Heliyon 9 (2023) e13633

Daftar isi tersedia diSains Langsung

Heliyon

beranda jurnal:www.cell.com/heliyon

Artikel Penelitian

Algoritma AdaBoost yang ditingkatkan untuk identifikasi kanker paru-paru berdasarkan hidung elektronik

Lijun HaoA,C,1,*, Geng HuangB,A,1

A*Sekolah Ilmu dan Teknik Kesehatan, Universitas Shanghai untuk Sains dan Teknologi, Shanghai, 200093, Cina*

B*Laboratorium Utama Pencitraan Molekuler Shanghai, Rumah Sakit Pusat Distrik Jiading yang Berafiliasi dengan Universitas Kedokteran dan Ilmu Kesehatan Shanghai, Shanghai, 201318, Tiongkok*

C*Perguruan Tinggi Instrumentasi Medis, Universitas Kedokteran dan Ilmu Kesehatan Shanghai, Shanghai, 201318, Cina*

INFO PASAL

Kata kunci:
Hidung elektronik
Kanker paru-paru
Meningkatkan pembelajaran
AdaBoost
Voting validasi silang
K-fold
TIDAK

ABSTRAK

Penelitian ini mengembangkan algoritma pembelajaran peningkatan kecerdasan yang lebih baik berdasarkan AdaBoost, yang dapat diterapkan untuk deteksi napas kanker paru-paru melalui hidung elektronik (eNose). Pertama, mengumpulkan sinyal napas dari relawan melalui eNose, termasuk individu sehat dan orang yang menderita kanker paru-paru. Selain itu, fitur sinyal diekstraksi dan dioptimalkan. Kemudian, sub-klasifikasi multi diperoleh, dan koefisiennya diturunkan dari kesalahan pelatihan. Untuk meningkatkan kinerja generalisasi, validasi silang K-fold digunakan saat membuat setiap sub-klasifikasi. Hasil prediksi subklasifikasi pada set pengujian kemudian dicapai dengan metode voting. Dengan demikian, pengklasifikasi AdaBoost yang ditingkatkan akan dibangun melalui integrasi heterogen. Hasil penelitian menunjukkan bahwa rata-rata presisi pengklasifikasi algoritma yang ditingkatkan untuk membedakan penderita kanker paru dan individu sehat dapat mencapai 98,47%, dengan sensitivitas 98,33% dan spesifisitas 97%. Dan dalam 100 tes independen dan acak, koefisien variasi kinerja pengklasifikasi hampir tidak melebihi 4%. Dibandingkan dengan algoritma terintegrasi lainnya, generalisasi dan stabilitas pengklasifikasi algoritma yang ditingkatkan lebih unggul. Jelas bahwa algoritma AdaBoost yang ditingkatkan dapat membantu menyaring kanker paru-paru secara lebih komprehensif. Selain itu, hal ini akan meningkatkan penggunaan eNose secara signifikan dalam identifikasi dini kanker paru-paru.

1. Perkenalan

Kanker paru-paru saat ini merupakan salah satu kanker paling umum dan mematikan di seluruh dunia. Menurut IARC (Badan Internasional untuk Penelitian Kanker), jumlah kasus baru kanker paru-paru di dunia pada tahun 2020 adalah 2,2 juta, dan jumlah kematian baru akibat kanker paru-paru adalah 1,8 juta [1]. Studi [2–4] telah menunjukkan bahwa tingkat kelangsungan hidup lima tahun pasien kanker paru stadium menengah adalah sekitar 60%, sedangkan pasien kanker paru stadium lanjut bahkan kurang dari 5%. Namun, tingkat kelangsungan hidup lima tahun pasien kanker paru stadium awal dapat meningkat hingga lebih dari 90% setelah pengobatan [5]. Oleh karena itu, diagnosis dini kanker paru sangatlah penting.

Saat ini, terdapat banyak teknik untuk skrining kanker paru-paru, antara lain sinar-X, computerized tomography (CT), positron Emission Tomography (PET), dan Magnetic Resonance Tomography (MRI). Namun, masing-masing memiliki kelemahannya masing-masing, seperti risiko radiasi

* Penulis yang sesuai. Sekolah Ilmu dan Teknik Kesehatan, Universitas Shanghai untuk Sains dan Teknologi, Shanghai, 200093, Cina. *Alamat email:* haolj@sumhs.edu.cn (L.Hao).

[1]Para penulis memberikan kontribusi yang sama terhadap pekerjaan ini.

https://doi.org/10.1016/j.heliyon.2023.e13633
Diterima 26 Agustus 2022; Diterima dalam bentuk revisi 1 Februari 2023; Diterima 6 Februari 2023
Tersedia online 21 Februari 2023

sinar-X; risiko radiasi dan tingginya tingkat positif palsu CT; resolusi spasial MRI yang rendah. Kombinasi teknik diagnosis PET/CT dapat mengkarakterisasi dan menentukan stadium kanker paru-paru, namun mahal dan tidak dapat berperan dalam skrining kanker paru-paru secara dini [6].

eNose adalah instrumen pintar yang dikembangkan dalam beberapa tahun terakhir. Ini dirancang untuk mendeteksi dan membedakan bau kompleks menggunakan serangkaian sensor. Ini adalah metode yang sepenuhnya non-invasif dan hampir tidak terbatas dalam hal frekuensi, akses, dan biaya.7,8]. Karena dapat menghubungkan senyawa organik volatil (VOC) napas tertentu atau jejak napas (yaitu pola VOC) dengan status kesehatan [9]. De Vries dkk. [10] pada tahun 2019 menggunakan eNose untuk menganalisis VOC dari pasien kanker paru-paru guna memprediksi apakah pasien yang menjalani imunoterapi akan mencapai remisi objektif. Keakuratan prediksinya bisa mencapai 85%. Namun, perangkat eNose tidak memberikan informasi tentang komposisi VOC napas tertentu, namun perangkat tersebut mengidentifikasi profil tertentu atau "cetakan bau" dari keseluruhan komposisi napas.

Penelitian telah menunjukkan kegunaan perangkat eNose untuk mendeteksi kanker paru-paru. Berbeda dengan teknik deteksi komposisi gas tradisional, perangkat eNose membuat model diagnostik matematis untuk mendeteksi kanker paru-paru dengan VOC. Untuk secara efektif membedakan respon gas pasien kanker paru-paru dan individu sehat yang terdeteksi oleh eNose, para peneliti telah merancang berbagai algoritma. Huber dkk. [11] pada tahun 2014 menerapkan kombinasi analisis komponen utama (PCA), analisis komponen independen (ICA), dan analisis regresi logistik (LR), yang dapat membedakan pasien kanker paru-paru dengan orang sehat dengan akurasi 80% tetapi spesifisitas hanya 48%. Chen lu dkk. [12] pada tahun 2015 menerapkan analisis regresi logistik untuk secara efektif membedakan dua jenis sampel napas dengan akurasi masing-masing 80,6% dan spesifisitas 74%. Dekel Shlomi dkk. [13] 2017 menerapkan support vector machine (SVM) untuk diferensiasi dua jenis sampel napas dengan akurasi 79,1% dan spesifisitas hingga 88,9%. Maribel dkk. [14] pada tahun 2019 menerapkan metode diskriminan PCA dan Fisher untuk membedakan pasien kanker paru-paru dari individu sehat dan akurasi serta spesifisitasnya dapat ditingkatkan menjadi 82,2% dan 91%. Namun, kinerja algoritma ini belum cukup untuk memenuhi kebutuhan klinis. Selain itu, pengklasifikasi ini didasarkan pada sampel kecil. Namun, kinerja generalisasinya belum diuji.

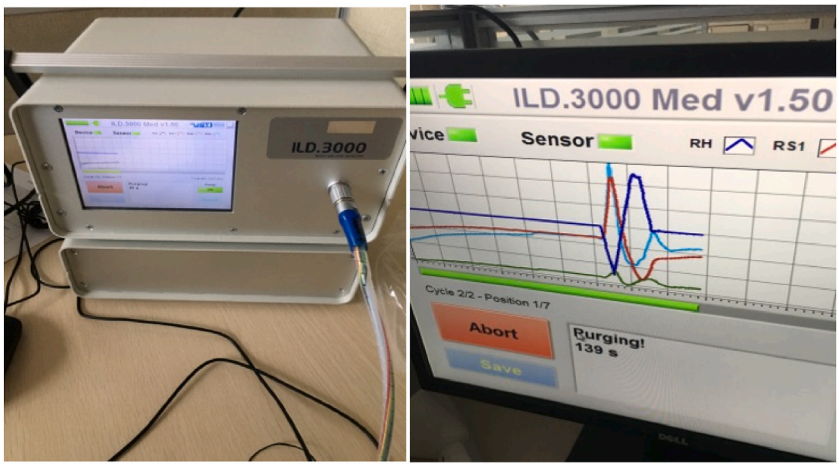
Dalam makalah tersebut, algoritma AdaBoost (ImAdaBoost) yang ditingkatkan telah diusulkan untuk membangun pengklasifikasi yang dapat membedakan nafas pasien kanker paru-paru dan individu sehat. Berdasarkan algoritma AdaBoost tradisional, kami menerapkan validasi silang K-fold dan metode pemungutan suara pada desain sub-klasifikasi yang inovatif [15]. Dan pengklasifikasi yang terintegrasi dan ditingkatkan dibentuk dengan memberi bobot pada beberapa sub-pengklasifikasi yang heterogen [16,17]. Kemudian, kami merancang eksperimen untuk membandingkan kinerja algoritma yang ditingkatkan dengan algoritma lainnya. Selanjutnya, eksperimen dirancang untuk menguji stabilitas dan kinerja generalisasi dari algoritma yang ditingkatkan. Melalui banyak pengujian acak, kurva fluktuasi kinerja dianalisis, dan stabilitas serta kinerja generalisasi algoritme diuji [18]. Akhirnya, kumpulan sampel pengujian baru dikumpulkan dan diprediksi oleh algoritma yang diusulkan di makalah.

2. Bahan dan metode

2.1. Perangkat eNose

Dalam penelitian ini, perangkat eNose (ditunjukkan padaGambar 1) yang kami terapkan adalah peralatan komersial yang cocok untuk deteksi gas umum. Perangkat (ILD.3000, UST Sensors GmbH Company, Jerman) dilengkapi dengan tiga sensor elektrokimia, semikonduktor oksida logam, dan sensor suhu yang dapat dikontrol [19]. Gambar (a) adalah sistem perangkat keras perangkat dan Gambar (b) menunjukkan antarmuka kumpulan perangkat.

Ketiga sensor gas pada perangkat tersebut merupakan inti dari perangkat tersebut. Sensor tersebut adalah sensor seri GGS1000, yang sensitif terhadap gas yang mudah terbakar; sensor seri GGS3000 yang dapat mendeteksi hidrokarbon khususnya untuk C1, C2⋯...C8; dan sensor seri GGS7000, yang dapat mendeteksi NO_2[20]. Sensor suhu yang dapat dikontrol Rt dirancang untuk menyediakan lingkungan suhu yang sesuai untuk meningkatkan



Gambar 1.Perangkat eNose.

kemampuan respon sensor terhadap gas. Kisaran variasi suhu adalah dari 200 ◦ C hingga 400 ◦ C [19].

Seluruh proses pengambilan data nafas seorang relawan dengan perangkat eNose melalui 5 tahapan, antara lain pembilasan sensor, pengukuran pasien, dan lain sebagainya. Setelah sistem melakukan pemanasan, total waktu adalah 17 menit 50 detik. Selama proses pengumpulan, hanya corong sekali pakai yang digunakan dan tidak ada alat intervensi yang digunakan, sehingga tidak membahayakan tubuh manusia.

2.2. Akuisisi data dan pra-pemrosesan

Kumpulan data pelatihan merupakan data napas 142 relawan termasuk 91 pasien kanker paru-paru dan 51 orang sehat. Informasi semua relawan dicatat secara anonim. Sampel tes yang baru dikumpulkan mencakup data napas 12 pasien kanker paru-paru dan 10 orang sehat.

Kriteria inklusi relawan kedua kelompok terdiri dari: (1) individu>18 tahun, mampu memahami dan membaca formulir persetujuan; (2) pasien yang menderita kanker paru primer, tidak ada bukti lain adanya kanker metastasis; (3) tidak ada riwayat merokok dan penyalahgunaan alkohol dalam tiga bulan terakhir; (4) dalam keadaan puasa. Informasi dasar tentang relawan ditampilkan diTabel 1.

Penelitian ini disetujui oleh komite etik Rumah Sakit Shanghai Changzheng (Nomor berkas persetujuan 2018SL029). Semua sukarelawan diberitahu tentang tujuan penelitian. Instruksi diberikan dan persetujuan lisan diperoleh dari masing-masing sukarelawan sebelum pengumpulan data pernapasan.

Ketiga kurva seperti yang ditunjukkan padaGambar 2adalah tiga sinyal respons gas yang dihembuskan secara bersamaan yang dikumpulkan oleh perangkat eNose, yang dapat dicatat sebagai $kamu_A$,$kamu_B$, Dan$kamu_C$.

Seperti yang ditunjukkan diGambar 2, sinyal yang dikumpulkan oleh sensor yang berbeda sangat bervariasi, dan sinyal yang dikumpulkan oleh sensor yang sama juga sangat bervariasi. Untuk memfasilitasi perbandingan data dan analisis statistik, semua sinyal dinormalisasi di sini. Dalam normalisasi,maks($kamu_{aku}$)Danmenit ($kamu_{aku}$) yang digunakan masing-masing diambil dari nilai maksimum dan minimum sinyal yang dikumpulkan oleh seluruh sukarelawan di semua sensor. Kemudian, sinyal yang dikumpulkan oleh masing-masing sensor dapat diubah menjadi rentang [0, 1] melalui Persamaan.(1).

$$kamu^{Sayib}_{aku} = \frac{kamu^{Sayib}_{k} - \text{menit}\,(kamu_{aku})}{\text{maks}(kamu_{aku}) - \text{menit}\,(kamu_{aku})} \quad (1)$$

dimana, i dapat diambil sebagai A, B dan C, masing-masing mewakili tiga sensor; i menunjukkan sampel ke-i yang dikumpulkan oleh sensor tertentu. Dan menit ($kamu_{aku}$)Danmaks($kamu_{aku}$)mewakili nilai minimum dan maksimum dari semua sinyal sampel yang dikumpulkan oleh sensor yang sama.

Setelah sinyal dinormalisasi, fitur-fiturnya seperti domain waktu, domain frekuensi, dan statistik diekstraksi lebih lanjut. Dalam penelitian ini, 14 fitur domain waktu (nilai maksimum dan posisi yang sesuai, nilai minimum dan posisi yang sesuai, rata-rata, nilai puncak-ke-puncak, varians, deviasi standar, faktor bentuk gelombang, faktor pulsa, faktor puncak, faktor margin, dan area), 14 fitur domain frekuensi (frekuensi pusat gravitasi, varians frekuensi, selisih rata-rata akar kuadrat, spektrum frekuensi dan spektrum daya yang dihitung dengan berbagai metode) dan 10 fitur statistik (deviasi kutub, median, kuantil, pluralitas, koefisien variasi, skewness, kurtosis , koefisien autokorelasi dan entropi informasi, dan korelasi antara dua sinyal sensor). Dengan menggabungkan semua fitur yang diekstraksi dari tiga sinyal sensor, serangkaian fitur dimensi tinggi (1557) yang sesuai dengan satu sampel dapat diperoleh.

2.3. Pengoptimalan fitur

Untuk menghindari bencana dimensi dan meningkatkan kinerja pengklasifikasi, PCA dan algoritma genetika (GA) diterapkan masing-masing untuk optimasi fitur.

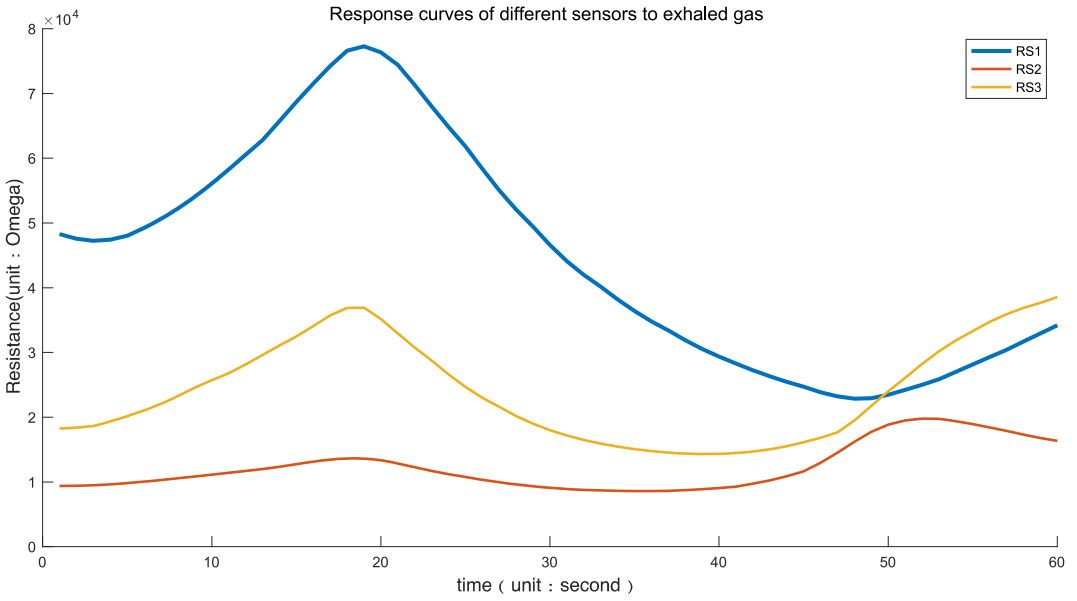
PCA adalah metode reduksi dimensi fitur umum [21]. Tujuannya adalah memetakan data berdimensi tinggi ke ruang berdimensi rendah melalui semacam proyeksi linier, yaitu mengganti n fitur asli dengan jumlah M fitur yang lebih sedikit. Varians data diharapkan menjadi yang terbesar dalam dimensi yang diproyeksikan sehingga fitur M baru tidak terlalu terkait satu sama lain.

GA adalah algoritma optimasi fitur berdasarkan seleksi alam yang muncul dalam beberapa tahun terakhir [22]. Seleksi fitur berdasarkan GA dapat dilakukan melalui empat langkah: pembangkitan populasi awal, seleksi fitur berdasarkan fungsi kebugaran, persilangan, dan mutasi. Dalam penelitian tersebut, fungsi kebugaran dibangun berdasarkan tingkat kesalahan pengklasifikasi Bayesian. Probabilitas crossover ditetapkan ke angka acak, probabilitas mutasi ditetapkan ke 0,5, dan jumlah iterasi ditetapkan ke 20.

Tabel 1
Informasi dasar para relawan.

	Sampel pelatihan			Sampel uji		
	Pria	Perempuan	Usia rata-rata (tahun)±SD	Pria	Perempuan	Usia rata-rata (tahun)±SD
Kanker paru-paru	63	28	55.8±12.1	11	1	62.7±6.6
Kontrol	34	17	51.6±14.6	6	4	55.9±11.2

* Dalam tabel, SD berarti simpangan baku.



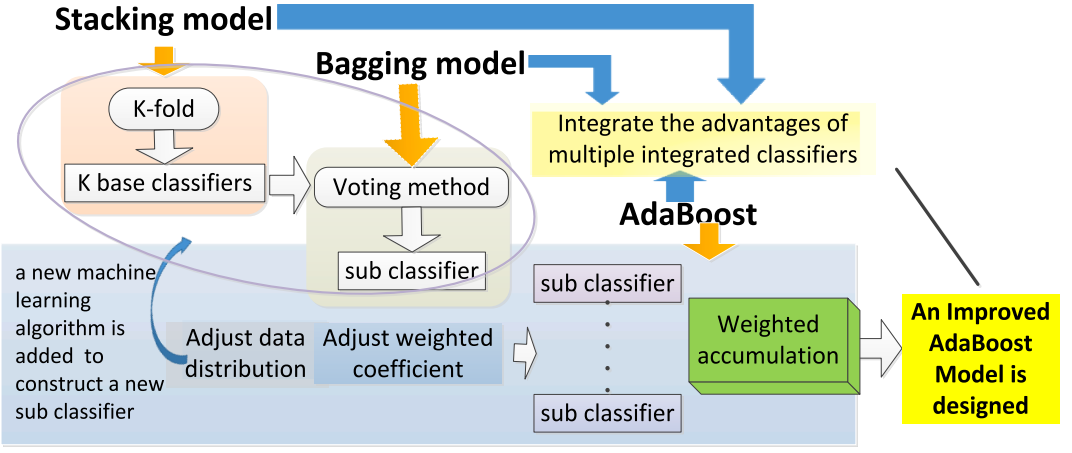
Gambar 2.Sinyal nafas dikumpulkan oleh Sensor.

2.4. Desain pengklasifikasi

Deteksi kanker paru-paru berdasarkan VOC yang dideteksi oleh eNose pada dasarnya adalah masalah klasifikasi. Dalam studi tersebut, model klasifikasi yang sangat kuat pertama kali dibangun berdasarkan sinyal ekspirasi pasien kanker paru-paru dan individu sehat. Ketika sampel baru diperoleh, sampel tersebut dapat dimasukkan sebagai masukan ke pengklasifikasi untuk menentukan apakah sampel tersebut merupakan pasien kanker paru-paru atau bukan.

Algoritma pembelajaran ansambel merupakan algoritma kombinatorial [23]. Algoritme pertama-tama menyusun serangkaian sub-pengklasifikasi (peserta didik yang lemah) dan kemudian menggabungkan pembelajar yang lemah dengan menggunakan strategi berbeda untuk membuat prediksi keseluruhan. Tingkat kesalahan model agregat ini akan berkurang. Selain itu, karena adanya saling penghambatan antar model, kinerja generalisasi ditingkatkan dan fenomena overfitting dapat dihindari.

Saat ini, algoritme integrasi yang representatif mencakup bagging, stacking, dan boosting. Algoritme bagging membangun beberapa pengklasifikasi berdasarkan sampel pelatihan yang berbeda untuk masing-masing memprediksi sampel baru dan memperoleh hasil prediksi akhir dengan pemungutan suara [24]. Model susun adalah kerangka model terintegrasi hierarki dua lapis. Ide dasarnya adalah menggabungkan hasil prediksi beberapa model tunggal menjadi satu model, untuk mengurangi kesalahan generalisasi pada satu model [25]. Pada lapisan pertama, validasi silang digunakan untuk memprediksi dan menghasilkan set pelatihan baru serta set pengujian baru. Pada lapisan kedua, set pelatihan baru digunakan untuk membangun pengklasifikasi, dan set pengujian baru diprediksi untuk mendapatkan hasil prediksi akhir. Algoritma peningkatan yang representatif adalah algoritma AdaBoost. Algoritme pertama-tama melatih subklasifikasi dan menyesuaikan distribusi data berdasarkan kesalahan pelatihan. Kemudian subklasifikasi baru dibangun berdasarkan distribusi data baru. Ulangi proses ini hingga jumlah subklasifikasi yang diperlukan tercapai



Gambar 3.Desain model klasifikasi ImAdaBoost berdasarkan beberapa model terintegrasi.

diperoleh. Terakhir, subklasifikasi ini diberi bobot dan digabungkan untuk mendapatkan hasil prediksi akhir. Ketiga algoritma tersebut mempunyai kelebihan dan kekurangan masing-masing. Algoritme bagging terutama digunakan untuk meningkatkan kinerja generalisasi dan memecahkan masalah overfitting. Pada model penumpukan, karena data pelatihan yang digunakan pada kedua lapisan berbeda, maka hasilnya lebih kuat. Sebaliknya, algoritma AdaBoost dapat meningkatkan akurasi pelatihan dan mengurangi underfitting.

Untuk sampel kecil, untuk mendapatkan pengklasifikasi presisi tinggi dengan kinerja generalisasi tinggi, kami mencoba menggabungkan tiga pengklasifikasi terintegrasi yang umum digunakan dalam penelitian kami.

Seperti yang ditunjukkan diGambar 3, Pertama, validasi silang K-fold digunakan untuk melatih beberapa pembelajar dasar, dan kemudian sub-pengklasifikasi diperoleh berdasarkan metode pemungutan suara. Selanjutnya koefisien subklasifikasi diperoleh berdasarkan teori AdaBoost, dan distribusi data sampel pelatihan disesuaikan. Dalam putaran pelatihan baru, algoritma pembelajaran mesin baru akan ditambahkan untuk melatih sub-pengklasifikasi baru. Setelah mencapai waktu pelatihan yang telah ditentukan, pelatihan dihentikan, dan semua sub-pengklasifikasi diberi bobot dan digabungkan untuk mendapatkan hasil prediksi akhir.

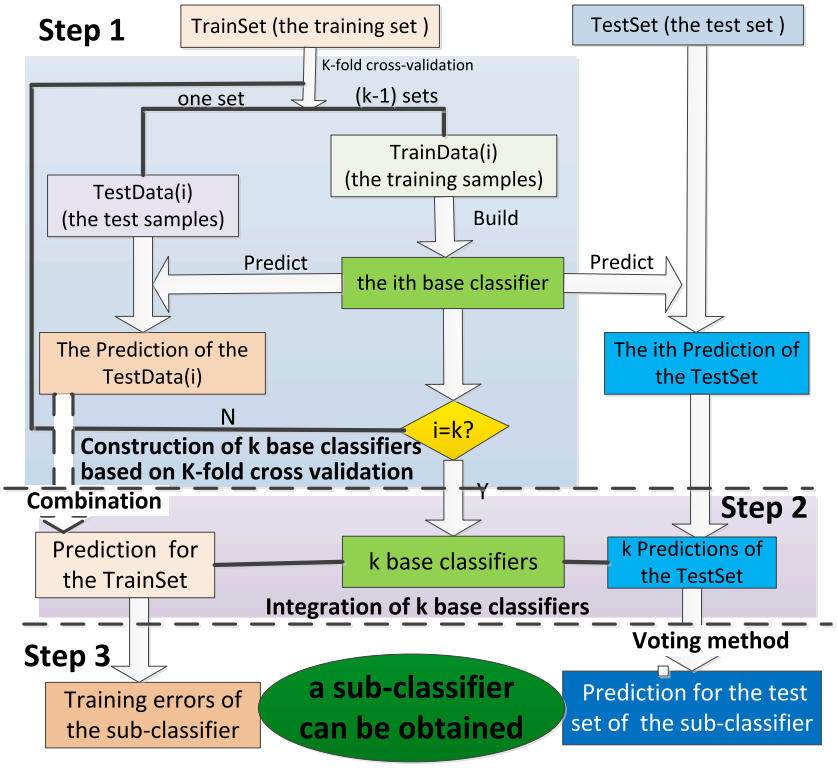
3. Teori

Ide dari algoritma AdaBoost adalah untuk menggabungkan keluaran dari beberapa pengklasifikasi "lemah" (sub-pengklasifikasi) dengan cara yang berbobot untuk menghasilkan klasifikasi yang efektif. Kemampuan beradaptasinya terletak pada sampel yang salah diklasifikasikan oleh subklasifikasi sebelumnya akan diperkuat dengan bobot yang lebih tinggi, dan sampel berbobot akan digunakan untuk melatih kembali subklasifikasi berikutnya. Namun, dalam makalah ini, desain sub-klasifikasi telah diperbaiki berdasarkan validasi silang, metode pemungutan suara, dan beberapa algoritma klasifikasi yang berbeda.

3.1. Peningkatan desain sub-klasifikasi

Seperti yang ditunjukkan diGambar 4, pembangunan subklasifikasi berdasarkan metode validasi silang K-fold terdiri dari tiga langkah khusus. Pada gambar, set pelatihan dilambangkan sebagai 'TrainSet' dan set pengujian dilambangkan sebagai 'TestSet'.

Pada langkah pertama, validasi silang K-fold diterapkan pada data pelatihan ('TrainSet') dan k pengklasifikasi dasar diperoleh. Pertama, bagi 'TrainSet' menjadi k set secara acak. Kemudian pilih satu kelompok sebagai sampel uji "TestData" dan kelompok lainnya (k-1) sebagai sampel pelatihan "TrainData". Selanjutnya, berdasarkan algoritma klasifikasi, dapat diperoleh sekelompok pengklasifikasi berdasarkan validasi silang K-fold. Pada saat yang sama, pengklasifikasi dasar dapat memprediksi prediksi sampel pengujian "TestData" dan set pengujian "TestSet". Akhirnya, diperoleh k pengklasifikasi dasar yang berbeda. Dan prediksi dari k kumpulan sampel uji "TestData" yang berbeda akan diperoleh oleh pengklasifikasi yang berbeda ini. Faktanya, k sampel pengujian yang berbeda ini merupakan set pelatihan. Pada langkah kedua, gabungkan prediktif



Gambar 4.Konstruksi subklasifikasi berdasarkan validasi silang K-fold dan metode voting.

nilai set pengujian untuk mendapatkan nilai prediksi sampel pelatihan. Untuk test set “TestSet”, kita juga akan mendapatkan hasil prediksi k-group. Pada langkah ketiga, kita dapat menentukan hasil prediksi akhir dari subklasifikasi ke set pengujian dengan cara voting. Pada saat yang sama, kesalahan prediksi subklasifikasi ke sampel pelatihan juga dapat diperoleh. Dengan demikian, sub-klasifikasi dibuat.

Mengikuti proses yang dijelaskan di atas, lebih banyak sub-klasifikasi akan dirancang. Sementara itu, untuk menggabungkan keunggulan beberapa algoritme klasifikasi, algoritme yang berbeda dapat dipilih saat membangun subklasifikasi yang berbeda. Dengan demikian, sejumlah subklasifikasi heterogen dapat dibangun secara berurutan.

3.2. Membangun pengklasifikasi terintegrasi berdasarkan AdaBoost

Inti dari algoritma AdaBoost adalah kombinasi tertimbang dari beberapa sub-klasifikasi [25]. Di bagian3.1, kami mendapatkan kesalahan sampel pelatihan dari setiap sub-pengklasifikasi. Dan berdasarkan kesalahan pelatihan, dapat diperoleh koefisien pembobotan subklasifikasi dan bobot sampel akan disesuaikan. Pengklasifikasi yang ditingkatkan akhirnya diperoleh dengan memberi bobot pada kumpulan sub-pengklasifikasi ini.

Dibandingkan dengan algoritme AdaBoost tradisional, hasil subklasifikasi lebih rumit untuk diperoleh dan perlu diperoleh dari beberapa pengklasifikasi dasar melalui pemungutan suara. Langkah-langkah intinya adalah sebagai berikut.

Langkah 1, pilih algoritme pembelajaran mesin dan bagi set pelatihan dengan K-fold seperti yang dijelaskan pada 3.1. Satu lipatan data dipilih sebagai sampel uji secara bergantian, dan k sesi pelatihan dilakukan untuk mendapatkan k pengklasifikasi dasar yang berbeda berdasarkan algoritma yang dipilih satu per satu, membentuk kelompok pengklasifikasi dasar, yang dicatat sebagai sub-ke-i. penggolong.

Nilai prediksi$G_{Saya}(J)$dari kelompok sub-klasifikasi untuk set pelatihan diperoleh dengan Persamaan.(2).

$$G_{Saya}(J) = [G_{1_{Saya}}, G_{2Saya}\cdots, G_{k_{Saya}}] \quad (Saya=1\cdots T,\ j=1\cdots M) \tag{2}$$

dimana, k adalah jumlah pengklasifikasi dasar.$G_{1\ Saya}, G_{Saya}\cdots, G_{kSaya}$adalah prediksi pengklasifikasi dasar k untuk satu lipatan data di set pelatihan, masing-masing, yang digabungkan untuk membentuk subklasifikasi ke-i untuk semua sampel di set pelatihan.*T*adalah jumlah subklasifikasi. Dan m adalah jumlah sampel pelatihan.

Tingkat kesalahan pelatihane_{Saya}sesuai dengan kelompok pengklasifikasi dasar kemudian dapat dihitung dengan Persamaan.(3).

$$e_{Saya} = \sum_{k} D_{J_{saya}(k)} \quad k = 1,2,\cdots,m\left(g_{Saya}(J) / \right)_{=kamu_i} \tag{3}$$

dimana, k melintasi semua sampel di set pelatihan yang nilai prediksinya tidak sesuai dengan nilai sebenarnya.$D_{J\ Saya}(k)$adalah distribusinya koefisien sampel k dalam proses pembuatan subklasifikasi ke-i. Penilai awalnya adalah$D_{J\ Saya=M}$. Untuk klasifikasi biner masalahnya, tingkat kesalahan pada dasarnya adalah jumlah bobot sampel ini.

Langkah 2, hitung koefisien bobotα_{Saya}dari subklasifikasi ke-i menggunakan fungsi eksponensial sebagai fungsi kerugian dengan Persamaan. (4) [26].

$$\alpha_{Saya} = \text{catatan}\left(\frac{1}{2}\frac{1-e_{Saya}}{e_{Saya}}\right) \tag{4}$$

Selanjutnya, gunakan Persamaan.(5)untuk menyesuaikan bobot sampel$D_{J\ Saya+1}$untuk melatih subklasifikasi ke-(i+1).

$$\begin{cases} D_{J\ Saya+1} = D_{J_{saya}} e^{-\alpha_{Saya} kamu_i G_{Saya}(J)} / jumlah \\ jumlah = \sum_{j=1}^{M} D_{J_{Saya+1}} \end{cases} \tag{5}$$

Di mana,$D_{J\ Saya}$adalah bobot sampel set pelatihan yang sesuai dengan kelompok pengklasifikasi dasar ke-i, sementara$D_{J\ Saya+1}$adalah sampel pelatihan yang disesuaikan bobot yang sesuai dengan subklasifikasi ke-(i+1).α_{Saya}adalah koefisien bobot subklasifikasi ke-i, sedangkan$kamu_J$Dan$G_{Saya}(J)$adalah nilai sebenarnya dan nilai prediksi dari sampel j.*jumlah*adalah faktor normalisasi.

Langkah 3, dapatkan hasil prediksi setiap subklasifikasi untuk set pengujian berdasarkan Persamaan.(6)dengan metode pemungutan suara.

$$H_{Saya}(aku) = \begin{cases} \frac{1}{P}\sum \left(H_{Saya}(J) < 0.5, hal > \frac{k}{2}\right) \\ \frac{1}{k-P}\sum \left(H_{Saya}(J) \geq 0.5, P \leq \frac{k}{2}\right) \end{cases} \tag{6}$$

Di mana,$H_{T\ Saya}(J)$adalah nilai prediksi pengklasifikasi dasar t untuk sampel j yang akan diuji.*P*adalah jumlah pengklasifikasi basis-k yang nilai prediksi sampel yang akan diuji kurang dari 0,5. Jika*P*lebih dari separuh jumlah pengklasifikasi dasar, keluarannya adalah rata-rata dari semua probabilitas yang diprediksi kurang dari 0,5; jika tidak, probabilitas keluaran adalah rata-rata dari semua probabilitas yang diprediksi lebih besar dari atau sama dengan 0,5.

Setelah nilai prediksi sub-pengklasifikasi T untuk sampel pengujian diperoleh, nilai prediksi akhir dari pengklasifikasi dapat dicapai sesuai dengan Persamaan.(7)dengan menimbangnya.

$$(aku) = tanda\left[\sum_{Saya=1}^{T} \alpha_{Saya} \cdot H_{Saya}(aku)\right] \quad aku = 1,2\cdots N, N \text{berarti jumlah set tes} \tag{7}$$

Di mana,$H_{Saya}(aku)$adalah himpunan prediksi untuk himpunan pengujian subklasifikasi ke-i.α_{Saya}adalah koefisien bobot subklasifikasi ke-i.$H(aku)$adalah hasil prediksi dari pengklasifikasi terintegrasi.

4. Eksperimen dan hasil

Mengidentifikasi pasien kanker paru-paru melalui pernafasan pada dasarnya adalah masalah klasifikasi. Dalam penelitian ini, kami berharap dapat membedakan pasien kanker paru-paru dari orang sehat dengan membuat model. Oleh karena itu, kelompok kanker paru-paru dan kelompok sehat masing-masing diberi label 1 dan 0.

Untuk mengevaluasi secara kuantitatif penerapan algoritma perbaikan yang diusulkan dalam makalah ini untuk deteksi kanker paru-paru, serangkaian tes dan eksperimen dirancang.

Pertama, kinerja algoritma ini diverifikasi oleh kumpulan data publik kardiovaskular.

Kemudian, sampel pelatihan pernapasan dikumpulkan dan diproses sebelumnya (termasuk 91 pasien kanker paru-paru dan 51 kontrol). Dan fitur-fiturnya diekstraksi lebih lanjut dari sinyal dan masing-masing dioptimalkan dengan PCA dan GA. Dengan demikian, diperoleh beberapa kumpulan fitur dengan dimensi berbeda.

Selanjutnya sampel pelatihan dibagi secara acak menjadi set pelatihan dan set tes sesuai dengan rasio tertentu. Pengklasifikasi dibuat menggunakan set pelatihan. Dalam studi tersebut, enam algoritma berbeda, seperti SVM, metode k-Nearest Neighbor (KNN), random forest (RF), LR, linear discriminant analysis (LDA), dan back propagation neural network (BPNN), dipilih untuk merancang subklasifikasi. . Selain itu, validasi silang lima kali lipat digunakan untuk membangun sub-klasifikasi. Metode spesifiknya dijelaskan di Bagian3.

Untuk mendapatkan hasil evaluasi yang lebih obyektif, sampel pelatihan dibagi secara acak beberapa kali. Setiap divisi relatif independen. Pengklasifikasi yang berbeda dibangun berdasarkan set pelatihan yang berbeda. Dan kinerja rata-rata dari pengklasifikasi yang berbeda ini diambil sebagai indeks evaluasi.

Terakhir, kumpulkan satu set sampel uji baru. Membuat pengklasifikasi menggunakan set pelatihan asli dan membuat prediksi terhadapnya. Jalankan sebanyak 20 kali dan ambil nilai rata-rata sebagai kinerja prediksi.

4.1. Pengujian algoritma ImAdaBoost

Algoritme yang ditingkatkan yang dirancang dalam makalah ini adalah untuk meningkatkan kinerja deteksi pengklasifikasi dalam sampel kecil. Kinerjanya diuji dan dibandingkan dengan algoritma lain melalui kumpulan data kardiovaskular terbuka. Dalam setiap pengujian, 0,48% data (sekitar 316 buah) diambil secara acak dari kumpulan data kardiovaskular sebagai set pelatihan, dan kemudian 0,02% data (sekitar 130 buah) diambil sebagai set pengujian. Dan set pelatihan dan set pengujian dihasilkan secara acak dan independen satu sama lain. Proses ini diulangi sebanyak 20 kali. Kinerja rata-rata dari algoritma yang berbeda [27–30] berdasarkan kumpulan data publik kardiovaskular yang ditunjukkan pada Meja 2.

Dibandingkan dengan algoritma pembelajaran mendalam yang populer saat ini, kinerja algoritma secara keseluruhan kurang baik. Namun karena kecepatan kerjanya yang cepat dan persyaratan konfigurasi komputer yang rendah, akan sangat membantu jika digunakan dalam skrining universal untuk membantu identifikasi penyakit dengan cepat. Seperti yang Terlihat DiMeja 2, akurasi, spesifisitas, dan presisi dari algoritma yang ditingkatkan jelas lebih baik daripada algoritma lain dalam sampel kecil.

Setelah memverifikasi keefektifan algoritme, kami selanjutnya menguji performa klasifikasi, performa generalisasi, dan ketahanan algoritme untuk identifikasi kanker paru-paru melalui napas.

4.2. Penentuan dimensi optimasi fitur dan algoritma optimasi

Pemilihan fitur adalah langkah pertama yang penting dalam pembelajaran mesin. Ini dapat mengoptimalkan efek dan kinerja model klasifikasi. Oleh karena itu, kami terlebih dahulu menerapkan algoritma PCA dan algoritma GA untuk mengoptimalkan fitur dalam dimensi yang berbeda. Kemudian, dengan membandingkan kinerja rata-rata sebanyak 20 kali, algoritma pengoptimalan fitur dan dimensi pengoptimalan yang sesuai diperoleh

Meja 2
Perbandingan algoritma yang berbeda berdasarkan kumpulan data publik kardiovaskular.

Penggolong	Ketepatan	Kepekaan	Kekhususan	Presisi	Skor F1
ImAdaBoost	71.88	65.92	77.84	75.10	70.07
AdaBoost	66,88	66.92	66.84	67.46	66.71
Mengantongi	64,96	66.30	63.61	64.82	65.21
Menumpuk	69,92	68.53	71.30	70,99	69.42
KNN	52.07	69.69	34.46	51.54	59.18
SVM	69.61	64.61	74.61	72.42	67,94
LDA	54.46	53.23	55.69	55.03	53.26
Federasi Rusia	68.84	68.46	69.23	69.46	68.66
LR	66,65	64.46	68.84	68.03	65,95
BP	51.15	98,84	3.46	68.03	80.4

bertekad. Dalam pengujian tersebut, 20% data (termasuk 10 orang sehat dan 18 penderita kanker paru-paru) diambil sebagai set pengujian dan 80% sisanya sebagai set pelatihan.

Tabel 3 dan 4menunjukkan hasil kinerja rata-rata setelah menggunakan metode berbeda untuk mengoptimalkan fitur ke dalam dimensi berbeda. Dalam tabel, Dim berarti dimensi fitur yang dioptimalkan; akurasi adalah persentase deteksi yang benar; sensitivitas adalah persentase penderita kanker paru yang dapat dideteksi dengan benar; sedangkan spesifisitas adalah persentase individu sehat yang dapat dideteksi dengan benar, dan presisi adalah persentase penilaian yang benar sebagai kanker paru-paru. Dalam mendeteksi penyakit, perlu meningkatkan sensitivitas sekaligus memastikan presisi. Sedangkan skor F1 merupakan rata-rata harmonis antara recall dan presisi, serta memperhitungkan sensitivitas dan presisi sehingga keduanya dapat mencapai level tertinggi pada saat yang bersamaan. Dan AUC adalah area di bawah kurva ROC. Semakin besar AUC, semakin baik efek pengklasifikasiannya [30,31].

Seperti dapat dilihat dari dua tabel, apa pun algoritma optimasinya, AUC pengklasifikasi akan optimal ketika dimensi fitur dioptimalkan menjadi 30. Dan jelas bahwa kinerja pengklasifikasi berdasarkan algoritma GA lebih baik dari itu. dari algoritma PCA.

Oleh karena itu, dalam pengujian berikut, kami memilih algoritme GA untuk mengoptimalkan fitur hingga 30 dimensi dan menganalisis lebih lanjut performa lain dari algoritme yang ditingkatkan.

4.3. Penentuan jumlah sampel pelatihan

Algoritma ImAdaBoost telah terbukti dapat diterapkan pada konstruksi pengklasifikasi dengan set pelatihan sampel kecil. Dalam studi tersebut, kami terus menyesuaikan rasio sampel pelatihan dan sampel pengujian, dengan harapan menemukan jumlah sampel pelatihan yang optimal. Selain itu, perubahan kinerja pengklasifikasi dapat diuji ketika jumlah sampel yang terdeteksi meningkat.

Dalam percobaan, rasio pembagian acak sampel pelatihan terus disesuaikan dan diuji secara terpisah. Rasio set pelatihan dan set pengujian ditetapkan masing-masing sebagai 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, dan 9:1. . Dan kemudian 20 tes dilakukan dalam setiap kasus. Kinerja rata-rata dibandingkanGambar 5.

Seperti yang ditunjukkan diGambar 5. Terlihat bahwa kinerja pengklasifikasi secara bertahap menjadi lebih baik seiring dengan bertambahnya jumlah sampel dalam set pelatihan. Ketika sampel pengujian melebihi sampel pelatihan, terjadi penurunan kinerja pengklasifikasi yang signifikan seiring dengan peningkatan sampel pengujian. Khususnya, spesifisitas menurun paling signifikan, dengan perubahan sekitar 30%. Namun, ketika sampel pelatihan menyumbang 60% atau lebih dari total sampel, kinerja pengujian pengklasifikasi relatif stabil. Namun, ketika sampel pelatihan menyumbang 90%, seluruh performa mencapai 100%, yang pada titik ini mungkin terjadi overfitting.

Oleh karena itu, dalam evaluasi kinerja pengklasifikasi selanjutnya, kami mengambil 80% sampel untuk sampel pelatihan dan 20% untuk sampel pengujian.

4.4. Kinerja klasifikasi

Selanjutnya, kami membandingkan kinerja algoritma AdaBoost yang ditingkatkan dengan algoritma klasifikasi terintegrasi lainnya. Algoritme terintegrasi mencakup model Penumpukan dan mode Bagging, seperti yang ditunjukkan pada gambarGambar. 6Dan7.

Gambar 6adalah perbandingan kinerja rata-rata 20 kali lipat dari empat pengklasifikasi terintegrasi. Di antara kelima indikator tersebut, algoritma yang ditingkatkan memiliki akurasi, sensitivitas, dan skor F1 terbaik. Masing-masing sebesar 97,85%, 98,33%, dan 98,34%.Gambar 7adalah kurva ROC dari empat algoritma terintegrasi [31]. Dan rata-rata AUC dari empat algoritma terintegrasi (ImAdaBoost, AdaBoost tradisional, model Stacking, dan model Bagging) adalah 0.996,0.989,0.992, dan 0.988. Meskipun presisi algoritma yang ditingkatkan sedikit lebih rendah dibandingkan model penumpukan, sensitivitasnya jauh lebih tinggi dibandingkan model Penumpukan. Dalam penelitian ini, kami berharap dapat mendeteksi sebanyak mungkin kanker paru-paru, yang berarti sensitivitas algoritme harus tinggi. Namun, pada saat yang sama, kami tidak ingin spesifisitas algoritme menjadi terlalu rendah sehingga menimbulkan kepanikan yang tidak perlu. Oleh karena itu, kami mencoba meningkatkan sensitivitas sebanyak mungkin sambil memastikan presisi. Perbandingan tersebut menunjukkan bahwa kinerja algoritma yang ditingkatkan sudah optimal.

4.5. Kinerja dan ketahanan generalisasi

Selain itu, mengevaluasi apakah suatu algoritma efektif tidak hanya bergantung pada kinerja tetapi juga pada kinerja generalisasi dan ketahanannya. Dalam studi tersebut, stabilitas dan kinerja generalisasi pengklasifikasi diuji berdasarkan fluktuasi perubahan kinerja dalam 100 pengujian. Dalam setiap pengujian, sampel pelatihan dibagi secara acak dan independen.

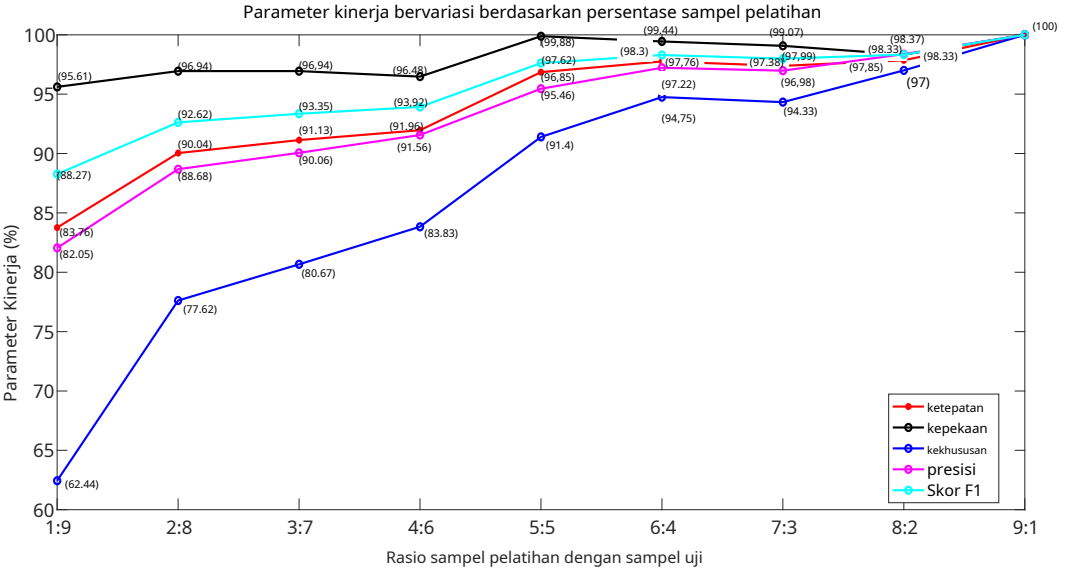
Kami menganalisis volatilitas kinerja pengklasifikasi dengan menghitung koefisien variasi setiap kinerja selama 100 pengujian. Jika koefisien variasi lebih besar dari 15%, hal ini menunjukkan bahwa kinerja pengklasifikasi tidak stabil [32]. Itu

Tabel 3
Perbandingan kinerja rata-rata algoritma ImAdaBoost berdasarkan PCA.

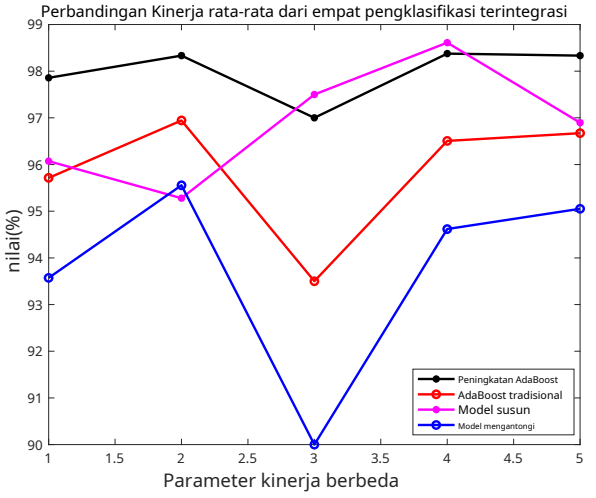
Redup	Ketepatan	Kepekaan	Kekhususan	Presisi	Skor F1	AUC
10	74.64	77.22	70	82,74	79.44	0,845
20	68.57	63,88	77	85.8	70.51	0,855
30	76.07	70	87	91.26	78.29	0,901
40	77,85	78,88	76	85.57	81.72	0,857

Tabel 4
Perbandingan kinerja rata-rata algoritma ImAdaBoost berdasarkan GA.

Redup	Ketepatan	Kepekaan	Kekhususan	Presisi	Skor F1	AUC
10	91.42	93,88	87	92,99	93.37	0,926
20	91.42	92,77	89	94.17	93.25	0,958
30	97,86	98.33	97	98.47	98.34	0,996
40	92.14	98,88	80	90.09	94.25	0,95



Gambar 5.Performa pengklasifikasi dengan rasio pembagian berbeda untuk data pelatihan.



Gambar 6.Perbandingan kinerja rata-rata dari empat klasifikasi terintegrasi rs.

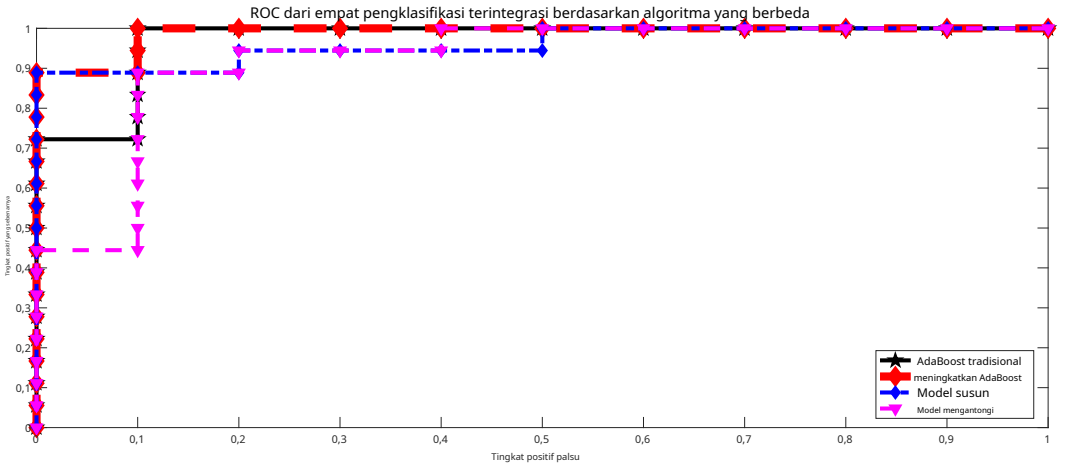
semakin kecil koefisien variasi, semakin stabil kinerja pengklasifikasi, dan taruhan serta ketahanan pengklasifikasi [33].

Dalam analisis bagian sebelumnya, ditemukan bahwa kinerja alg yang ditingkatkan jauh lebih baik dibandingkan pengklasifikasi terintegrasi lainnya.Tabel 5menunjukkan perbandingan pengujian koefisien 100 ketika kedua pengklasifikasi diterapkan secara terpisah. Dengan membandingkan koefisien-koefisien tersebut, dapat disimpulkan bahwa algoritma ImAdaBoost memiliki kinerja yang lebih stabil. Kinerja generalisasi dan ketahanannya lebih baik dibandingkan model susun.

ter kinerja generalisasi

oritma dan model susunnya

Banyak variasi kinerja pada suara yang diklasifikasi berdasarkan es dari algoritma yang ditingkatkan



Gambar 7.ROC dari empat pengklasifikasi terintegrasi.

Tabel 5
Koefisien variasi setiap parameter kinerja pada pengklasifikasi berbeda.

penggolong	Ketepatan	Kepekaan	Kekhususan	Presisi	Skor F1	AUC
ImAdaBoost	2.657	1.891	6.483	3.269	2.005	0,35
Model susun	2.803	1.827	7.723	3.781	2.078	0,32

4.6. Prediksi pengklasifikasi untuk tes baru

Untuk menguji pengklasifikasi dalam praktiknya, serangkaian sampel uji baru dikumpulkan. Sampel terdiri dari 9 orang sehat dan 12 orang penderita kanker paru. Pengklasifikasi dibuat secara acak menggunakan 80% data pelatihan sebagai set pelatihan. Untuk mendapatkan hasil yang lebih obyektif, dilakukan 20 tes secara acak. Sampel pelatihan diekstraksi secara acak dan independen dalam beberapa pengujian. Rata-rata kinerja diskriminatif klasifikasi ditunjukkan padaTabel 6. Statistik positif palsu dan negatif palsu dalam 20 tes diberikanGambar 8.

Seperti yang bisa dilihat dariGambar 8(a), jumlah negatif palsu yang terjadi lebih kecil dibandingkan dengan algoritma AdaBoost tradisional dan model bagging. Kemampuan algoritma yang ditingkatkan untuk mendeteksi kanker paru-paru sebanding dengan algoritma penumpukan. Namun, frekuensi positif palsu lebih tinggi dibandingkan algoritma AdaBoost tradisional dan model tumpukan, seperti yang ditunjukkan pada Gambar 8(B). Algoritme yang ditingkatkan memiliki kemampuan yang lebih lemah untuk mendeteksi individu yang sehat. Namun, menganalisis perubahan fluktuasi AUC, seperti yang ditunjukkan padaGambar 9, dapat diketahui bahwa kinerja algoritma yang ditingkatkan lebih stabil, dengan koefisien variasi sebesar 1,72.

Mungkin karena sedikitnya jumlah sampel uji, keuntungan dari algoritma yang ditingkatkan dalam mengidentifikasi kanker paru-paru tidak tercermin. Pada pekerjaan selanjutnya, kami akan menguji kinerja algoritma yang ditingkatkan dengan memperoleh lebih banyak sampel.

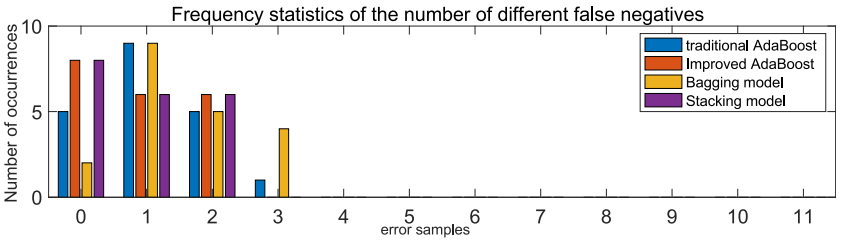
5. Diskusi

Perangkat eNose merupakan instrumen diagnostik baru yang dikembangkan dalam beberapa tahun terakhir, yang dapat diterapkan di bidang pengujian makanan, pemantauan lingkungan, dan diagnosis medis. Perangkat ini memantau dan mendiagnosis penyakit manusia dengan mengumpulkan VOC dalam napas manusia, yang memiliki keunggulan non-invasif, pengoperasian sederhana, dan biaya pemeriksaan rendah. Namun, karena kompleksitas metabolisme manusia dan keragaman penyakit, serta dampak sampel pelatihan yang berbeda terhadap kinerja pengklasifikasi, eNose belum banyak digunakan secara klinis dalam deteksi penyakit. Untuk meningkatkan kinerja algoritma pendeteksian, peneliti telah berfokus pada tiga aspek: pertama, untuk meningkatkan peralatan akuisisi perangkat keras, menemukan karakteristik gas yang berhubungan dengan penyakit dan pilihan hancurkan sensornya; kedua, mengumpulkan sampel sebanyak mungkin secara terus-menerus untuk dijadikan dasar bagi t dia universalitas algoritma; ini. ketiga, melakukan inovasi dan penyempurnaan algoritma untuk membangun algoritma pendeteksian yang lebih cerdas dan tangguh

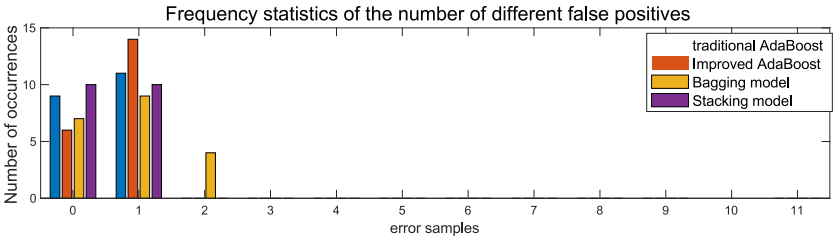
Membedakan dan mengklasifikasikan data nafas pasien kanker paru dan individu sehat adalah t dia pekerjaan inti paru-paru cerdas

Tabel 6
Performa rata-rata dalam 20 tes (%).

ketepatan	Kepekaan	Kekhususan	Presisi	Skor F1	AUC
92.38	92.5	92.22	94.27	93.17	0,984

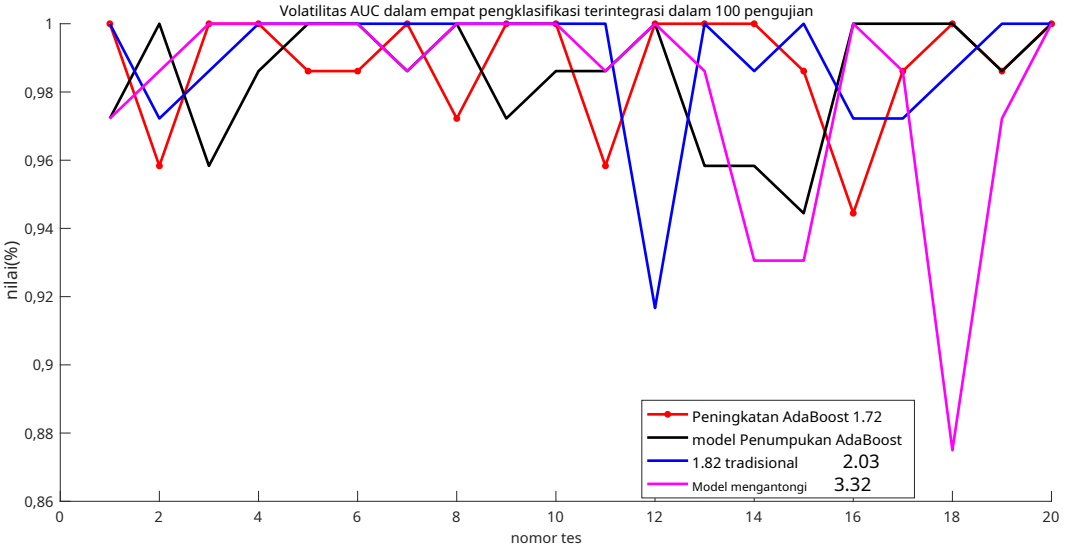


(a) Frequency statistics of the false negatives number



(b) Frequency statistics of the false positives number

Gambar 8.Statistik jumlah kesalahan deteksi untuk sampel baru.



Gambar 9.Fluktuasi AUC di 20 baru tes sampel.

deteksi kanker. Analisis mekanisme metabolisme manusia menunjukkan bahwa paru-paru bersifat patologis [34]. Namun, hubungan informasi ini, yaitu penyakit saya, tidak terlalu jelas. Oleh karena itu, bukanlah tugas yang mudah untuk memanfaatkan intrinsi membangun model diskriminatif yang sesuai. Akuisisi data besar sa pengklasifikasi, tetapi itu tidak mudah. Ketika jumlah sampel sudah pasti, penting untuk mengekstrak lebih banyak fitur dari bentuk gelombang sampel yang dikumpulkan. Namun, terlalu banyak fitur juga dapat menyebabkan penurunan akurasi klasifikasi dan memperlambat komputasi, sehingga optimalisasi fitur juga diperlukan. Dalam studi tersebut, waktu proses pengklasifikasi meningkat pesat setelah pengoptimalan fitur sampel dibandingkan dengan kumpulan fitur asli. Waktu untuk 20 kali lari dikurangi dari lebih dari 2 jam menjadi sekitar 7 menit. Selain itu juga dapat diketahui dari hasil percobaan pada Bagian4.2yang mengubah algoritma optimasi fitur dan dimensi optimasi res menghasilkan perbedaan yang signifikan dalam kinerja yang Clas pengukur. Namun , hanya dua algoritma optimasi fitur, PCA dan GA, Kami terapkan di makalah. Apakah ada lebih banyak konve fitur ini tidak optimal algoritma asi (misalnya algoritma optimasi yang bukan P perlu menjelajahi dimensi fitur yang optimal) ne eds fselidiki segera melakukan asi [35]. SAYA Dalam studi ini, kami memfokuskan pekerjaan pada desain dan konstruksi pengklasifikasi kinerja tinggi.

komposisi gas yang dihembuskan pasien ancer akan berubah karena mplisit dalam sinyal pernafasan yang tidak stasioner, dengan hubungan c antara sinyal ekspirasi dan penyakit dan contoh memfasilitasi konstruksi yang disesuaikan secara universal

Algoritme pembelajaran terintegrasi ImAdaBoost dirancang berdasarkan algoritma AdaBoost tradisional. Secara teori, dalam dua algoritma AdaBoost, sub-pengklasifikasi diberi bobot untuk digabungkan sebagai pengklasifikasi yang ditingkatkan. Namun, konstruksi subklasifikasinya berbeda. Dalam algoritma ImAdaBoost, subklasifikasi diperoleh berdasarkan metode evaluasi dan pemungutan suara 5 kali lipat. Dan dalam algoritma AdaBoost tradisional, sub-klasifikasi dilatih berdasarkan sampel yang diekstraksi dengan metode bootstrap.

Untuk meningkatkan kinerja pengklasifikasi, pertama-tama kami menggunakan ide integrasi untuk mengintegrasikan keunggulan beberapa algoritme pembelajaran mesin. Untuk pengklasifikasi terintegrasi, sub-pengklasifikasi adalah komponen pentingnya. Dalam studi tersebut, subklasifikasi heterogen dibangun menggunakan enam algoritma pembelajaran mesin tradisional. Proses integrasi juga merupakan proses pengurangan tingkat kesalahan secara terus menerus. Hal ini telah diverifikasi dengan kumpulan data publik kardiovaskular. Seperti yang ditunjukkan diMeja 2, akurasi prediksi algoritma ImAdaBoost lebih baik dibandingkan sub-klasifikasinya. Namun, karena hanya 12 atribut yang dikumpulkan dalam kumpulan data publik, hal ini mungkin mengakibatkan kinerja algoritme secara keseluruhan tidak terlalu baik. Namun, dapat ditemukan bahwa sensitivitas dan spesifisitas algoritma ini lebih lemah dibandingkan beberapa algoritma. Namun presisinya tetap optimal.

Dalam mengidentifikasi pasien kanker paru-paru dan individu sehat berdasarkan sinyal pernapasan, algoritme ini memiliki spesifisitas yang agak lebih rendah dibandingkan algoritme penumpukan, namun sensitivitasnya jauh lebih baik. Hasilnya adalah pasien kanker paru-paru lebih kecil kemungkinannya untuk terlewatkan, namun orang yang sehat lebih besar kemungkinannya untuk salah didiagnosis. Sebagai alat skrining tambahan, jelas lebih penting untuk mengurangi diagnosis yang terlewat dan lebih bermakna untuk skrining penyakit. Tentu saja, alasan fenomena ini mungkin juga karena kecilnya ukuran sampel orang sehat. Sampel positif dan negatif tidak seimbang. Dalam penelitian selanjutnya, kami juga akan menguji lebih lanjut gagasan ini dengan meningkatkan jumlah individu yang sehat.

Selain itu, kinerja pengklasifikasi berkaitan erat dengan sampel pelatihan. Mengevaluasi apakah suatu algoritma efektif tidak hanya bergantung pada kinerja tetapi juga pada stabilitas algoritma tersebut [36,37]. Untuk meningkatkan ketahanan algoritma dan meningkatkan kinerja generalisasi pengklasifikasi, pendekatan validasi silang K-fold kemudian digunakan dan beberapa pengklasifikasi dasar dilatih dengan sampel data yang berbeda secara bergantian untuk membangun sub-pengklasifikasi. Kemudian dalam penelitian tersebut, kinerja dan stabilitas algoritma yang ditingkatkan dianalisis secara lebih sistematis dan komprehensif, hal yang jarang terjadi pada penelitian sebelumnya. Berdasarkan hasil 100 percobaan independen acak, stabilitas kinerja pengklasifikasi bahkan lebih baik daripada model susun, seperti yang ditunjukkan padaTabel 5di bagian4.5.

Dalam studi yang tersedia [38], keakuratan penggunaan eNose untuk mengidentifikasi kanker paru-paru sekitar 85%. Sebagai perbandingan, algoritma yang diusulkan dalam makalah ini sangat meningkatkan kemampuan eNose untuk menyaring kanker paru-paru. Namun, kajian dalam makalah ini dapat lebih ditingkatkan, seperti meningkatkan ukuran sampel dan jenis sampel serta mengumpulkan informasi sampel yang lebih komprehensif.

Untuk lebih meningkatkan kinerja algoritma diskriminasi nafas, dalam penelitian selanjutnya, kami akan melakukan penelitian berikut: pertama, melakukan lebih banyak pengumpulan sampel nafas multisenter dan mengurangi dampak ketidakseimbangan sampel; kedua, untuk meningkatkan konstruksi algoritma klasifikasi diskriminasi nafas yang sangat kuat berdasarkan sampel kecil dan untuk meningkatkan kemampuan generalisasi algoritma [15,36]; dan ketiga, mencoba kombinasi deteksi napas dan teknik deteksi lainnya untuk menambang fitur multi-omics guna memberikan fitur sampel yang lebih bermakna untuk algoritme klasifikasi. Selain itu, kami juga akan menerapkan lebih banyak algoritme pengoptimalan fitur, seperti algoritme pengoptimalan fitur Relief, untuk mengoptimalkan fitur guna meningkatkan kinerja pengklasifikasi.

6. Kesimpulan

Dalam makalah tersebut, kami mengusulkan peningkatan algoritme napas kanker paru-paru AdaBoost berdasarkan teori integrasi dan penguatan untuk tugas membedakan sampel napas pasien kanker paru-paru dari orang sehat melalui eNose. Dalam algoritma yang ditingkatkan, sekelompok pengklasifikasi k-base pertama-tama diperoleh dengan validasi silang K-fold berdasarkan sekelompok sampel pelatihan, dan kemudian sub-pengklasifikasi diperoleh lebih lanjut dengan metode pemungutan suara. Ketika setiap subklasifikasi diperoleh, koefisien bobotnya juga akan diperoleh secara sinkron sesuai dengan kesalahan pelatihan, dan distribusi data sampel pelatihan akan disesuaikan. Selain itu, sub-pengklasifikasi yang lebih heterogen akan diperoleh berdasarkan beberapa kelompok pengklasifikasi dasar. Terakhir, sub-pengklasifikasi ini digabungkan untuk mendapatkan pengklasifikasi yang ditingkatkan dan terintegrasi.

Secara umum, algoritme ini tidak hanya meningkatkan kinerja pengklasifikasi dengan mengintegrasikan beberapa pengklasifikasi heterogen, tetapi juga meningkatkan kinerja generalisasi pengklasifikasi dengan validasi silang k-fold. Dibandingkan dengan algoritme tradisional, algoritme ini meningkatkan kinerja pengklasifikasi untuk mengidentifikasi kanker paru-paru dan individu sehat dengan benar. Algoritme yang diusulkan akan membantu mempromosikan metode skrining kanker paru-paru non-invasif melalui eNose, dan mempromosikan penerapan metode pemeriksaan tambahan penyakit klinis berdasarkan eNose. Namun, untuk meningkatkan ketahanan metode, algoritma optimasi fitur dan algoritma klasifikasi cerdas untuk sampel kecil perlu dipelajari lebih lanjut. Untuk memajukan algoritme ini ke dalam aplikasi klinis, lebih banyak sampel dengan kuantitas dan jenis yang lebih banyak perlu dikumpulkan, dan lebih banyak pengujian serta eksperimen konfirmasi perlu dilakukan.

Kontribusi penulis

Para penulis ini memberikan kontribusi yang sama untuk pekerjaan ini.

Hao lijun:Melakukan percobaan; Menganalisis dan menafsirkan data; Menulis makalahnya.

Huang Gang (Rekan penulis pertama):Menyusun dan merancang eksperimen; Materi dan alat analisis yang disumbangkan; Perolehan Pendanaan, dan Pengawasan.

Pernyataan pendanaan

Gang Huang didukung oleh proyek Konstruksi Laboratorium Kunci Pencitraan Molekuler Shanghai [18DZ2260400], Program Utama dari Yayasan Ilmu Pengetahuan Alam Nasional Tiongkok [81830052].

Pernyataan ketersediaan data

Tidak ada.

Pernyataan pernyataan minat

Para penulis menyatakan tidak ada konflik kepentingan.

Lampiran A. Data tambahan

Data tambahan untuk artikel ini dapat ditemukan online dihttps://doi.org/10.1016/j.heliyon.2023.e13633.

Referensi

[1]J. Ferlay, dkk., Statistik kanker untuk tahun 2020: gambaran umum, Int. J.Kanker 149 (4) (2021) 778–789.
[2]D. Sun, dkk., OA08.03 tingkat kelangsungan hidup 5 tahun pasien kanker paru non-sel kecil pasca operasi dengan dua pola tindak lanjut yang berbeda, J. Thorac. Onkol. 16 (10) (2021) S860–S861.
[3]Q. Pei, dkk., Kecerdasan buatan dalam aplikasi klinis untuk kanker paru-paru: diagnosis, pengobatan dan prognosis, Clin. kimia. Laboratorium. medis. 60 (12) (2022) 1974–1983.
[4]M. Zoair, dkk., Nilai (18)F parameter FDG-PET/CT pada tindak lanjut jangka panjang untuk pasien dengan kanker paru-paru non-sel kecil, Innov. Bedah. Sains. 7 (2) (2022) 35–43.
[5]S. Kort, dkk., Studi prospektif multi-pusat tentang diagnosis subtipe kanker paru-paru melalui analisis napas, Kanker Paru 125 (2018) 223–229.
[6]M. Donaghy, S. Stolberg, S. Grundy, PET-CT sebelum biopsi pada jalur diagnostik kanker paru-paru, Kanker Paru 139 (2020).
[7]B. Liu, dkk., Deteksi kanker paru-paru melalui napas melalui hidung elektronik ditingkatkan dengan pendekatan pemilihan fitur kelompok jarang, Sens. Actuators B Chem. 339 (2021).
[8]K. Chen, dkk., Mengenali kanker paru-paru dan stadiumnya menggunakan sistem hidung elektronik yang dikembangkan sendiri, Comput. biologi. medis. 131 (2021), 104294.
[9]W. Biehl, dkk., Pengenalan pola VOC terhadap kanker paru-paru: evaluasi komparatif dari berbagai strategi berbasis anjing dan eNose menggunakan bahan sampel yang berbeda, Acta Oncol. 58 (9) (2019) 1216–1224.
[10]PJ Mazzone, dkk., Diagnosis kanker paru-paru dengan analisis hembusan napas dengan rangkaian sensor kolorimetri, Thorax 62 (7) (2007) 565–568.
[11]AJ Hubers, dkk., Gabungan hipermetilasi dahak dan analisis eNose untuk diagnosis kanker paru-paru, J. Clin. jalan. 67 (8) (2014) 707–711.
[12]L. Chen, dkk., Model prediksi senyawa organik yang mudah menguap dalam napas yang dihembuskan untuk diagnosis kanker paru-paru, Tumor 35 (4) (2015) 404–413.
[13]D. Shlomi, dkk., Deteksi kanker paru-paru dan mutasi EGFR dengan sistem hidung elektronik, J. Thorac. Onkol. 12 (10) (2017) 1544–1551.
[14]M. Rodriguez-Aguilar, dkk., Kromatografi gas ultracepat yang digabungkan dengan hidung elektronik untuk mengidentifikasi biomarker yang mudah menguap dalam napas yang dihembuskan dari pasien penyakit paru obstruktif kronik: studi percontohan, Biomed. Kromatografi. 33 (12) (2019) e4684.
[15]TT Nguyen, dkk., Kerangka pengklasifikasi berganda berbobot berdasarkan proyeksi acak, Inf. Sains. 490 (2019) 36–58.
[16]HBF David, A. Suruliandi, SP Raja, Kerangka kerja bertumpuk untuk ansambel algoritma klasifikasi heterogen, J. Circ. sistem. Hitung. 30 (15) (2021).
[17]W. Wang, D. Sun, Algoritma AdaBoost yang ditingkatkan untuk klasifikasi data tidak seimbang, Inf. Sains. 563 (2021) 358–374.
[18]JH Morra, dkk., Perbandingan AdaBoost dan mesin vektor pendukung untuk mendeteksi penyakit Alzheimer melalui segmentasi hipokampus otomatis, IEEE Trans. medis. Gambar. 29 (1) (2010) 30–43.
[19]A. Voss, dkk., Mendeteksi penggunaan ganja pada permukaan kulit manusia melalui sistem hidung elektronik, Sensor 14 (7) (2014) 13256–13272.
[20]Lj Hao, M. Zhang, G. Huang, Fitur optimalisasi sinyal nafas yang dihembuskan berdasarkan pearson-BPSO, Mobile Inf. sistem. 2021 (2021) 1–9.
[21]J. Fu, dkk., Klasifikasi pola menggunakan model penciuman dengan pemilihan fitur PCA pada hidung elektronik: studi dan penerapan, Sensor 12 (3) (2012) 2818–2830.
[22]MR Gauthama Raman, dkk., Sistem deteksi intrusi yang efisien berdasarkan hipergraf - algoritma genetika untuk optimasi parameter dan pemilihan fitur dalam mesin vektor pendukung, Knowl. Sistem Dasar. 134 (2017) 1–12.
[23]MS Nawaz, B. Shoaib, MA Ashraf, Prediksi penyakit kardiovaskular cerdas yang diberdayakan dengan optimalisasi penurunan gradien, Heliyon 7 (5) (2021), e06948.
[24]MN Uddin, RK Halder, Sistem dinamis multilayer berbasis metode ansambel untuk memprediksi penyakit kardiovaskular menggunakan pendekatan pembelajaran mesin, Inform. medis. Tidak Terkunci 24 (2021).
[25]TT Nguyen, dkk., Sebuah novel 2 tahap yang menggabungkan model pengklasifikasi dengan penumpukan dan pemilihan fitur berbasis algoritma genetika, dalam: Intelligent Computing Methodologies, 2014, hlm. 33–43.
[26]YS Jeon, DH Yang, DJ Lim, FlexBoost- Algoritme peningkatan fleksibel dengan fungsi kerugian adaptif, IEEE Access 7 (2019) 125054–125061.
[27]V. Gupta, M. Mittal, KNN dan PCA classifier dengan pemodelan Autoregresif selama interpretasi sinyal EKG yang berbeda, Procedia Comput. Sains. 125 (2018) 18–24.
[28]E. Kasbohm, dkk., Strategi untuk identifikasi pola terkait penyakit senyawa organik yang mudah menguap: prediksi paratuberkulosis pada model hewan menggunakan hutan acak, J. Breath Res. 11 (4) (2017), 047105.
[29]YY Liu, dkk., Perbandingan regresi logistik, klasifikasi dan pohon regresi, serta model jaringan saraf dalam memprediksi Kekerasan Kembali, J. Quant. Kriminol. 27 (4) (2011) 547–573.
[30]V. Dominic, D. Gupta, S. Khare, Analisis kinerja efektif teknik pembelajaran mesin untuk penyakit kardiovaskular, Appl. medis. Inf. 36 (1) (2015) 23–32.
[31]T. Takenouchi, O. Komori, S. Eguchi, Perpanjangan kurva karakteristik operasi penerima dan klasifikasi optimal AUC, Neural Comput. 24 (10) (2012) 2789–2824.
[32]A. Fawzi, O. Fawzi, P. Frossard, Analisis ketahanan pengklasifikasi terhadap gangguan permusuhan, Mach. Mempelajari. 107 (3) (2017) 481–508.
[33]V. Paliwal, NR Babu, Prediksi batas stabilitas pada milling dengan mempertimbangkan variasi parameter dinamis dan koefisien gaya potong spesifik, Procedia CIRP 99 (2021) 183–188.
[34]X. Chen, dkk., Menghitung indeks senyawa organik volatil (VOC) dalam pernafasan untuk skrining kanker paru-paru dan deteksi dini, Kanker Paru-paru 154 (2021) 197–205.
[35]H. Lu, dkk., Algoritma ansambel hibrid yang menggabungkan AdaBoost dan algoritma genetika untuk klasifikasi kanker dengan data ekspresi gen, IEEE ACM Trans. Hitung. biologi. Bioinf 18 (3) (2021) 863–870.

[36]S. Wu, H. Nagahashi, AdaBoost yang Dihukum: meningkatkan kesalahan generalisasi AdaBoost yang lembut melalui distribusi margin, IEICE Trans. Sistem Info. E98.D (11) (2015) 1906–1915.
[37]A. Mahabub, Teknik canggih untuk mendeteksi berita palsu menggunakan Ensemble Voting Classifier dan perbandingan dengan pengklasifikasi lainnya, SN Appl. Sains. 2 (4) (2020).
[38]R. de Vries, dkk., Prediksi respons terhadap terapi anti-PD-1 pada pasien dengan kanker paru-paru non-sel kecil dengan analisis hidung elektronik pada napas yang dihembuskan, Ann. Onkol. 30 (10) (2019) 1660–1666.

Hao Lijun,kandidat doktor di Universitas Shanghai untuk Sains dan Teknologi dan juga bekerja di Universitas Kedokteran & Ilmu Kesehatan Shanghai. Dia lulus dari Institut Instrumen Biomedis Universitas Shanghai Jiaotong pada tahun 2007. Dia terutama terlibat dalam pengajaran dan penelitian di bidang teknik biomedis. Dalam hal pengajaran, ia berpartisipasi dalam pembangunan Shanghai, kursus berkualitas tinggi, basis pelatihan pendidikan publik Shanghai, dll.; berpartisipasi dalam persiapan dan penyelesaian beberapa buku teks. Dan dalam hal penelitian ilmiah, dia telah memimpin proyek Chenguang dari Komisi Pendidikan Kota Shanghai dan beberapa proyek permulaan penelitian ilmiah tingkat sekolah dan berpartisipasi dalam beberapa proyek Komisi Sains dan Teknologi dan dana ilmu pengetahuan alam. Dalam lima tahun terakhir, ia telah menerbitkan lebih dari sepuluh makalah sebagai penulis pertama, dan mengajukan serta menyetujui banyak paten.

Geng Huang,MD, Profesor kelas dua, dan pengawas doktoral, merangkap sebagai direktur Institut Kedokteran Nuklir Klinis Universitas Shanghai Jiaotong, presiden Perguruan Tinggi Kedokteran Nuklir Asia, ketua Asosiasi Pendidikan Kedokteran Shanghai, dan pemimpin Asosiasi Pendidikan Medis Shanghai. spesialisasi klinis utama nasional kedokteran pencitraan, disiplin ilmu utama Shanghai dan disiplin ilmu kelas satu Shanghai. Dia telah memenangkan para ahli muda dan paruh baya dengan kontribusi luar biasa dari Kementerian Kesehatan, talenta terkemuka Shanghai, talenta terkemuka Shanghai Medical, rencana 100 orang Shanghai, dan gelar lainnya. Dia telah melaksanakan lebih dari 30 proyek seperti National Natural Science Foundation dan proyek-proyek utama, serta proyek “973”. Sejauh ini, Beliau telah menerbitkan lebih dari 200 makalah di jurnal dalam dan luar negeri, termasuk lebih dari 80 makalah yang termasuk dalam SCI atau EI; Lebih dari 10 paten resmi; Dia mengedit lebih dari 10 buku teks dan monografi yang direncanakan untuk perguruan tinggi kedokteran. Selain itu, ia telah memenangkan lebih dari 10 penghargaan, termasuk hadiah kedua Penghargaan Kemajuan Sains dan Teknologi Nasional dan hadiah pertama Penghargaan Sains dan Teknologi Medis Huaxia. Bidang penelitian utamanya adalah penyelidikan molekuler, omics pencitraan tumor, teknik biomedis.