



Daftar isi tersedia diSains Langsung

# Jurnal Informatika Patologi

beranda jurnal:[www.elsevier.com/locate/jpi](http://www.elsevier.com/locate/jpi)

## Paru-paru XML-GBM: Aplikasi berbasis pembelajaran mesin yang dapat dijelaskan untuk diagnosis kanker paru-paru



Sarreha Tasmin Rikta<sup>A</sup>, Khandaker Mohammad Mohi Uddin<sup>A,\*</sup>, Nitish Biswas<sup>A</sup>, Rafid Mostafiz<sup>B</sup>, Fateha Sharmin<sup>C</sup>, Samrat Kumar Dey<sup>D</sup>

<sup>A</sup>Departemen Ilmu dan Teknik Komputer, Universitas Internasional Dhaka, Dhaka 1205, Bangladesh

<sup>B</sup>Institut Teknologi Informasi, Universitas Sains dan Teknologi Noakhali, Noakhali, Bangladesh

<sup>C</sup>Departemen Kimia, Universitas Chittagong, Chittagong, Bangladesh

<sup>D</sup>Sekolah Sains dan Teknologi, Universitas Terbuka Bangladesh, Gazipur 1705, Bangladesh

### ARTIKEL

### INFORMASI

### ABSTRAK

#### Kata kunci:

Kanker paru-paru

ROS pembelajaran mesin yang dapat

dijelaskan

BENTUK

GBM

Aplikasi seluler

Kanker paru-paru telah menjadi penyebab utama kematian terkait kanker di seluruh dunia. Deteksi dini dan diagnosis kanker paru-paru dapat sangat meningkatkan peluang kelangsungan hidup pasien. Pembelajaran mesin semakin banyak digunakan di sektor medis untuk mendeteksi kanker paru-paru, namun kurangnya interpretasi model ini masih menjadi tantangan yang signifikan. Pembelajaran mesin yang dapat dijelaskan (XML) adalah pendekatan baru yang bertujuan untuk memberikan transparansi dan interpretasi untuk model pembelajaran mesin. Seluruh percobaan telah dilakukan pada kumpulan data kanker paru-paru yang diperoleh dari Kaggle. Hasil model prediktif dengan teknik penyeimbangan kelas ROS (Random Oversampling) digunakan untuk memahami gambaran klinis paling relevan yang berkontribusi terhadap prediksi kanker paru-paru menggunakan teknik penjelasan pembelajaran mesin yang disebut SHAP (SHapley Additive exPlanation). Hasilnya menunjukkan kekokohan kapasitas GBM dalam mendeteksi kanker paru-paru, dengan akurasi 98,76%, presisi 98,79%, recall 98,76%, F-Measure 98,76%, dan tingkat error 0,16%. Terakhir, aplikasi seluler dikembangkan dengan menggabungkan model terbaik untuk menunjukkan kemanjuran pendekatan kami.

### Perkenalan

Jenis kanker yang paling mematikan secara global adalah kanker paru-paru. Ini adalah salah satu penyebab utama kematian akibat kanker pada wanita dan pria.<sup>1,2</sup>Salah satu jenis kanker yang bermula di paru-paru adalah kanker paru-paru. Ketika sel-sel tubuh mulai berkembang biak di luar kendali, kanker pun berkembang. Kanker paru-paru sering kali berkembang dalam jangka waktu yang lama dan terutama menyerang orang berusia antara 55 dan 65 tahun.<sup>3</sup>Kanker paru-paru sel kecil (SCLC) dan kanker paru-paru non-sel kecil (NSCLC) adalah dua jenis utama kanker paru-paru. NSCLC menyumbang sekitar 80% –85% kasus kanker paru-paru. Paling sering, perokok atau mantan perokok menderita kanker paru-paru jenis ini. Lebih dari 85% kasus kanker paru-paru disebabkan oleh perokok aktif atau sebelumnya.<sup>4</sup>Dibandingkan dengan bentuk kanker paru-paru lainnya, penyakit ini lebih umum terjadi pada wanita dibandingkan pria dan lebih cenderung menyerang orang yang lebih muda. Sebaliknya, SCLC, juga dikenal sebagai karsinoma sel oat, menyumbang 10% –15% dari seluruh kasus kanker paru-paru. Laju pertumbuhan SCLC dan pembentukan tumor besar yang berpotensi menyebar jauh ke seluruh tubuh sebenarnya berkorelasi langsung dengan kebiasaan merokok. Seringkali bermula di bronkus di tengah dada. Jumlah keseluruhan rokok yang dihisap berdampak pada angka kematian akibat kanker paru-paru.<sup>5</sup>Menurut Organisasi Kesehatan Dunia

(WHO), kanker paru-paru merupakan penyebab utama kematian terkait kanker pada tahun 2020, yang merenggut 1,80 juta nyawa.

Baik angka kematian maupun jumlah orang yang terkena penyakit ini diperkirakan akan terus meningkat seiring dengan bertambahnya jumlah penduduk dunia. Tingkat kelangsungan hidup 5 tahun untuk kanker paru-paru hanya 18%, hal ini menunjukkan pentingnya deteksi dini dan diagnosis. Pencitraan medis, seperti pemindaiannya tomografi komputer (CT), biasanya digunakan dalam diagnosis kanker paru-paru. Namun, interpretasi gambar medis dapat menjadi tantangan dan memakan waktu bagi ahli radiologi. Angka kematian ini dapat diturunkan dengan deteksi dini dan pengobatan. Algoritme pembelajaran mesin bisa sangat bermanfaat dalam situasi tersebut untuk memperkirakan keganasan dengan tepat.<sup>6–8</sup>Namun, karena kompleksitas dan sifat black-box dari banyak algoritma pembelajaran mesin, sulit untuk memahami bagaimana model membuat prediksi atau keputusan dan untuk mengidentifikasi potensi kesalahan atau bias. Hal ini dapat menjadi perhatian utama dalam bidang kesehatan, keuangan, dan peradilan pidana, dimana konsekuensi dari kesalahan model bisa sangat parah. Keterbatasan lain dari model black-box adalah bahwa model tersebut sering kali sensitif terhadap pilihan hyperparameter, yang dapat membuatnya sulit untuk dioptimalkan dan digeneralisasikan ke data baru. Selain itu, model kotak hitam rentan terhadap overfitting, yang dapat menyebabkan performa buruk pada data yang tidak terlihat. Keseluruhan,

\* Penulis koresponden di: Departemen Ilmu dan Teknik Komputer, Universitas Internasional Dhaka, Dhaka 1205, Bangladesh.  
Alamat email:jilanicsejnu@gmail.com (KMM Uddin),samrat.sst@bou.ac.bd (SK Dey).

model kotak hitam telah sangat sukses dalam berbagai aplikasi, namun kurangnya interpretasi dan transparansi merupakan masalah utama. Pembelajaran mesin yang dapat dijelaskan (XML) adalah pendekatan baru yang bertujuan untuk memberikan transparansi dan interpretasi model kotak hitam, yang dapat membantu mengatasi keterbatasan ini.

Oleh karena itu, teknik post hoc akhir-akhir ini menjadi semakin populer sebagai solusi permasalahan penyajian model kotak hitam dengan cara yang dapat dimengerti oleh manusia. Penjelasan seperti ini sering digunakan untuk membantu pakar domain dalam menemukan bias diskriminatif dalam model blackbox.<sup>9,10</sup>

Metode lokal model-agnostik yang berkonsentrasi pada penjelasan prediksi spesifik dari pengklasifikasi kotak hitam tertentu, seperti LIME<sup>11</sup> dan BENTUK,<sup>12</sup> adalah salah satu teknik yang paling terkenal. Teknik-teknik ini menghasilkan gangguan pada kejadian tertentu dalam data dan melacak dampak gangguan ini pada keluaran pengklasifikasi kotak hitam untuk mengukur kontribusi fitur individual terhadap prediksi tertentu. Pendekatan ini telah diterapkan di berbagai bidang, termasuk hukum, kedokteran, keuangan, dan sains<sup>13-15</sup> karena sifatnya yang umum, untuk menjelaskan berbagai pengklasifikasi, termasuk jaringan saraf dan model ansambel yang canggih.

4 prinsip yang dapat dijelaskan,<sup>16</sup> Namun, digunakan dalam percobaan ini untuk memperkirakan kanker paru-paru. Transparansi adalah prinsip utama pertama yang menjelaskan model dengan cara yang jelas dan mudah dipahami, misalnya dengan menyoroti fitur terpenting yang mengarah pada diagnosis tertentu. Hal ini dapat dicapai dengan menggunakan teknik seperti pentingnya fitur, pemilihan fitur, dan metode interpretasi model seperti LIME, SHAP, dan lain-lain. Menurut prinsip kedua bernama Fairness, model tidak boleh mendiskriminasi kelompok pasien tertentu berdasarkan faktor seperti usia, ras, atau jenis kelamin. Hal ini dapat dicapai dengan menggunakan algoritma yang sadar akan keadilan, seperti algoritma yang secara eksplisit mengoptimalkan metrik keadilan kelompok, atau dengan menggunakan metode koreksi bias, seperti pengambilan sampel ulang atau pelatihan permusuhan. Dan prinsip ketiga adalah Robustness. Menurut prinsip ini, model harus kuat terhadap perubahan kecil pada data masukan dan menghasilkan prediksi yang konsisten bahkan ketika disajikan dengan data baru atau data yang belum terlihat. Hal ini dapat dicapai dengan menggunakan teknik seperti validasi silang, regularisasi, dan ensembling untuk meningkatkan generalisasi model. Sesuai dengan prinsip terakhir, akuntabilitas, model harus mampu memberikan penjelasan atas prediksinya jika terjadi kesalahan atau kesalahan, dan harus dapat memahami penyebab kesalahan tersebut dan mengambil tindakan perbaikan. Hal ini dapat dicapai dengan menggunakan teknik seperti pemantauan model, audit model, dan tata kelola model untuk memastikan bahwa model berperilaku sesuai harapan dan setiap masalah atau bias diidentifikasi dan ditangani secara tepat waktu. Prinsip-prinsip pembelajaran mesin yang dapat dijelaskan di atas akan membantu membangun model yang dapat dipercaya dan dipahami sehingga dapat membantu meningkatkan akurasi prediksi dan memastikan bahwa model tersebut adil, kuat, dan akuntabel.

Namun, kami mengusulkan nilai SHAP (SHapley Additive exPlanation) dalam eksperimen ini untuk meningkatkan kepercayaan, tanggung jawab, proses debug, dan banyak tugas lainnya. Konsep teori permainan<sup>17</sup> dan penjelasan lokal digunakan untuk membentuk landasan SHAP. Strategi yang dapat dijelaskan telah disorot dalam jurnal terkenal dan juga mendapatkan popularitas dalam aplikasi medis. Cosgriff dan Celi<sup>18</sup> menunjukkan cara menganalisis model jaringan saraf dalam menggunakan metodologi penjelasan menggunakan catatan pasien elektronik frekuensi tinggi. Model penjelasan dijelaskan dalam Lundberg dkk.,<sup>19</sup> sebagai cara untuk melengkapi model ML dalam memperkirakan angka kematian pasien gagal ginjal. Saat ini, analisis data gambar, sinar-X, CT scan, ultrasound, dan teknik pencitraan lainnya menggunakan model yang dapat dijelaskan. Lundberg dkk.<sup>20</sup> jelaskan bagaimana model meramalkan hipoksemia selama operasi berlangsung. Daftar teknik yang digunakan dalam pengobatan dibahas lebih mendalam di Singh et al.<sup>21</sup> Selain itu, penelitian terhadap kondisi paru-paru sedang berlangsung. Banyak artikel telah ditulis tentang penggunaan kecerdasan buatan (AI) untuk menangani kondisi paru-paru. Xi dkk.<sup>22</sup> menggunakan aerosol yang dihembuskan untuk mendeteksi penyakit struktural paru-paru menggunakan algoritma pembelajaran mesin seperti Random Forests (RF) dan Support Vector Machines. Dalam penyelidikan ekstensif, model RF juga digunakan.<sup>23</sup> Untuk menemukan kanker paru-paru, para ilmuwan menyarankan perangkat e-nose buatan tangan. Meski sudah banyak penelitian mengenai hal ini, kami tetap terdorong

untuk bekerja dengan kanker paru-paru. Artikel ini berfokus pada penjelasan model yang digunakan pada kanker paru-paru sehingga orang dapat memahami cara kerja model secara internal dan yakin dengan prediksi eksperimen tersebut. Ringkasan proposal ditunjukkan pada Gambar 1.

Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost), dan Light Gradient Boosting (LightGBM) merupakan 3 pengklasifikasi berbasis ansambel yang digunakan dalam penelitian ini. Di antara semua pengklasifikasi, GBM memiliki akurasi terbaik, yaitu 98,76%. Terakhir, aplikasi smartphone dikembangkan untuk mengintegrasikan model terbaik. Cara penerapan proposal ditunjukkan pada Gambar 2.

Berikut ini adalah kontribusi terhadap usulan upaya penelitian kanker paru-paru:

1. GBM mencapai akurasi yang lebih tinggi yaitu 98,76% dalam penyelidikan ini.
2. Untuk hasil terbaik, 3 pengklasifikasi—XGBoost, LightGBM, dan GBM—are digunakan dalam skenario ini.
3. Teknik keseimbangan data, penskalaan fitur, PCA (Analisis Komponen Utama), dan penyetelan hyperparameter telah digunakan untuk mencapai tingkat akurasi tertinggi.
4. SHAP digunakan untuk membuat keluaran peningkatan gradien dapat dimengerti, bermakna, dan dapat dipercaya oleh manusia.
5. Terakhir, aplikasi ponsel pintar yang mudah digunakan yang dapat menghitung hasil berdasarkan masukan waktu nyata telah dikembangkan.

Ada 4 bagian untuk bisa proyek penelitian ini. Bagian 2 menunjukkan pekerjaan terkait dan Bagian 3 menyediakan bahan dan metode. 4 subbagian di Bagian 3 adalah deskripsi kumpulan data dan pra-pemrosesan data, penyetelan PCA dan Hyperparameter, model pembelajaran mesin, dan SHAP. Analisis dan pembahasan hasil dibahas pada Bagian 4 yang berisi 5 subbagian. Subbagian ini adalah pengaturan lingkungan, akurasi klasifikasi, evaluasi model, analisis hasil SHAP, dan pembuatan aplikasi seluler. Tugas akhirnya selesai di Bagian 5.

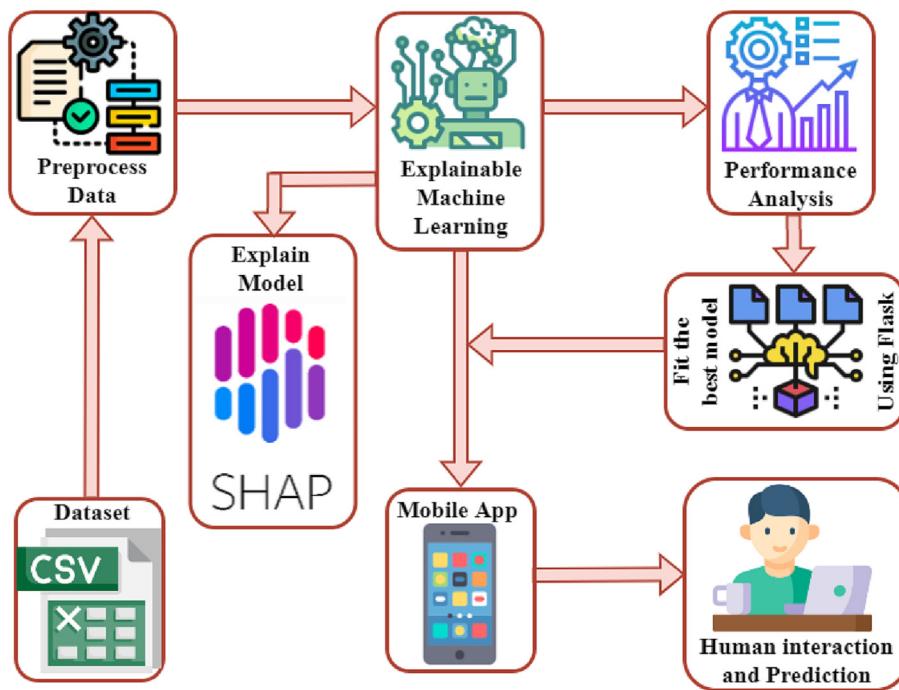
#### Pekerjaan yang berhubungan

Ada peningkatan minat terhadap penggunaan teknik eXplainable Machine Learning (XML) untuk prediksi kanker paru-paru dalam beberapa tahun terakhir. Beberapa penelitian penting di bidang ini meliputi:

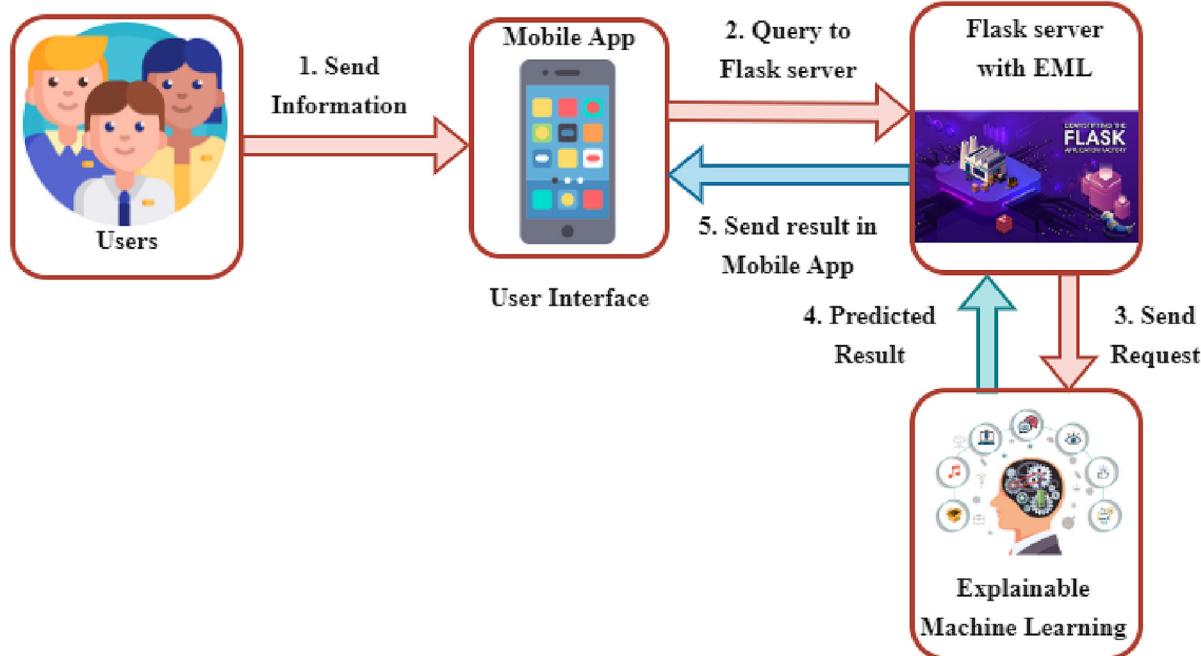
Dalam sebuah penelitian, Masrur Sobhan dan Ananda Mohan Mondal<sup>24</sup> mengusulkan jalur untuk mengidentifikasi kelas kanker paru-paru yang signifikan dan gen spesifik pasien yang dapat mendukung pengembangan obat-obatan yang efektif untuk pasien kanker paru-paru. Mereka menggunakan 2 varian SHAP yang dikenal sebagai "penjelasan pohon" dan "penjelasan gradien", yang masing-masing menggunakan algoritme klasifikasi "pengklasifikasi berbasis pohon", XGBoost, dan "pengklasifikasi berbasis pembelajaran mendalam", jaringan saraf konvolisional. 100 gen teratas spesifik kelas dan gen yang diekspresikan secara berbeda, keduanya merupakan biomarker berbasis populasi. Hanya sedikit gen yang ditemukan dimiliki oleh pasien, menunjukkan bahwa setiap individu dengan kanker paru-paru diwakili oleh serangkaian gen spesifik pasien berbeda yang ditemukan. Pengujian ini menunjukkan bahwa XGBoost mencapai akurasi 96,3%.

Dalam penelitian ini,<sup>25</sup> Model pembelajaran mesin (ML) digunakan untuk memprediksi lama rawat inap (LOS) pasien kanker paru-paru. Metodologi ini bertujuan untuk mengatasi ketidakseimbangan kumpulan data untuk metode berbasis klasifikasi yang menggunakan rekam medis elektronik (EHR). Mereka memperkirakan rata-rata lama rawat inap (LOS) untuk pasien ICU dengan kanker paru-paru menggunakan kumpulan data MIMIC-III dan algoritma ML yang diawasi. Model Random Forest (RF) berkinerja lebih baik dibandingkan model lainnya selama 3 tahap kerangka kerja dan memberikan hasil yang diharapkan. Mereka mendeskripsikan hasil model prediktif (RF) menggunakan teknik keseimbangan kelas SMOTE untuk memahami faktor klinis paling signifikan yang berkontribusi dalam memprediksi LOS kanker paru-paru menggunakan model RF yang memanfaatkan SHAP.

Studi lain oleh Jamie et al.<sup>26</sup> menggunakan 3 teknik XAI—SHAP, LIME, dan Scoped Rules—untuk menunjukkan betapa bermanfaatnya menambahkan lampiran tersier ke model ML yang dapat dijelaskan dan memberikan interpretasi data untuk kumpulan data EHR skala besar. Simulacrum, kumpulan data sintetis yang dihasilkan oleh Data Kesehatan



Gambar 1.Ikhtisar pekerjaan yang diusulkan.



Gambar 2.Alur kerja aplikasi seluler.

Insight CiC menggunakan data kanker anonim yang disediakan oleh National Cancer Registration and Analysis Service (NCRAS) Kesehatan Masyarakat Inggris, berfungsi sebagai sumber data untuk eksperimen ini. Mereka membandingkan fitur EHR berdasarkan relevansi prediksi tertimbang yang dihitung oleh model XAI. Namun, dalam penelitian ini, 3 pengklasifikasi—Regresi Logistik, XGBoost, dan EBM—digunakan, dengan XGBoost menampilkan performa terbaik dalam hal akurasi klasifikasi.

Dengan menggunakan contoh model yang digunakan untuk menilai risiko kanker paru-paru dalam skrining kanker paru-paru dengan tomografi komputer dosis rendah, Katarzyna dkk.<sup>27</sup> mengusulkan pendekatan terpilih dari bidang XAI dalam pekerjaan lain. Ini

metode ini membantu pemahaman yang lebih baik mengenai perbandingan 3 model prediksi risiko kanker paru yang digunakan dalam skrining kanker paru, yaitu model BACH, model PLCOm2012, dan model LCRAT. Kinerja dan akurasi model penelitian tidak dibahas oleh penulis. Mereka hanya berkonsentrasi pada pemahaman bagaimana model bertindak untuk berbagai pasien. Untuk penyelidikan ini, mereka menggunakan database kanker paru-paru dalam negeri.

Elias Dritsas dan Maria Trigka<sup>28</sup> menggunakan berbagai pengklasifikasi pembelajaran mesin, termasuk NB, BayesNet, SGD, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF, dan AdaBoostM1, untuk mengidentifikasi orang-orang yang berisiko tinggi terkena penyakit paru-paru. kanker. Untuk mengidentifikasi model dengan

kinerja prediktif tertinggi, pengklasifikasi ini dinilai dalam hal akurasi, presisi, perolehan, F-Measure, dan AUC. Sumber kumpulan datanya diperoleh dari Kaggle. Model RotF dari eksperimen ini memiliki performa pada level tertinggi.

Penelitian lain mengenai penelitian kanker paru dilakukan oleh Muntasir dkk. di mana mereka memeriksa sejumlah penelitian sebelumnya yang berkaitan dengan model prediksi kanker paru-paru dan membandingkan hasilnya dengan model mereka. Mereka menciptakan pendekatan pembelajaran XGBoost, LightGBM, AdaBoost, dan bagging ensemble, antara lain, untuk memprediksi kanker paru-paru. Pendekatan validasi model dilakukan dengan menggunakan K-fold 10 crossvalidation. Akurasi terbaik pada percobaan ini adalah 94,42% yang dicapai dengan XGBoost.

Patra<sup>30</sup> memeriksa berbagai pengklasifikasi pembelajaran mesin, termasuk Radial Basis Function Network (RBF), K-Nearest Neighbors (KNN), J48, Support Vector Machine (SVM), Logistic Regression, Artificial Neural Network (ANN), Nave Bayes, dan Random Forest, untuk memprediksi kanker paru-paru. Kumpulan data, yang mencakup 32 kejadian dan 57 atribut, dikumpulkan dari "repositori UCI". RBF mencapai akurasi sebesar 81,25%, yang menurut penulis lebih tinggi dibandingkan semua algoritma lainnya.

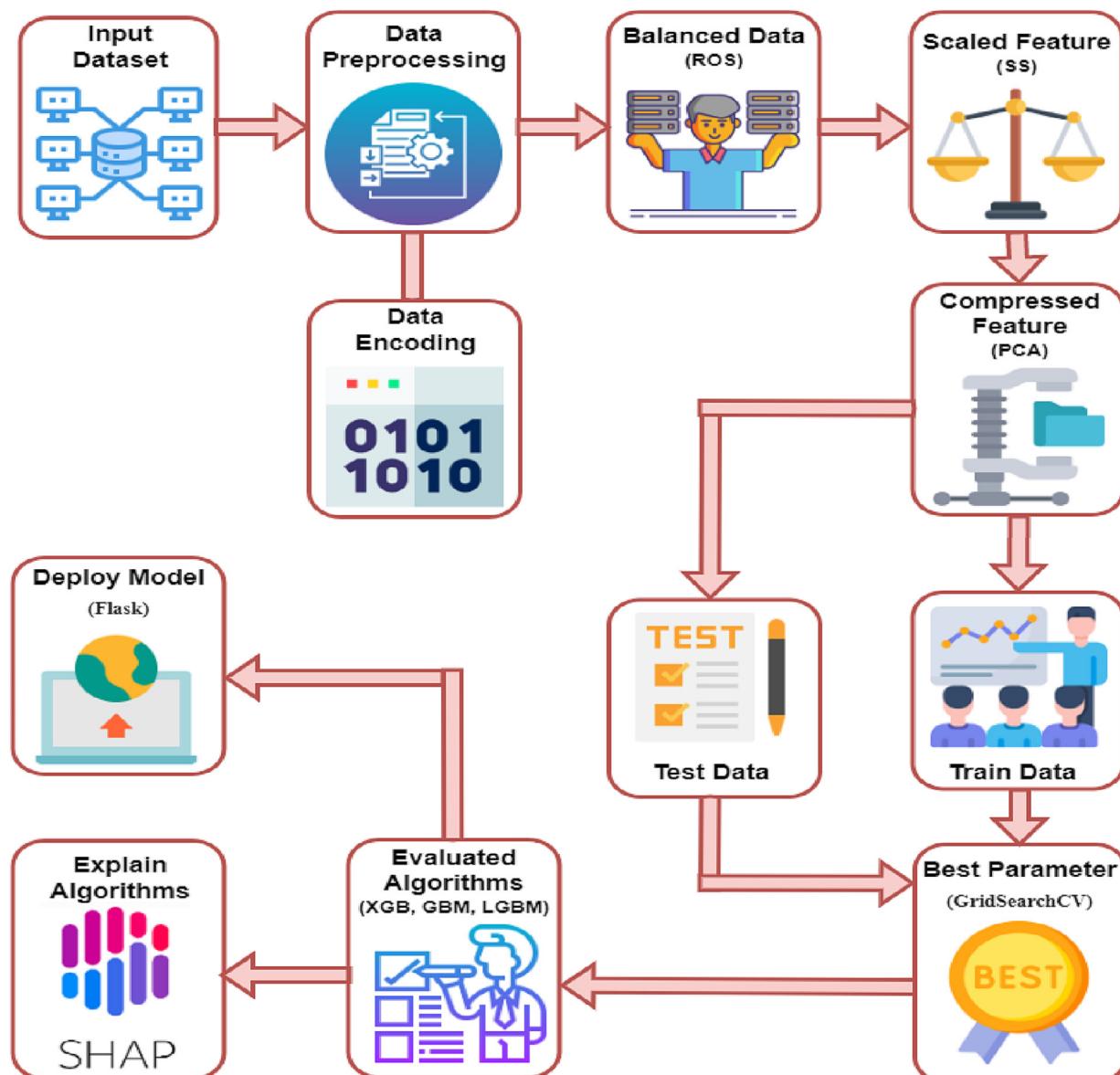
Studi lain oleh Sim et al.<sup>31</sup> menyarankan studi tentang kualitas hidup terkait kesehatan (HRQOL) dalam prediksi kelangsungan hidup kanker paru-paru 5 tahun menggunakan berbagai

model pembelajaran mesin, termasuk Decision Tree, Logistic Regression, Bagging, Random Forest, dan AdaBoost. Untuk menilai kinerja model, 2 set fitur berbeda digunakan dengan validasi silang K-fold 5. Data dari 809 penyintas operasi kanker paru-paru yang menjalani operasi dibandingkan dengan penampilan model. Temuan percobaan ini menunjukkan bahwa AdaBoost memiliki akurasi tertinggi yaitu 94,8%.

Secara keseluruhan, studi-studi ini terutama berfokus pada menunjukkan potensi penggunaan pembelajaran mesin yang dapat dijelaskan untuk penelitian kanker paru-paru, karena dapat meningkatkan kinerja dan kemampuan interpretasi model pembelajaran mesin. Mereka memberikan wawasan yang dapat ditafsirkan mengenai mekanisme yang mendasari penyakit ini dan faktor-faktor yang berkontribusi terhadap perkembangannya.

## Bahan dan metode

Bagian ini menjelaskan pendekatan yang kami rekomendasikan dan teknik yang digunakan untuk mendeteksi kanker paru-paru. **Gambar 3** menggambarkan pendekatan yang diusulkan untuk memprediksi kanker paru-paru. Pendekatan yang disarankan adalah dengan melihat pengkodean label, metode pemilihan fitur dengan skalar standar, di mana setiap nilai fitur dalam data memiliki variansi nol dan satuan, analisis Komponen Utama (PCA), yang mengompresi data, penyetelan hyperparameter dengan pencarian grid, yang digunakan untuk menentukan model dan kinerja terbaik



Gambar 3. Prosedur kerja metodologi yang diusulkan.

matriks, yang digunakan untuk meningkatkan efektivitas model klasifikasi. Hasil eksperimen ini menunjukkan bahwa penggunaan semua pengklasifikasi pembelajaran mesin untuk memprediksi kanker paru-paru adalah arah yang menjanjikan. Di bawah ini adalah gambaran algoritma yang digunakan dalam penelitian ini untuk mempermudah pemahaman.

**Algoritma:** Prosedur Kerja Prediksi Kanker Paru XML Memasukkan:

Kumpulan Data Kanker Paru Kaggle

**Keluaran:** Prediksi Nilai XML Kanker Paru (Ya atau Tidak)

### 1. Mulai

```

2. data←memuat kumpulan data;
3. bentuk←bentuk beban;
4. Prosedur DO_EXPLAINABLE(model, x )
5. penjelasan←shap.Penjelasan( model, x );
6. nilai_bentuk←penjelasan( x );
7.jikaplotSAMABATANG
8. shap.bar( shap_values );
9.lain jikaplotSAMAHANGAT
10. shap.beeswarm( shap_values );
11.kalau tidak
12. shap.waterfall( shap_values );
13.prosedur akhir
14.pra-pemrosesan:
15.jikadata.dtype adalah objek atau string yang sama
16. pengkodean data;
17.x←data.drop[paru-paru];
18.tahun←data.paru-paru;
19.penyimbangan_data:
20. x_os, y_os = RandomOverSampling(x,y);
21.penskalaan_fitur:
22. berskala_x←penskalaan_fitur_(x_os);
23.fitur_optimasi:
24. pca_X = PCA(n_components = 9).fit(scaled_X);
25.x1, x2, y1, y2←split_data dari pca_X dan y_os;
26.untuk saya inrentang(len(model)):
27. memeriksa HTP;
28. model←train_model menggunakan x1 dan y1;
29. meramalkan←pengujian_model menggunakan x2 dan y2;
30. menghitung metrik evaluasi kinerja;
31.LAKUKAN_MENJELASKAN(model, x);
32.Akhir

```

Deskripsi kumpulan data dan pra-pemrosesan data

Kumpulan data yang disebut "Kanker Paru-Paru" diperoleh dari Kaggle<sup>32</sup> dan berisi 309 kejadian dan 16 atribut, dimana 15 atribut bersifat prediktif dan 1 atribut kelas. Kanker paru-paru adalah atribut kelas, dan atribut prediktifnya adalah, secara berurutan, jenis kelamin, usia, merokok, jari kuning, kecemasan, tekanan teman sebaya, penyakit kronis, kelelahan, alergi, mengi, alkohol, batuk, sesak napas, kesulitan menelan, dan nyeri dada. **Tabel 1** menjelaskan setiap fitur kumpulan data.

Salah satu karakteristiknya adalah Gender dan Kanker Paru-paru sama-sama mengandung nilai kategorikal yang telah diubah menjadi nilai numerik (0,1) pada tahap pra-pemrosesan data melalui pengkodean label. Kebisingan kumpulan data, nilai atau informasi yang hilang, dan data yang tidak seimbang<sup>33</sup> dapat mengurangi keakuratan hasil. Itulah sebabnya item yang tidak diinginkan dari kumpulan data ini dihilangkan sebelum model pembelajaran mesin dijalankan. Output terbaik untuk kumpulan data dicapai dengan pra-pemrosesan data. Namun, kumpulan data ini tidak mengandung nilai apa pun yang hilang, namun kumpulan data ini sepenuhnya tidak seimbang. Pengambilan sampel berlebihan secara acak (ROS)<sup>34</sup> digunakan di sini untuk mengatasi masalah kumpulan data yang tidak seimbang. Ini melibatkan duplikasi contoh dari kelas minoritas untuk menyeimbangkan distribusi kelas. Hal ini dilakukan dengan memilih contoh secara acak dari kelas minoritas dan menambahkannya ke dataset hingga distribusi kelas seimbang. Untuk melakukan eksperimen ini, kelas minoritas telah ditingkatkan sebesar 70%. Setelah oversampling, total 216 baris dalam dataset memiliki nilai 1 yang menunjukkan bahwa ditemukan keganasan, sedangkan 270 baris

Tabel 1

Deskripsi setiap karakteristik dalam dataset.

Nama atribut	Keterangan
Jenis kelamin	Ciri ini menunjukkan apakah seseorang berjenis kelamin laki-laki atau perempuan.
Usia	Usia seseorang dicatat menggunakan fitur ini.
Merokok	Ciri-ciri ini menunjukkan apakah peserta tersebut merokok atau tidak.
Jari kuning	Ciri ini menunjukkan apakah jari penggunanya berwarna kuning atau tidak.
Kecemasan	Ada tidaknya rasa cemas ditunjukkan oleh ciri ini.
Tekanan teman sebaya	Karakteristik ini membuat pengguna mengetahui apakah mereka rentan terhadap tekanan sebaya atau tidak.
Penyakit kronis	Fitur ini menunjukkan apakah peserta mengidap penyakit kronis atau tidak.
Kelelahan	Tingkat kelelahan peserta ditunjukkan dengan atribut ini.
Alergi	Ciri-ciri ini menunjukkan apakah seseorang mempunyai alergi atau tidak.
Mengi	Fitur ini menunjukkan apakah seseorang menderita mengi atau tidak. Fitur ini menunjukkan apakah orang tersebut minum atau tidak.
Alkohol	Fitur ini menunjukkan apakah peserta sedang batuk atau tidak.
Batuk	Fitur ini menunjukkan apakah peserta menderita sesak napas atau tidak.
Sesak napas	Fitur ini menunjukkan apakah peserta mengalami masalah menelan atau tidak.
Kesulitan menelan	Fitur ini menunjukkan apakah pengguna mengalami masalah menelan atau tidak.
Nyeri dada	Fitur ini menunjukkan apakah seseorang mengalami nyeri dada atau tidak.
Kanker paru-paru	Fitur ini menentukan apakah seseorang telah didiagnosis menderita kanker paru-paru atau belum.

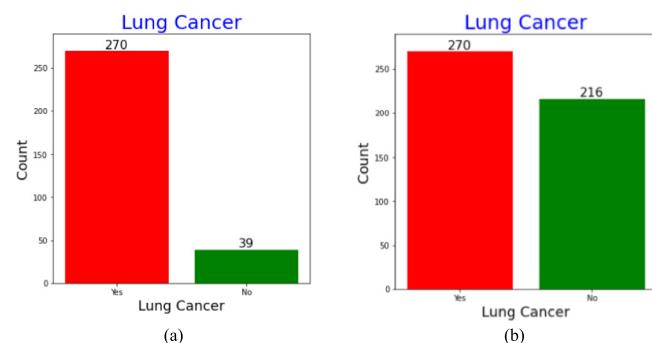
dalam kumpulan data memiliki nilai 0, yang menunjukkan bahwa tidak ada kanker yang ditemukan.

Gambar 4 menunjukkan data tidak seimbang sebelum tidak seimbang dan setelah tidak seimbang.

### Penyetelan PCA dan hyperparameter

Dataset percobaan ini terdiri dari 16 fitur yang berdimensi tinggi. Karena overfitting, berbagai fitur ini mempersulit pencapaian hasil yang optimal. Untuk meningkatkan kinerja hasil, Analisis Komponen Utama (PCA) diterapkan pada kumpulan data, yang mengurangi 16 karakteristik menjadi 9. PCA adalah teknik pengurangan dimensi yang sering digunakan untuk mengurangi dimensi kumpulan data besar.<sup>35</sup> Hal ini dilakukan dengan memadatkan kumpulan variabel yang besar menjadi kumpulan variabel yang lebih kecil sambil mempertahankan sebagian besar data dalam kumpulan variabel yang lebih besar.<sup>36</sup>

Namun, pembelajaran mesin menggunakan penyetelan hyperparameter,<sup>37</sup> dimana nilai parameter dipilih sebelum algoritma diajarkan. Kumpulan hyperparameter khusus tersebut memaksimalkan performa model dan memberikan hasil yang lebih baik dengan lebih sedikit kesalahan dengan meminimalkan fungsi kerugian yang telah ditetapkan sebelumnya. GridSearchCV dan penyetelan hyperparameter digabungkan dalam penelitian ini. GridSearchCV adalah metode di perpustakaan scikit-learn untuk Python yang digunakan untuk melakukan pencarian mendalam pada area tertentu.



Gambar 4. Data seimbang: (a) Sebelum penyeimbangan data dan (b) setelah penyeimbangan data.

ruang parameter untuk estimator. Teknik penyetelan hyperparameter ini melakukan pencarian validasi silang yang komprehensif untuk menemukan nilai ideal untuk hyperparameter yang diinginkan.<sup>38</sup> Model mengevaluasi dan memverifikasi setiap kumpulan nilai kamus yang unik.<sup>39</sup> Namun untuk mendapatkan akurasi yang menjanjikan dari GridSearchCV, kami menentukan parameter untuk GBM, XGBoost, dan LightGBM. Parameter GBM terbaik n\_estimators=100, learning\_rate=1, dan max\_depth=1 dipilih untuk menghasilkan akurasi 98,76%.

GridsearchCV kemudian digunakan untuk membangun XGBoost, dengan parameter terbaik adalah objektif="biner: logistik", random\_state=45, eval\_metric="auc," dan n\_estimators=100, menghasilkan akurasi 98,27%. Mengikuti pengembangan model LightGBM dengan parameter default, prediksi dibuat pada set pengujian yang tidak terlihat. Namun akurasinya buruk. Akibatnya, model dibangun menggunakan GridsearchCV, yang mencapai akurasi 98,89% ketika nilai ideal untuk num\_leaves=31, learning\_rate=1, dan n\_estimators=100 digunakan. Validasi silang 10 kali lipat digunakan untuk melakukan eksperimen ini. Oleh karena itu, model terbaik dengan akurasi tertinggi dipilih untuk setiap kumpulan hyperparameter.

#### Pengklasifikasi pembelajaran mesin

Pemodelan data dilakukan di sini menggunakan 3 algoritme pembelajaran mesin berbasis ansambel: Gradient GBM, LightGBM, dan XGBoost. GBM adalah proses pembelajaran yang secara bertahap menyesuaikan model baru untuk memberikan perkiraan variabel respons yang lebih tepat. Untuk membuat perkiraan akhir, ini menggabungkan prediksi dari banyak pohon keputusan. Setiap node pohon keputusan menggunakan subset informasi yang berbeda untuk memutuskan pemisahan mana yang terbaik. Hal ini menunjukkan bahwa tidak ada dua pohon yang persis sama, dan sebagai hasilnya, berbagai sinyal dapat diekstraksi dari data oleh setiap pohon. Setiap pohon berikutnya juga memperhitungkan kesalahan atau kesalahan apa pun yang dihasilkan oleh pohon sebelumnya. Jadi, setiap pohon keputusan yang muncul setelahnya dibangun menggunakan kelemahan dari pohon sebelumnya. Algoritme mesin peningkat gradien membangun pohon secara berurutan dengan cara ini.

Metode peningkatan terkenal lainnya adalah XGBoost. XGBoost sebenarnya hanyalah versi modifikasi dari algoritma GBM. Tujuan dari pengklasifikasi ini adalah untuk mengklasifikasikan data secara akurat dengan menghitung pengklasifikasi lemah secara berulang.<sup>40</sup> Terapkan kriteria akurasi dan kerugian logistik untuk memilih model terbaik dalam ruang hipotesis dan membuat prediksi terbaik dari data pengujian berdasarkan kriteria evaluasi menggunakan data sampel yang disediakan dan tidak bergantung satu sama lain. Faktanya, ini terdiri dari sejumlah metode regularisasi yang mengurangi overfitting dan meningkatkan kinerja secara umum.

Data dalam jumlah besar dapat ditangani dengan mudah dengan LightGBM. Pemisahan optimal dipilih oleh LightGBM menggunakan strategi berbasis histogram untuk mempercepat proses pelatihan. Variabel kontinu apa pun dipisahkan ke dalam wadah atau keranjang, bukan menggunakan nilai individual. Ini mempersingkat periode pelatihan dan menggunakan lebih sedikit memori. Namun, 3 algoritma pembelajaran mesin diperiksa dalam penelitian ini untuk mengetahui kapasitasnya dalam memprediksi munculnya kanker paru-paru, dengan GBM menunjukkan tingkat akurasi tertinggi.

#### SHAP (Penjelasan Aditif SHapley)

SHAP adalah metode komprehensif untuk menganalisis hasil model pembelajaran mesin apa pun yang dikembangkan oleh Lundberg et al.<sup>41</sup> SHAP menyediakan cara untuk menghitung kontribusi setiap fitur dan didasarkan pada teori permainan dan penjelasan lokal. Model tersebut menghasilkan nilai prediksi untuk setiap sampel prediksi, dan nilai SHAP adalah skor yang diberikan untuk setiap fitur dalam kumpulan data.<sup>42</sup> Untuk mendukung iML, SHAP dibuat dan tersedia sebagai seperangkat alat python. Untuk setiap fitur, SHAP menyediakan daftar nilai Shapley untuk datum tertentu. Hal ini didasarkan pada anggapan bahwa prediksi dapat digambarkan dengan anggapan bahwa setiap fitur adalah "pemain" dalam permainan dimana prediksi adalah pembayarannya.<sup>43</sup> Nilai Shapley, sebuah strategi dari teori permainan koalisi, menjelaskan bagaimana mendistribusikan "pembayaran" secara merata ke seluruh karakteristik. Banyak faktor berbeda yang akan muncul dalam kumpulan data kami yang ada. Setiap variabel yang berbeda dapat dipandang sebagai pemain dalam pengertian teori permainan. Manfaat dari beberapa peserta yang bekerja sama untuk menyelesaikan suatu proyek dapat dilihat dari hasil prediksinya

#### Meja 2

Pengaturan lingkungan dari sistem yang diusulkan.

Sumber	Detail
CPU	Intel® Inti™CPU i3-1005G1 @ 1,20GHz
RAM	12GB
GPU	Grafik Intel® UHD
Perangkat lunak	Anakonda
Bahasa	ular piton

dihadirkan dengan memanfaatkan kumpulan data ini untuk melatih model. Nilai Shapley mendistribusikan keuntungan kerjasama secara merata dengan memperhatikan kontribusi masing-masing pemain. Ukuran standar relevansi fitur hanya menunjukkan fitur mana yang signifikan, dan pengaruhnya terhadap hasil prediksi tidak diketahui. Manfaat utama dari nilai SHAP adalah bahwa nilai tersebut dapat menunjukkan dampak positif dan negatif dari dampak atribut pada setiap sampel.

#### Hasil dan Diskusi

Pertama, data dibagi menjadi pelatihan (65%) dan pengujian (35%). Model optimal dengan akurasi tertinggi diperiksa menggunakan berbagai metode pembelajaran mesin, termasuk penskalaan fitur, PCA, ROS, dan penyetelan hyperparameter. Model terbaik dipilih menggunakan semua teknik pembelajaran mesin ini.

#### Pengaturan lingkungan

Eksperimen ini melibatkan beberapa sumber daya. Meja 2 menyajikan bahan-bahan yang digunakan untuk pengembangan model penelitian ini.

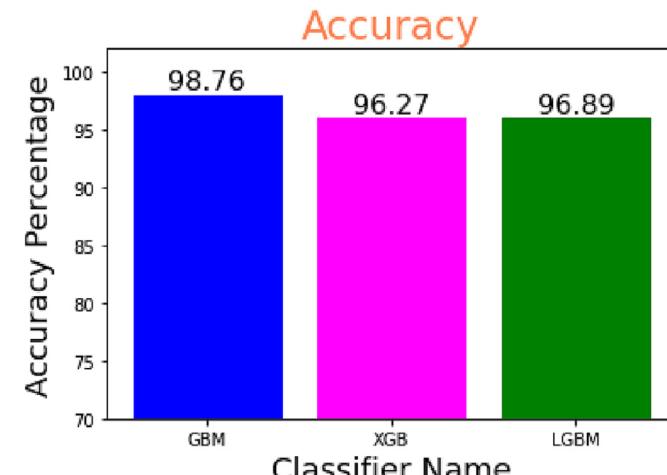
#### Akurasi klasifikasi

Efisiensi sistem klasifikasi dinilai menggunakan sejumlah matriks terkenal, seperti akurasi, perolehan (juga dikenal sebagai sensitivitas), presisi, dan skor F1.<sup>44</sup> Tabel 3 menampilkan seberapa baik kinerja GBM, XGBoost, dan LightGBM dalam hal teknik pembelajaran mesin. Namun,

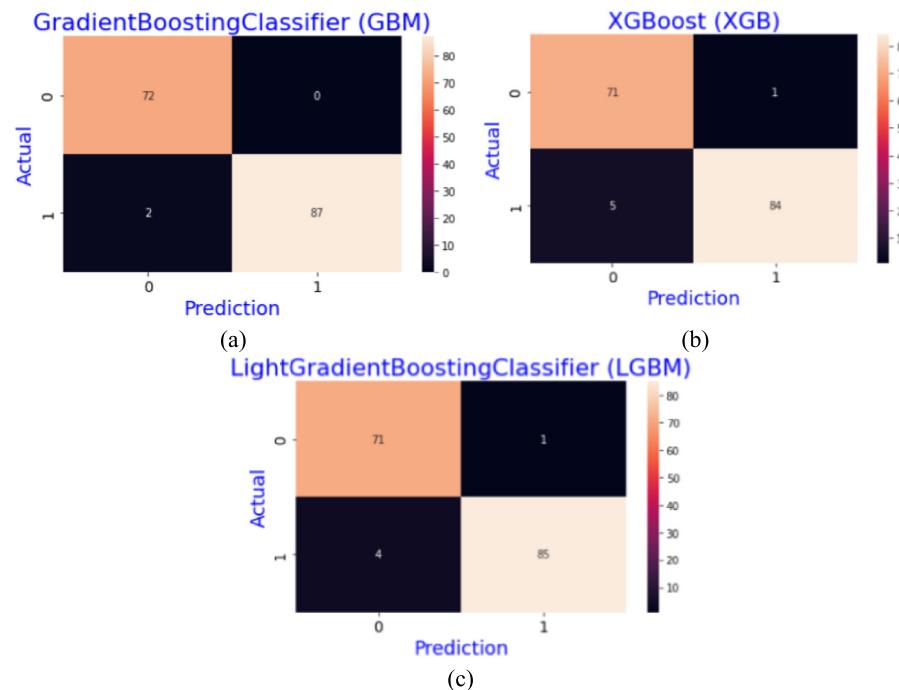
Tabel 3

Evaluasi metode pembelajaran mesin yang dapat dijelaskan.

Metode	Presisi	Mengingat	F_Ukur	Ketepatan	Kesalahan
GBM	98,79%	98,76%	98,76%	98,76%	0,012%
XGB	96,41%	96,27%	96,28%	96,27%	0,037%
LGBTM	96,97%	96,89%	96,89%	96,89%	0,031%



Gambar 5. Akurasi untuk GBM, XGB, dan LGBM.



Gambar 6. Matriks kebingungan untuk: (a) GBM dan (b) XGB, dan (c) LGBM.

kami telah menyertakan hasil Akurasi, Presisi, Perolehan, dan Skor F1 untuk tujuan mengamati kinerja model.

Penelitian perbandingan menunjukkan bahwa pengklasifikasi GBM mengungguli pengklasifikasi lainnya dengan akurasi 98,76%. Semua metrik menunjukkan kinerja rendah untuk pengklasifikasi XGB. Dan di antara pengklasifikasi tersebut, LGBM menduduki peringkat kedua tertinggi. Karena akurasi merupakan indikator data seimbang yang andal, maka akurasi dianggap sebagai metrik performa utama eksperimen. Namun, Mesin Peningkat Gradien mencapai akurasi paling seimbang dalam penyelidikan ini, seperti yang ditunjukkan pada Gambar 5.

#### Evaluasi model

Komponen kunci dalam menciptakan model pembelajaran mesin yang hebat adalah evaluasi model. Dalam eksperimen ini, berbagai metrik evaluasi, seperti matriks konfusi (akurasi, presisi, perolehan, ukuran F, Error), dan kurva AUC-ROC, digunakan untuk menilai performa atau kaliber model. Jumlah prediksi benar-positif, benar-negatif, positif palsu, dan negatif palsu yang dibuat oleh algoritme ditentukan oleh matriks konfusi. True Positives adalah jumlah kejadian di mana algoritma berhasil mengidentifikasi kelas positif, sedangkan True Negatives adalah jumlah kejadian di mana metode mengantisipasi kelas negatif dengan tepat. Positif palsu adalah jumlah kejadian ketika algoritme memprediksi kelas positif padahal kelas sebenarnya negatif, dan negatif palsu adalah jumlah kejadian ketika sistem memprediksi kelas negatif padahal kelas sebenarnya positif. Berbagai indikator kinerja, termasuk akurasi, presisi, perolehan, dan skor F1, dapat dihitung menggunakan matriks.<sup>45</sup> Gambar 6 menunjukkan matriks konfusi dari setiap pengklasifikasi.

Metrik kinerja ini memungkinkan kami menilai seberapa baik model kami memproses data yang diberikan. Matriks evaluasi ini didefinisikan dalam Persamaan(1)-(5).

$$\text{Ketepatan} \% = \frac{\text{dilthTN}}{\text{dilthTN} + \text{FPthFN}} \times 100 \quad (1)$$

$$\text{Presisi} \% = \frac{\text{dil}}{\text{dil} + \text{FP}} \times 100 \quad (2)$$

$$\text{Mengingat} \% = \frac{\text{dil}}{\text{dil} + \text{FN}} \times 100 \quad (3)$$

$$\text{F_Ukur} \% = \frac{2 \cdot \text{Ingat} \cdot \text{Presisi}}{\text{Mengingat} + \text{Presisi}} \times 100 \quad (4)$$

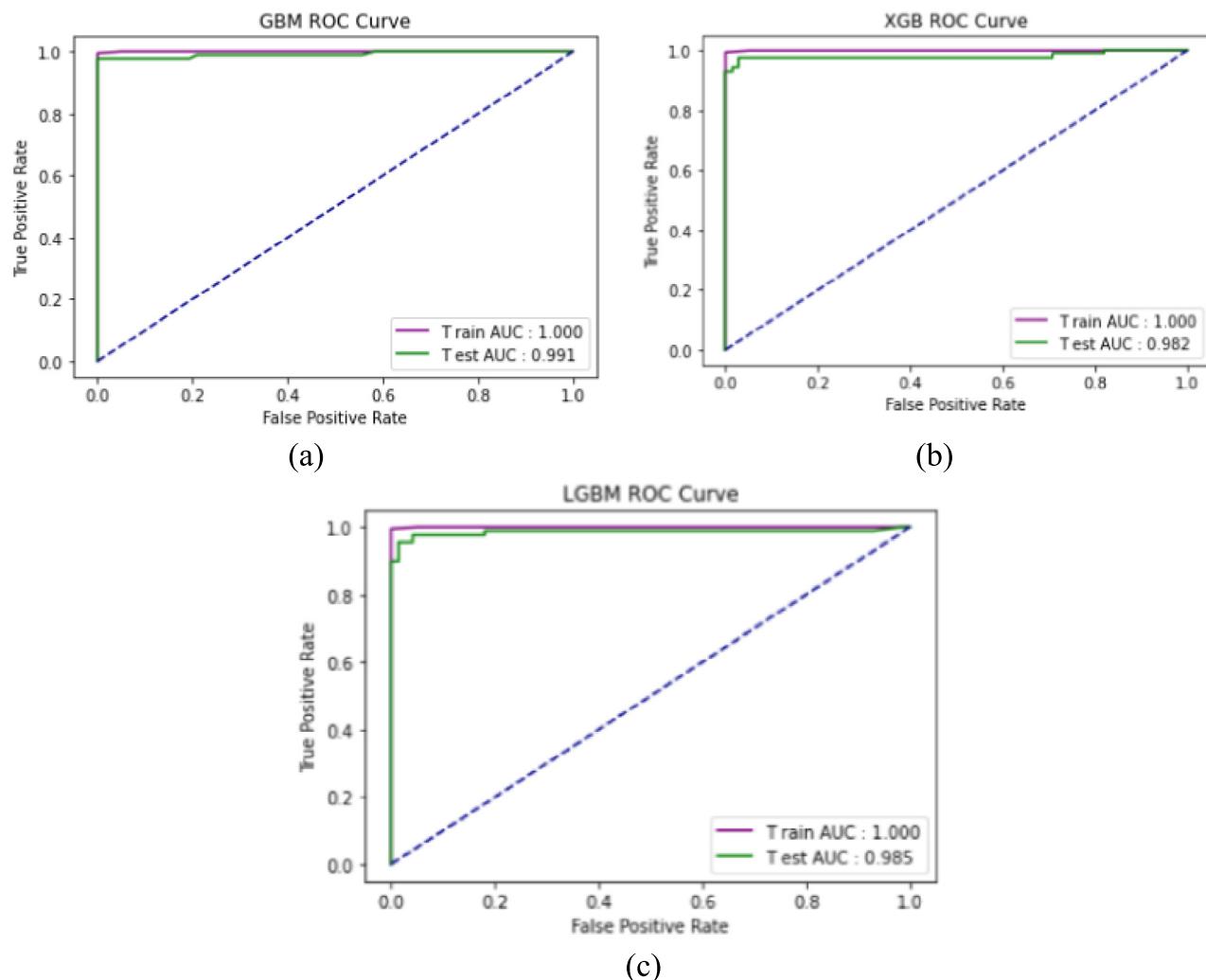
$$\text{Kesalahan} \% = \frac{\text{FPthFN}}{\text{dil} + \text{TN} + \text{FPthFN}} \times 100 \quad (5)$$

Dimana TP, FP, TN, dan FN masing-masing mewakili True Positives, False Positives, dan True Negatives. Kurva AUC-ROC ditunjukkan pada Gambar 7. Di sini, kurva AUC-ROC digunakan untuk menunjukkan seberapa baik kinerja model klasifikasi pada grafik. Ini adalah statistik yang disukai dan penting untuk mengevaluasi seberapa baik model kategorisasi bekerja.

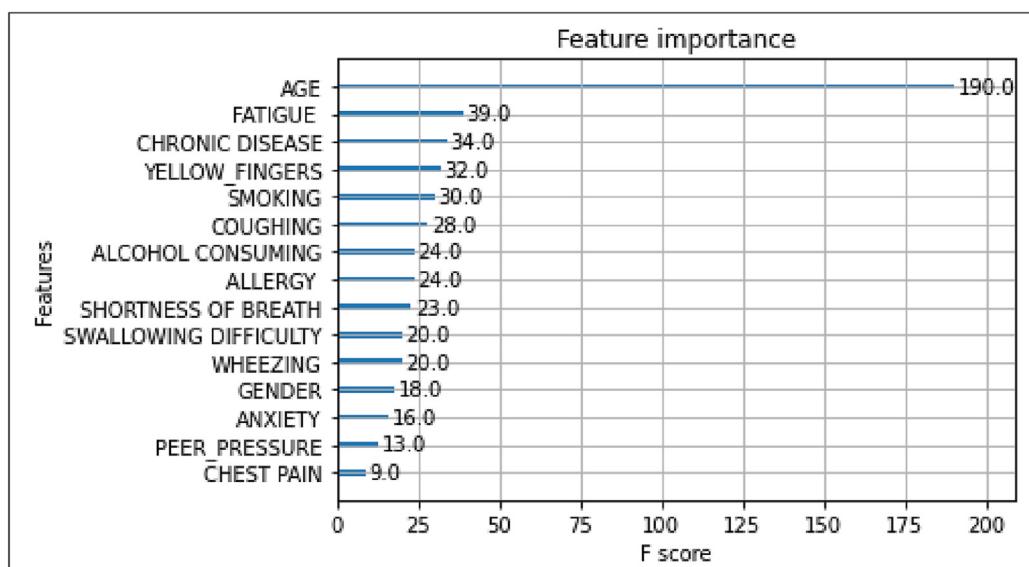
Masalah klasifikasi biner dapat dievaluasi menggunakan kurva ROC sebagai statistik. Kurva probabilitas ini, yang pada dasarnya membedakan "sinyal" dari "kebisingan", menampilkan TPR (True Positive Rate) versus FPR (False Positive Rate) pada nilai ambang batas yang berbeda. Kurva ROC diringkas menggunakan Area Under the Curve (AUC), yang mengukur kapasitas pengklasifikasi untuk membedakan kelas. Performa model dalam memisahkan kelas positif dan negatif berkorelasi terbalik dengan AUC.

#### Analisis hasil SHAP

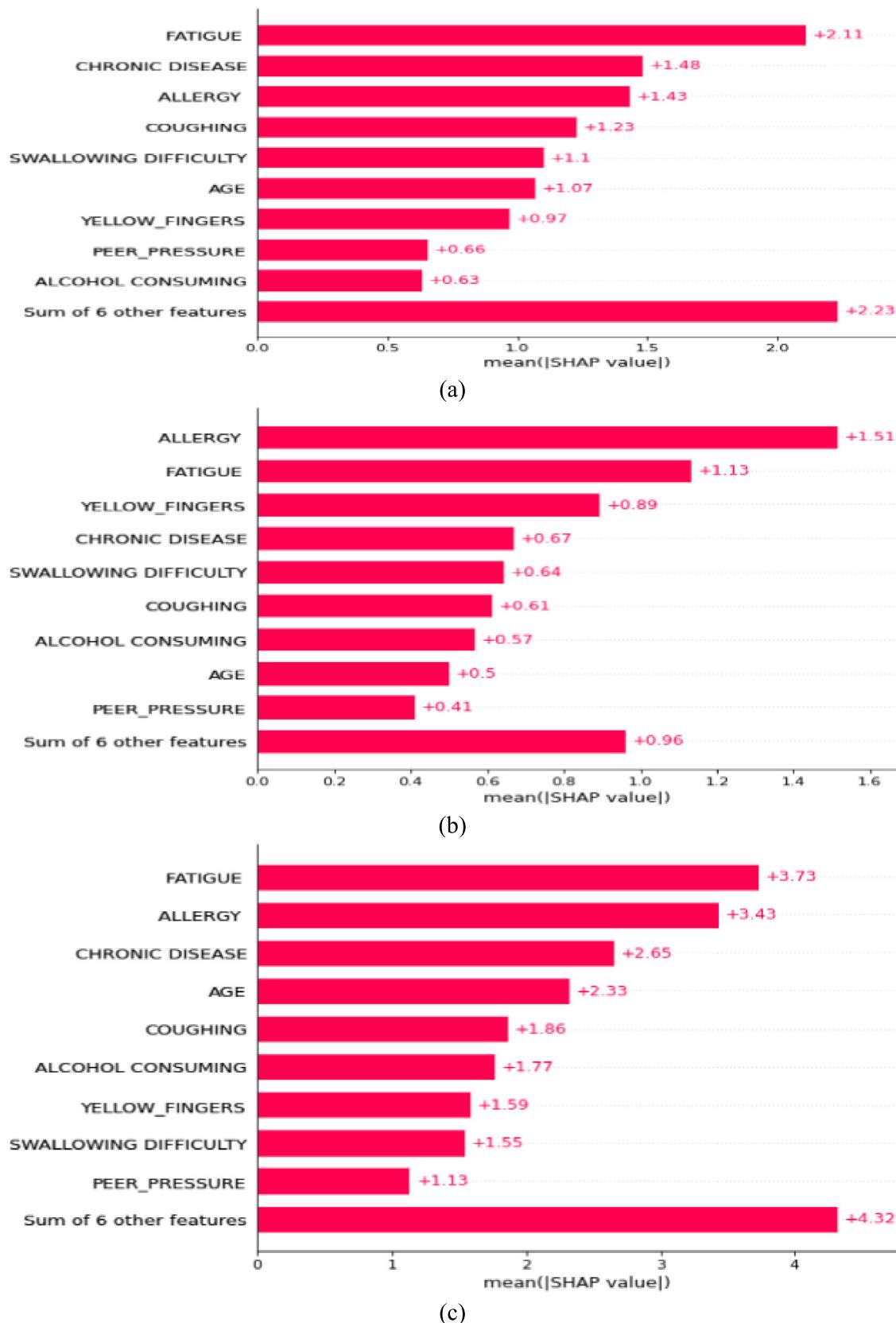
Terakhir, nilai faktor penjelasnya digunakan untuk menentukan penjelasan nilai Shapley tentang kanker paru-paru di set tes. Dalam disiplin pembelajaran mesin, semakin mudah dijelaskan suatu model, semakin mudah untuk memahami dan memahami prediksi yang telah dibuat. Untuk menjelaskan keluaran model dan menentukan sejauh mana suatu karakteristik tertentu berkontribusi terhadap hasil suatu peristiwa tertentu, SHAP digunakan di sini. Ini adalah alat canggih untuk kepentingan fitur yang memungkinkan untuk memahami fitur mana yang paling penting dalam mendorong prediksi tertentu,



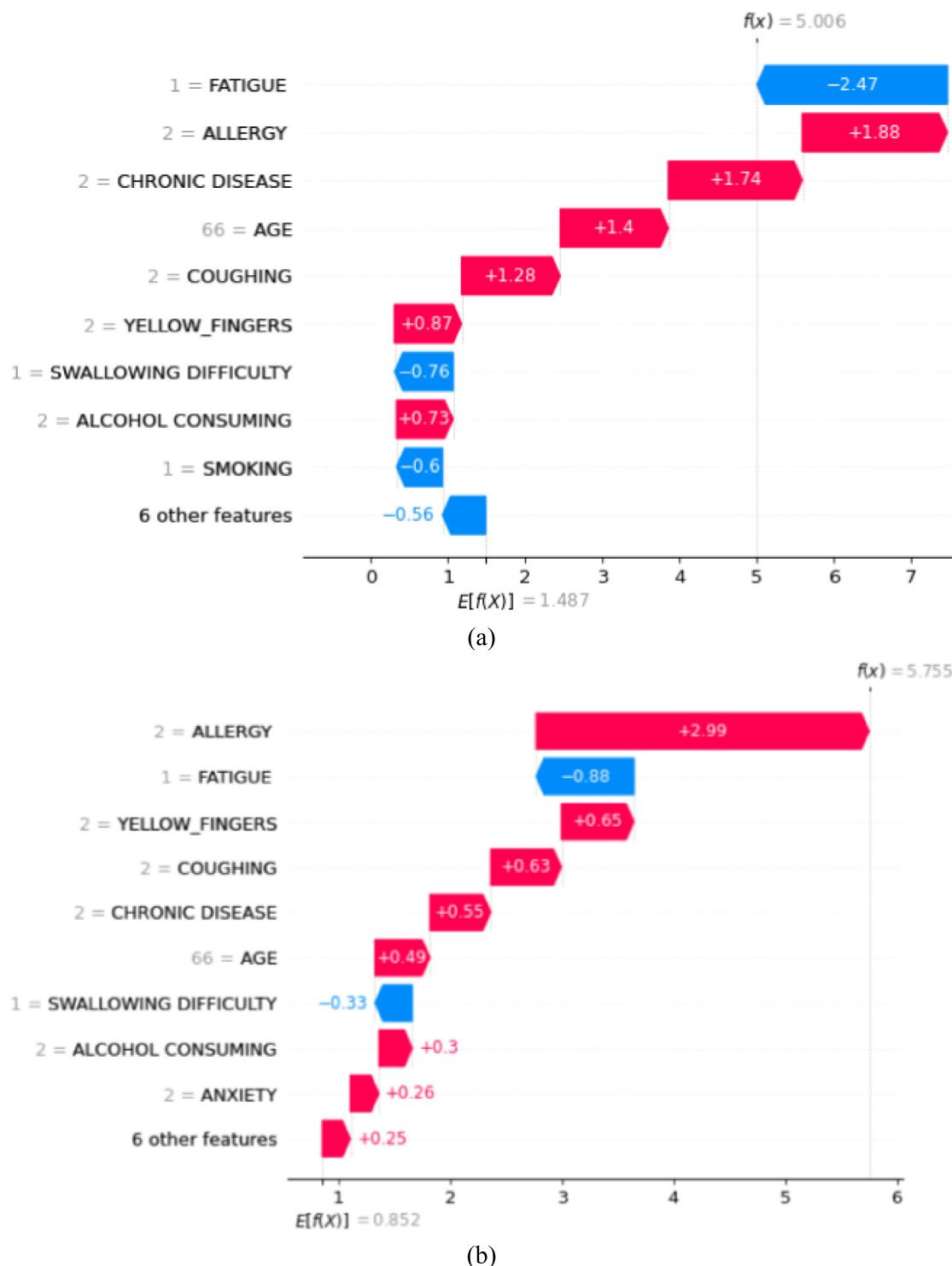
Gambar 7.Kurva ROC: (a) GBM, (b) XGB, dan (c) LGBM.



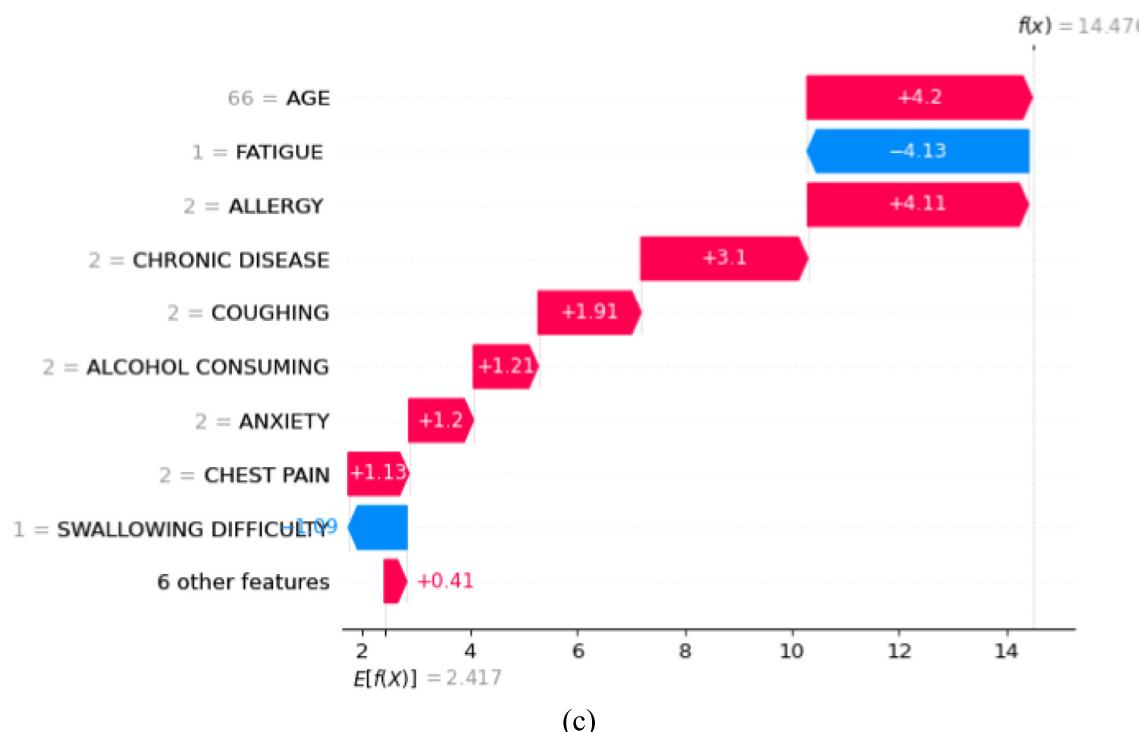
Gambar 8.Menampilkan pentingnya untuk prediksi kanker paru-paru.



Gambar 9. Plot SHAP Bar untuk: (a) GBM, (b) XGBoost, dan (c) LGBM.



Gambar 10. Plot air terjun SHAP untuk: (a) GBM, (b) XGB, dan (c) LGBM.



(c)

Gambar 10 (lanjutan).

dan bagaimana interaksi fitur yang berbeda berkontribusi terhadap prediksi model secara keseluruhan. Strategi kepentingan fitur dikembangkan sebagai hasil analisis ini. **Gambar 8** menggambarkan pentingnya variabel.<sup>46</sup>

Grafik tersebut menunjukkan bahwa AGE, yang menyumbang 190,0 dari total keseluruhan, merupakan kontributor utama dalam prediksi kanker paru-paru. KElelahan merupakan ciri terpenting berikutnya, disusul PENYAKIT KRONIS, JARI KUNING, dan seterusnya, yang disusun menurut signifikansinya. Faktor yang memberikan kontribusi paling kecil, dengan faktor 9,0, adalah CEST PAIN.

Lebih banyak plot grafik SHAP tersedia untuk membantu pemahaman manusia tentang hasil yang diharapkan. Plot batang GBM, XGBoost, dan LGBM ditampilkan di**Gambar 9**.

Kepentingan masing-masing model berbeda-beda. Komponen terpenting GBM, seperti yang ditunjukkan pada**Gambar 9(a)**, adalah FATIGUE, yang berarti nilai SHAP absolut yang jauh lebih tinggi dibandingkan karakteristik lainnya. PENYAKIT KRONIS menyumbang jumlah tertinggi kedua (+1.48). FATIGUE adalah kontributor sekunder untuk XGBoost, menurut 9(b), sedangkan ALERGI adalah kontributor utama, terhitung +1,51. Mirip dengan GBM, FATIGUE memainkan peran penting untuk 9(c), diikuti oleh ALERGI, PENYAKIT KRONIS, dan kondisi lainnya.

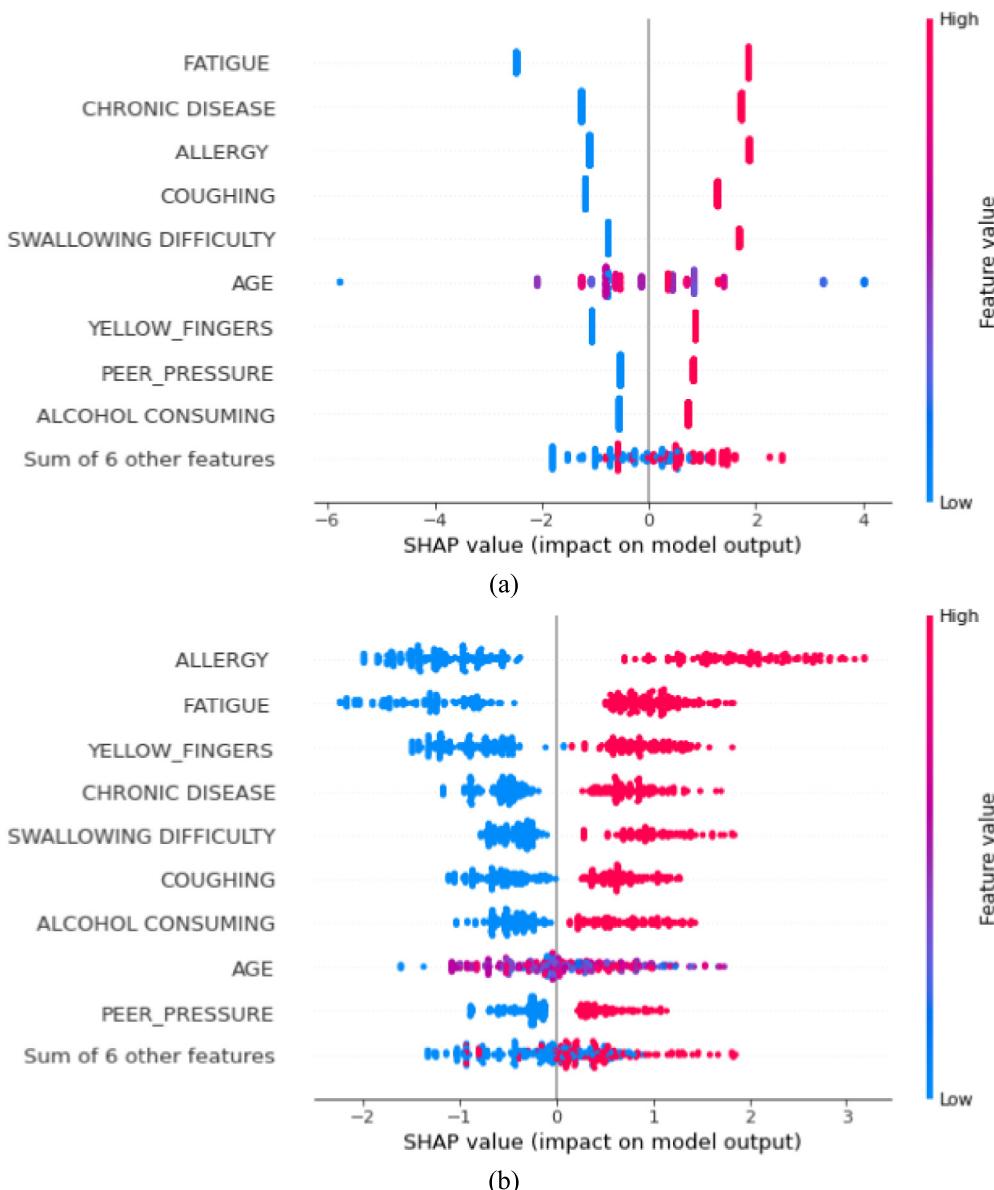
Teknik visualisasi lainnya adalah Waterfall Plot, yang memvisualisasikan kontribusi setiap fitur terhadap prediksi. Dalam Plot Air Terjun SHAP, fitur-fitur dicantumkan di sepanjang sumbu x dan nilai SHAP diwakili oleh batang yang membentang dari garis dasar (biasanya nol) hingga prediksi akhir. Nilai SHAP positif menunjukkan bahwa fitur tersebut berdampak positif terhadap prediksi, sedangkan nilai SHAP negatif menunjukkan bahwa fitur tersebut berdampak negatif. Ketinggian setiap batang mewakili besarnya kontribusi fitur terhadap prediksi. Ini memberikan cara yang jelas dan intuitif untuk memahami bagaimana setiap fitur berkontribusi terhadap prediksi dan bagaimana fitur tersebut berinteraksi

satu sama lain. Hal ini dapat membantu mengidentifikasi fitur mana yang paling penting, mana yang memiliki dampak positif atau negatif paling kuat, dan bagaimana kaitannya dengan prediksi akhir. **Gambar 10** mewakili plot air terjun untuk GBM, XGBoost, dan LGBM.

**Gambar 10(a)** menunjukkan bahwa FATIGUE memiliki nilai SHAP sebesar -2,47 yang berpengaruh negatif terhadap prediksi, sedangkan ALERGI berpengaruh positif terhadap prediksi dengan nilai SHAP sebesar +1,88, dan seterusnya. Semua nilai SHAP yang dijumlahkan akan sama dengan  $E[f(x)] - f(x)$ . ALERGI mempunyai pengaruh positif sebesar +2,99 terhadap prediksi **Gambar 10(B)**. Untuk memprediksi kanker paru, FATIGUE mempunyai dampak negatif sebesar -0,88 dan YELLOW\_FINGERS mempunyai dampak positif sebesar +0,65 untuk XGBoost. Sebagai perbandingan, AGE mempunyai kontribusi positif paling besar terhadap prediksi 10(c), sedangkan kontribusi gabungan dari 6 variabel lainnya adalah yang paling kecil.

Teknik plot SHAP lainnya adalah Beeswarm yang ditunjukkan pada **Gambar 11**. Ini adalah jenis plot sebar yang digunakan untuk menampilkan distribusi sejumlah besar observasi individu dengan cara meminimalkan tumpang tindih antar titik. Dalam plot ini, titik-titik data diwakili oleh titik-titik kecil yang ditempatkan sepanjang sumbu x, dengan sumbu y menunjukkan kepadatan titik-titik tersebut. Titik-titik tersebut disusun sedemikian rupa sehingga sedekat mungkin dengan nilai x tanpa tumpang tindih.

FATIGUE biasanya merupakan komponen paling signifikan untuk GBM, seperti yang ditunjukkan pada**Gambar 11(A)**. Kemudian, PENYAKIT\_KRONIS dan ALERGI masing-masing merupakan faktor terpenting kedua dan ketiga dalam prognosis. Di sisi lain, ALERGI memberikan kontribusi paling besar terhadap prediksi stroke untuk 11(b). Kemungkinan mengalami ramalan juga akan meningkat seiring dengan meningkatnya kadar ALERGI. Untuk XGBoost, FATIGUE dan YELLOW\_FINGERS masing-masing merupakan faktor risiko tertinggi kedua dan ketiga untuk kanker paru-paru. **Gambar 11(c)** menunjukkan bahwa FATIGUE mempunyai pengaruh paling besar terhadap prediksi. ALERGI dan CHRONIC\_DISEASE adalah 2 ciri paling krusial berikutnya.

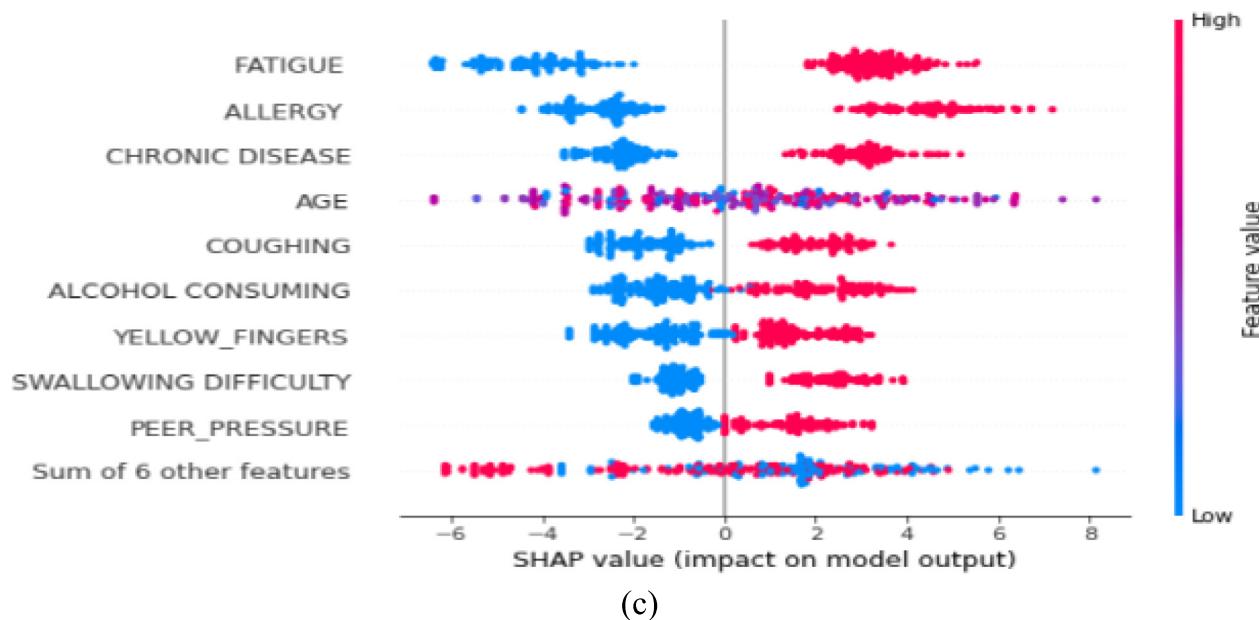


Gambar 11. Plot SHAP Beeswarm untuk: (a) GBM, (b) XGB, dan (c) LGBTM.

Seperti yang dapat kita lihat dari bagian ini, GBM memperoleh akurasi maksimum 98,76% dengan presisi 98,79%, recall 98,76%, dan F-measure 98,76% dengan tingkat kesalahan 0,012%. Tanda-tanda ini secara kolektif menunjukkan bahwa deteksi kanker paru-paru dapat dimodelkan menggunakan GBM dan secara keseluruhan sangat signifikan. Terakhir, set pengujian 35% digunakan untuk menguji GBM setelah dilatih ulang pada seluruh set pelatihan 65%. Data untuk kanker paru-paru telah terbukti mendapat manfaat dari penggunaan nilai SHAP untuk penjelasan model. Menarik untuk dicatat bahwa karya ini menunjukkan bagaimana gagasan pentingnya fitur berdasarkan nilai absolut SHAP dapat diperluas untuk digunakan sebagai metode pemilihan fitur. Properti yang dapat dijelaskan, yang digunakan dalam pendekatan ini, dapat bermanfaat untuk pemilihan fitur, sebuah langkah pra-pemrosesan yang umum dalam pembelajaran mesin. Kami mengantisipasi pemilihan fitur berdasarkan

Nilai-nilai SHAP akan menjadi strategi populer di kalangan praktisi pembelajaran mesin.

Namun, banyak penelitian telah dilakukan mengenai hal ini oleh banyak peneliti dengan menggunakan metode yang beragam dan menghasilkan berbagai temuan. Perkiraan kanker yang akurat sangat penting karena kanker paru-paru menyerang banyak orang di seluruh dunia. Pentingnya kanker paru-paru menginspirasi kami untuk membahas topik ini. Studi ini berfokus pada penggunaan XML untuk meramalkan kanker paru-paru dan menunjukkan implementasi yang berguna (aplikasi seluler) yang dapat memprediksi kanker berdasarkan masukan yang diberikan. Perbandingan pekerjaan ini dengan penelitian sebelumnya disediakan di Tabel 4 untuk memahami pengetahuan yang ada mengenai suatu topik dan mengidentifikasi keserjangan dalam literatur yang dapat diatasi oleh penelitian mereka.



Gambar 11 (lanjutan).

Tabel 4

Perbandingan karya kami dengan karya yang paling terkait.

Pengarang	Himpunan data	Model yang diusulkan	Pertunjukan
Prediksi kanker paru-paru (XML) Sobhan dkk., <sup>1</sup> (2022)	Basis data UCSC Xena, 1415 contoh	XGBoost	Akurasi: 96,3%
Alsinglawi dkk., <sup>2</sup> (2022)	Data MIMIC-III, 53.423 contoh	Federasi Rusia Model bekas BACH, PLCoM2012, dan LCART.	AUC: 98% (95,3%-100%), Penarikan: 98% (95,3%-100%). Tidak disebutkan performa dan akurasinya. Hanya fokus pada memahami bagaimana model bertindak untuk berbagai pasien.
Katarzyna dkk., <sup>3</sup> (2022)	dataset Kanker Paru Domestik, 34 393 individu		
Jamie dkk., <sup>4</sup> 2021	Kumpulan data simularcum, 1 322 100 contoh	XGBoost	Presisi: 78%, Penarikan:78% Akurasi: 78%
Prediksi kanker paru-paru (ML) Elias Dritsas dan Maria Trigka, <sup>5</sup> 2022	Kumpulan data Kaggle, 309 contoh	Hutan rotasi (RotF)	AUC: 99,3%, F-Measure, presisi, recall, dan akurasi: 97,1%.
Muntasir dkk., <sup>6</sup> 2022	Kumpulan data Kaggle, 309 contoh	XGBoost	AUC: 98,14%, Presisi:95,66% Akurasi:94,42% Penarikan: 94,46%
Patra, <sup>7</sup> 2020	Repositori UCI, contoh 32	Jaringan Fungsi Basis Radial	Akurasi: 81,25% Skor F: 81,3% AUC: 74,9% Presisi:81,3% Penarikan: 81,3% AUC: 94,9% Akurasi: 94,8% Akurasi: 98,76% F-score: 98,76% AUC: (train-1.0, test-0,991) Presisi: 98,79% Penarikan: 98,76%
Sim dkk., <sup>8</sup> 2020 Diusulkan	Data HRQOL, 809 individu Kaggle, 309 instance dan total 16 fitur	AdaBoost GBM (XML)	

## Implementasi aplikasi seluler

Komponen praktis dari penelitian ini ditunjukkan pada bagian ini. Aplikasi aplikasi percobaan yang dibuat menggunakan model terbaik digambarkan pada Gambar 12. React Native digunakan untuk membangun aplikasi ini. Program ini memiliki formulir umpan balik pengguna dengan kolom masukan yang memperkirakan kanker payudara dan mengumpulkan komentar pengguna. Model ini awalnya dibuat sebagai file pkl di notebook Jupyter setelah itu aplikasi flask digunakan untuk membuat api. Api ini digunakan untuk menambahkan model pembelajaran mesin ke aplikasi Android, dan hasilnya kemudian ditampilkan di layar. Gambar 12(a) dan (b) menggambarkan aplikasi seluler tempat siapa pun dapat mengirimkan masukan untuk memperkirakan hasilnya. Gambar 12(c) dan (d) menampilkan keluaran setelah memasukkan masukan.

## Kesimpulan dan pekerjaan masa depan

Dalam penelitian ini, kami telah melakukan upaya untuk menutup kesenjangan mengenai interpretasi model risiko ini. Di berbagai sektor, algoritma pembelajaran mesin yang dapat dijelaskan telah memperoleh daya tarik yang signifikan. Meskipun kanker paru-paru telah diatasi dengan menggunakan teknik XML. Namun, kami juga telah menunjukkan cara menggunakan teknik pembelajaran mesin yang dapat dijelaskan untuk memprediksi kanker paru-paru. Pendekatan XML menawarkan wawasan penelitian mengenai karakteristik yang paling penting untuk prognosis kanker. Untuk ini, GBM, XGBoost, dan LGBM dijelaskan menggunakan SHAP. Kemampuan untuk memenuhi berbagai kualitas yang diinginkan, seperti konsistensi, lokalitas, dan hilangnya, menjadikan SHAP pilihan populer untuk interpretasi model. Sangat penting untuk menunjukkan bagaimana cara kerja sistem medis.

The figure consists of four screenshots labeled (a), (b), (c), and (d).

- (a)**: A form for entering patient information. Fields include Name (Ayonti), Age (59), Gender (Female), Smoking (No), Yellow Fingers (No), Anxiety (No), and Peer Pressure (Yes). Buttons at the bottom are Home, Predict, and Result.
- (b)**: A form for entering symptoms. Fields include Alcohol Consuming (No), Coughing (Yes), Shortness of Breath (Yes), and Swallowing Difficulty (No). A Predict button is at the bottom. Buttons at the bottom are Home, Predict, and Result.
- (c)**: A screenshot showing a modal dialog with the message "Ayonti, don't worry, You don't have lung cancer!" and an OK button. Below the dialog are fields for Wheezing (Select Wheezing), Alcohol Consuming (Select Alcohol Consuming), Coughing (Ayonti), Swallowing Difficulty (Select Swallowing Difficulty), and Chest Pain (Select Chest Pain). Buttons at the bottom are Home, Predict, and Result.
- (d)**: A screenshot showing a summary table of input features and their values, followed by a message and two buttons. The table is as follows:

	ANXIETY	No
PEER PRESSURE	Yes	
CHRONIC DISEASE	No	
FATIGUE	Yes	
ALLERGY	No	
WHEEZING	Yes	
ALCOHOL CONSUMING	No	
COUGHING	Yes	
SHORTNESS OF BREATH	Yes	
SWALLOWING DIFFICULTY	No	
CHEST PAIN	Yes	

A message below the table says "Ayonti, don't worry, You don't have lung cancer!". Buttons at the bottom are Delete and Back.

Gambar 12.Bidang masukan aplikasi Android (a,b) (c,d) bidang hasil.

Dalam SHAP, relevansi fitur ditentukan oleh kontribusinya terhadap keluaran model, apa pun model yang digunakan. Kontribusi ini digunakan dalam kasus ini sebagai pendekatan pemilihan fitur dan untuk menentukan peringkat

fitur dalam hal relevansi. Dalam percobaan ini, SHAP terbukti lebih unggul dibandingkan metode pemilihan fitur populer lainnya. Temuan ini menunjukkan bahwa penggunaan SHAP sebagai mekanisme pemilihan fitur dapat menjadi strategi yang baik

solusi pembelajaran mesin yang perlu dapat ditafsirkan. Kami bermaksud untuk mengkaji lebih lanjut SHAP menggunakan pembelajaran mendalam dalam karya selanjutnya. Selain itu, analisis gambar dengan kemampuan menjelaskan akan dilakukan.

## Kontribusi Penulis

STR, NB, dan KMMU bertanggung jawab atas konseptualisasi dan desain penelitian. Mereka juga memiliki akses penuh terhadap semua data penelitian dan menerima tanggung jawab atas keakuratan pembuatan model dan data penelitian. Semua kontributor bekerja sama untuk menulis artikel. Laporan tersebut direvisi secara kritis dengan masukan dari STR, SKD, dan lain-lain. Seluruh hasil dan teknik penyajian data dihasilkan oleh NB dan KMMU. Versi final telah ditinjau dan disetujui oleh semua penulis, yang juga berkontribusi terhadap pengumpulan dan analisis data.

## Pendanaan

Tidak ada.

## Persetujuan Etis

Tidak dibutuhkan.

## Persetujuan untuk berpartisipasi

Tidak dibutuhkan.

## Ketersediaan data

Berdasarkan permintaan yang masuk akal, penulis terkait akan memberikan data yang mendukung temuan penelitian ini.

## Deklarasi kepentingan

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

## Referensi

- Cassidy A, Duffy SW, Myles JP, Liloglou T, Lapangan JK. Prediksi risiko kanker paru-paru: Alat untuk deteksi dini. *Jurnal Internasional Kanker* 1 Jan 2007;120(1):1–6.<https://doi.org/10.1002/ijc.22331>.
- Sistem Prediksi Diagnosis Kanker Paru Menggunakan Teknik Klasifikasi Data Mining. [On line]. Tersedia:[www.ijcsit.com](http://www.ijcsit.com).
- Qiang Y, Guo Y, Li X, Wang Q, Chen H, Cuic D. Studi Pendahuluan Aturan Diagnostik Kanker Paru Perifer Berdasarkan Teknik Data Mining. [On line]. Tersedia:[www.elsevier.com/loc/jnmu](http://www.elsevier.com/loc/jnmu).
- Shopland DR, Eyre HJ, Pechacek TF. ARTIKEL Kematian Akibat Kanker Akibat Merokok pada Tahun 1991: Apakah Kanker Paru Kini Menjadi Penyebab Utama Kematian Perokok di Amerika Serikat? [On line]. Tersedia:<http://jnci.oxfordjournals.org>.
- Karabatak M, Ince MC. Sebuah sistem paka untuk mendeteksi kanker payudara berdasarkan aturan asosiasi dan jaringan saraf. *Aplikasi Sistem Pakar* 2009;36(2 BAGIAN 2):3465–3469. <https://doi.org/10.1016/j.eswa.2008.02.064>.
- Stokowaty W, WojtaSB, Krajewska J, Stobiecka E, Dralle H, Musholt T, dkk. Pengklasifikasi dua miRNA membentuk karisoma tiroid folikular dari adenoma tiroid folikular. *Endokrinol Sel Mol* Januari 2015;399:43–49.<https://doi.org/10.1016/j.mce.2014.09.017>.
- Zhang R, bin Huang G, Sundararajan N, Saratchandran P. Klasifikasi multikategori menggunakan mesin pembelajaran ekstrim untuk diagnosis kanker ekspresi gen microarray. *IEEE/ACM Trans Comput Biol Bioinform* Juli 2007;4(3):485–494.<https://doi.org/10.1109/TCBB.2007.1012>.
- Wang Y, dkk. Seleksi gen dari data microarray untuk klasifikasi kanker - pendekatan pembelajaran mesin. *Comput Biol Chem* Februari 2005;29(1):37–46.<https://doi.org/10.1016/j.combiolchem.2004.11.001>.
- Kim B, dkk. Interpretabilitas Melampaui Atribusi Fitur: Pengujian Kuantitatif dengan Vektor Aktivasi Konsep (TCAV).2018.
- Tan S, Caruana R, Hooker G, Lou Y. Distill-and-compare: mengaudit model kotak hitam menggunakan distilasi model transparan. *AIES 2018 - Prosiding Konferensi AAAI/ACM 2018 tentang AI, Etika, dan Masyarakat; Desember 2018.* hal. 303–310.<https://doi.org/10.1145/3278721.3278725>.
- Ribeiro MT, Singh S, Guestrin C. 'Mengapa saya harus mempercayai Anda?' Menjelaskan prediksi pengklasifikasi apa pun. *Prosiding Konferensi Internasional ACM SIGKDD tentang Penemuan Pengetahuan dan Penambangan Data*, 13-17 Agustus-2016. Agustus 2016. hal. 1135–1144.<https://doi.org/10.1145/2939672.2939778>.
- Lundberg SM, Allen PG, Lee SI. Pendekatan Terpadu untuk Menafsirkan Prediksi Model. [On line]. Tersedia:<https://github.com/slundberg/shap>.
- Elshawi R, Al-Mallah MH, Sakr S. Tentang interpretabilitas model berbasis pembelajaran mesin untuk memprediksi hipertensi. *BMC Med Informasiakan Decis Mak* Juli 2019;19(1). <https://doi.org/10.1186/s12911-019-0874-0>.
- Ibrahim M, Louie M, Paisley J. Penjelasan Global Jaringan Neural Memetakan Lanskap Prediksi Pusat Pembelajaran Mesin Ceena Modarres, Capital One.2019. <https://doi.org/10.1145/aies2020>.
- Whitmore LS, George A, Hudson CM. Memetakan kinerja kimia pada struktur molekul menggunakan penjelasan yang dapat ditafsirkan secara lokal.November 2016.[Online]. Tersedia: <http://arxiv.org/abs/1611.07443>.
- Phillips PJ, dkk. Empat Prinsip Kecerdasan Buatan yang Dapat Dijelaskan Gaithersburg, MD. September 2021.<https://doi.org/10.6028/NIST.IR.8312>.
- Strumbelj E, Kononenko I. Menjelaskan model prediksi dan prediksi individu dengan kontribusi fitur. *Knowl Inf Syst* November 2014;41(3):647–665.<https://doi.org/10.1007/s10115-013-0679-x>.
- Cosgriff CV, Celi LA. Memanfaatkan hubungan temporal dalam prediksi kematian. *Lancet Digital Health* 1 April 2020;2(4):e152–e153.[https://doi.org/10.1016/S2589-7500\(20\)30056-X](https://doi.org/10.1016/S2589-7500(20)30056-X).Elsevier Ltd.
- Lundberg SM, dkk. Dari penjelasan lokal hingga pemahaman global dengan AI yang dapat dijelaskan untuk pemohonan. *Nat Mach Intell* Januari 2020;2(1):56–67.<https://doi.org/10.1038/s42256-019-0138-9>.
- Lundberg SM, dkk. Prediksi pembelajaran mesin yang dapat dijelaskan untuk pencegahan hipoksemia selama operasi. *Nat Biomed Eng* Oktober 2018;2(10):749–760.<https://doi.org/10.1038/s41551-018-0304-0>.
- Singh A, Sengupta S, Lakshminarayanan V. Model pembelajaran mendalam yang dapat dijelaskan dalam analisis citra medis. *J Pencitraan* 20 Juni 2020;6(6).<https://doi.org/10.3390/JIMAGING6060052>.MDPI AG.
- Xi J, Zhao W, Yuan JE, Cao B, Zhao L. Klasifikasi multi-resolusi gambar aerosol yang dihembuskan untuk mendeteksi penyakit paru obstruktif di saluran udara kecil. *Comput Biol Med* Agustus 2017;87:57–69.<https://doi.org/10.1016/j.combiomed.2017.05.019>.
- Li W, Jia Z, Xie D, Chen K, Cui J, Liu H. Mengenali kanker paru-paru menggunakan enose buatan sendiri: Sebuah studi komprehensif. *Hitung Biol Med* Mei 2020;120.<https://doi.org/10.1016/j.combiomed.2020.103706>.
- M. Sobhan dan AM Mondal, "Pembelajaran Mesin yang Dapat Dijelaskan untuk Mengidentifikasi Biomarker Khusus Pasien untuk Kanker Paru",<https://doi.org/10.1101/2022.10.13.512119>.
- Alsinglawi B, dkk. Kerangka kerja pembelajaran mesin yang dapat dijelaskan untuk prediksi lama rawat inap di rumah sakit kanker paru-paru. *Perwakilan Sains Desember* 2022;12(1).<https://doi.org/10.1038/s41598-021-04608-7>.
- Kobylinska K, Ataulowski T, Adamek M, Biecek P. Pembelajaran mesin yang dapat dijelaskan untuk model skrining kanker paru-paru. *Appl Sci (Swiss)* Februari 2022;12(4).<https://doi.org/10.3390/applkasi12041926>.
- Duell J, Fan X, Burnett B, Aarts G, Zhou SM. Perbandingan Penjelasan yang Diberikan Metode Kecerdasan Buatan yang Dapat Dijelaskan dalam Menganalisis Catatan Kesehatan Elektronik. [On line]. Tersedia:<http://www.ncbi.nlm.nih.gov/about>.
- Dritsas E, Trigka M. Prediksi risiko kanker paru-paru dengan model pembelajaran mesin. *Big Data dan Komputasi Kognitif* November 2022;6(4):139.<https://doi.org/10.3390/bdcc6040139>.
- Mamun M, Farjana A, al Mamun M, Ahammed MS. Model prediksi kanker paru menggunakan teknik pembelajaran ensemble dan analisis sistematis review. *Kongres AI IoT Dunia IEEE 2022, AIoT 2022*; 2022. hal. 187–193.<https://doi.org/10.1109/AIoT54504.2022.9817326>.
- Patra R. Prediksi kanker paru-paru menggunakan pengklasifikasi pembelajaran mesin. *Komunikasi dalam Ilmu Komputer dan Informasi. CCIS*, 2020. hal. 132–142.[https://doi.org/10.1007/978-981-15-6648-6\\_11](https://doi.org/10.1007/978-981-15-6648-6_11).
- ah Sim J, dkk. Dampak besar kualitas hidup terkait kesehatan terhadap prediksi kelangsungan hidup 5 tahun di kalangan penderita kanker paru-paru: penerapan pembelajaran mesin. *Sci Rep* Desember 2020;10(1).<https://doi.org/10.1038/s41598-020-67604-3>.
- Dataset Prediksi Kanker Paru. Tersedia daring:<https://www.kaggle.com/datasets/mysarahmadbh/lung-cancer?fbclid=IwAR0uQ5K3mEbQZJcwQGYqlJRSydvsk2oU1Sav5YvIt0ECqkx6-vPR43JAM>.
- Ahmed N, dkk. Prediksi diabetes berbasis pembelajaran mesin dan pengembangan aplikasi web pintar. *Int J Cognit Comput Eng* Juni 2021;2:229–241.<https://doi.org/10.1016/j.jicce.2021.12.001>.
- Mohammed R, Rawashdeh J, Abdullah M. Pembelajaran mesin dengan teknik oversampling dan undersampling: studi ikhtisar dan hasil eksperimen. *Konferensi Internasional Sistem Informasi dan Komunikasi ke-11 tahun 2020, ICICS 2020; April 2020. P. 243–248.*<https://doi.org/10.1109/ICICS49469.2020.239556>.
- Tharwat A. Analisis komponen utama - tutorial. *Pengenalan Pola Aplikasi Int J* 2016;3 (3):197.<https://doi.org/10.1504/japar.2016.079733>.
- Kumar S. Strategi Lindung Nilai yang Efektif Untuk Portofolio Obligasi Negara AS Menggunakan Analisis Komponen Utama. [On line]. Tersedia:<https://ssrn.com/abstract=4007786>.
- Biswas N, Uddin KMM, Rikta ST, Dey SK. Analisis komparatif pengklasifikasi pembelajaran mesin untuk prediksi stroke: Pendekatan analisis prediktif. *Anal Perawatan Kesehatan* November 2022;2:100116.<https://doi.org/10.1016/j.health.2022.100116>.
- Mir Ishrak A, Dhruba M, Haider N, dkk.Penerapan Pembelajaran Mesin dalam Penilaian Risiko Kredit: Pendahuluan Perbankan Cerdas.2018.
- Saleh H, dkk. Prediksi stroke menggunakan pembelajaran mesin terdistribusi berdasarkan apache spark. *Int J Adv Sci Technol* 2019;28(15):89–97.<https://doi.org/10.13140/RG.2.2.13478.68162>.
- Wang D, Zhang Y, Zhao Y. LightGBM: metode klasifikasi miRNA yang efektif pada pasien kanker payudara. *Seri Prosiding Konferensi Internasional ACM; Oktober 2017.* hal. 7-11. <https://doi.org/10.1145/3155077.3155079>.

41. Lundberg SM, dkk. Prediksi pembelajaran mesin yang dapat dijelaskan untuk pencegahan hipoksemia selama operasi. *Nat Biomed Eng* 2018;2(10):749–760.<https://doi.org/10.1038/s41551-018-0304-0>.
42. Li R, dkk. Interpretasi dan visualisasi berbasis pembelajaran mesin dari interaksi nonlinier dalam kelangsungan hidup kanker prostat. *Informatika Kanker JCO Clin* 2020;4:637–646. <https://doi.org/10.1200/cci.20.00002>.
43. Du Y, Rafferty AR, McAuliffe FM, Wei L, Mooney C. Sistem pendukung keputusan klinis berbasis pembelajaran mesin yang dapat dijelaskan untuk prediksi diabetes mellitus gestasional. *Perwakilan Sains Desember* 2022;12(1).<https://doi.org/10.1038/s41598-022-05112-2>.
44. Sokolova M, Lapalme G. Analisis sistematis ukuran kinerja untuk tugas klasifikasi. *Manajemen Proses Inf Juli* 2009;45(4):427–437.<https://doi.org/10.1016/j.ipm.2009.03.002>.
45. Luque A, Carrasco A, Martín A, DE LAS Heras A. Dampak ketidakseimbangan kelas dalam metrik kinerja klasifikasi berdasarkan matriks kebingungan biner. *Pengenalan Pola Juli* 2019;91:216–231.<https://doi.org/10.1016/j.patcog.2019.02.023>.
46. Fisher A, Rudin C, Dominici F. Semua Model Salah, Tapi Banyak yang Berguna: Mempelajari Pentingnya Variabel dengan Mempelajari Seluruh Kelas Model Prediksi Secara Bersamaan. 2019.