.

[29] Backert, S., dkk.,*Ekspresi gen diferensial dalam sel dan jaringan karsinoma usus besar terdeteksi dengan susunan cDNA.*Jurnal Internasional Kanker, 1999.**82**(6): hal. 868-874.

[30] Geng, Z., dkk.,*Optimalisasi energi dan pemodelan prediksi industri petrokimia: Jaringan saraf konvolusional yang ditingkatkan berdasarkan lintas fitur.*Energi, 2020.**194**: P. 116851.

[31] Han, Y., dkk.,*Evaluasi efisiensi energi industri petrokimia yang kompleks.*Energi, 2020: hal. 117893.

[32] Hu, X., dkk.,*Deteksi kebocoran baru dan pengelolaan kehilangan air pada jaringan pasokan air perkotaan menggunakan jaringan saraf multiskala.* Jurnal Produksi Bersih, 2020.**278**: P. 123611.

[33] Geng, Z., dkk.,*Peringatan dini dan pengendalian risiko keamanan pangan menggunakan jaringan saraf AHC-RBF yang ditingkatkan yang mengintegrasikan AHP-EW.*Jurnal Teknik Pangan.**292**: P. 110239.

[34] Geng, Z., dkk.,*Ekstraksi relasi semantik menggunakan LSTM sekuensial dan terstruktur pohon dengan perhatian.*Ilmu Informasi, 2020.**509**: P. 183-192.

[35] Chen, Z. dan J. Li.*Skema mesin vektor dukungan beberapa kernel untuk pemilihan fitur secara simultan dan klasifikasi berbasis aturan*. di dalam*Konferensi Asia Pasifik tentang Penemuan Pengetahuan dan Penambangan Data*. 2007. Peloncat.

[36] Rathore, S., M. Hussain, dan A. Khan,*GECC: Klasifikasi ansambel biopsi usus besar berdasarkan ekspresi gen.*

[37] Lu, H., dkk.,*Algoritme hutan rotasi yang hemat biaya untuk klasifikasi data ekspresi gen.*Neurokomputer, 2017.**228**: P. 270-276.

[38] Tan, S.,*K-tetangga terdekat berbobot tetangga untuk korpus teks tidak seimbang.*Sistem Pakar dengan Aplikasi, 2005.**28**(4): hal. 667-671.

[39] Ganganwar, V.,*Ikhtisar algoritme klasifikasi untuk kumpulan data tidak seimbang.*Jurnal Internasional Teknologi Berkembang dan Rekayasa Lanjutan, 2012.**2**(4): hal. 42-47.

[40] Mani, saya. dan saya. Zhang.*Pendekatan kNN terhadap distribusi data yang tidak seimbang: studi kasus yang melibatkan ekstraksi informasi*. di dalam*Prosiding lokakarya pembelajaran dari kumpulan data yang tidak seimbang*. 2003.

[41] Jadhav, AS,*Ukuran TPR-TNR berbobot baru untuk menilai kinerja pengklasifikasi.*Sistem Pakar dengan Aplikasi, 2020 : hal. 113391.

[42] Behzadian, M., dkk.,*Survei aplikasi TOPSIS yang canggih.*Sistem Pakar dengan aplikasi, 2012.**39**(17): hal. 13051-13069.

[43] Alon, U., dkk.,*Pola ekspresi gen yang luas terungkap melalui analisis pengelompokan tumor dan jaringan usus normal yang diperiksa dengan susunan oligonukleotida.*Prosiding National Academy of Sciences, 1999.**96**(12): hal. 6745-6750.

[44] Landi, MT, dkk.,*Tanda ekspresi gen dari merokok dan perannya dalam perkembangan dan kelangsungan hidup adenokarsinoma paru.*PloS satu, 2008. **3**(2): hal. e1651.

[45] Notterman, DA, dkk.,*Profil ekspresi gen transkripsional adenoma kolorektal, adenokarsinoma, dan jaringan normal diperiksa dengan susunan oligonukleotida.*Penelitian kanker, 2001.**61**(7): hal. 3124-3130.

[46] Golub, TR, dkk.,*Klasifikasi molekuler kanker: penemuan kelas dan prediksi kelas dengan pemantauan ekspresi gen.*sains, 1999.**286**(5439): P. 531-537.

[47] Pomeroy, SL, dkk.,*Prediksi hasil tumor embrio sistem saraf pusat berdasarkan ekspresi gen.*Alam, 2002.**415**(6870): hal. 436-442.

[48] Al-Shahrour, F., dkk.,*BABELOMICS: seperangkat alat web untuk anotasi fungsional dan analisis kelompok gen dalam eksperimen throughput tinggi.* Penelitian asam nukleat, 2005.**33**(tambahan 2): hal. W460-W464.

[49] Hassan, M., dkk.,*Segmentasi citra arteri karotis menggunakan c-means fuzzy spasial yang dimodifikasi dan pengelompokan ansambel.*Metode dan Program Komputer dalam Biomedis, 2012.**108**(3): hal. 1261-1276.

[50]*Scikit-learn: Pembelajaran Mesin dengan Python, Pedregosa dkk., JMLR 12, hlm. 2825-2830, 2011*. Tersedia dari: https:// scikitlearn.org/stable/about.html.

[51] Han, H., W.-Y. Wang, dan B.-H. Mao,*Borderline-SMOTE: metode pengambilan sampel berlebih yang baru dalam pembelajaran kumpulan data yang tidak seimbang.*Kemajuan dalam komputasi cerdas, 2005: hal. 878-887.

[52] Littlestone, N. dan MK Warmuth.*Algoritma mayoritas tertimbang*. di dalam*Yayasan Ilmu Komputer, 1989., Simposium Tahunan ke-30 tentang*. 1989.IEEE.

[53] Mahfouz, MA dan JA Nepomuceno,*Pewarnaan grafik untuk mengekstraksi gen diskriminatif dalam data kanker.*Sejarah genetika manusia, 2019.

[54] Au, W.-H., dkk.,*Pengelompokan atribut untuk pengelompokan, seleksi, dan klasifikasi data ekspresi gen.*Transaksi IEEE/ACM pada biologi komputasi dan bioinformatika, 2005.**2**(2): hal. 83-101.

[55] Zhou, K., dkk.,*Konstruksi kd-tree waktu nyata pada perangkat keras grafis.*Transaksi ACM pada Grafik (TOG), 2008.**27**(5): hal. 126.

[56] Tersedia dari: https://github.com/mamahfouz66/EKNN_Ensemble_KNN_Based_Classifier.

[57] Okun, O. dan H. Priisalu,*Kompleksitas kumpulan data dalam klasifikasi kanker berbasis ekspresi gen menggunakan ansambel k-tetangga terdekat.*Kecerdasan buatan dalam kedokteran, 2009.**45**(2): hal. 151-162.

[58] Mahfouz, M.,*Rfknn: Knn Kasar-Fuzzy untuk Klasifikasi Big Data.*Jurnal Internasional Penelitian Lanjutan dalam Ilmu Komputer, 2018.**9**(2): hal. 274-279 DOI: 10.26483/ijarcs.v9i2.5667.