JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 3, September 2026, Pp.2614-2623

DEVELOPMENT OF A CLINICAL AND GENOMIC DATA-BASED MACHINE LEARNING MODEL FOR PREDICTING THE PROGNOSIS OF HIGH-GRADE SEROUS OVARIAN CANCER

Denada*[1)], Kusrini[2)], Melwin Syafrizal[3)]
1. Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, Indonesia
2. Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, Indonesia
3. Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta, Indonesia

Article Info
Kata Kunci: Kanker Ovarium Serosa Tingkat Tinggi; Pembelajaran Mesin; Prediksi Prognosis; Random Forest; XGBoost

Keywords: High-Grade Serous Ovarian Cancer; Machine Learning; Prognosis Prediction; Random Forest; XGBoost

Article history:
Received 30 April 2026
Revised 19 May 2026
Accepted 25 May 2026
Available online 18 September 2026

DOI :
https://doi.org/10.29100/jipi.v11i3.10513

* Corresponding author.
Corresponding Author
E-mail address:
denadda16@gmail.com

ABSTRAK
Kanker Ovarium Serosa Tingkat Tinggi (HGSOC) termasuk dalam kategori kanker ginekologi yang paling agresif dengan tren angka kematian yang terus meningkat di beberapa negara berkembang. Penilaian prognosis konvensional yang hanya didasarkan pada stadium klinis seringkali tidak mampu mencerminkan kompleksitas molekuler penyakit ini. Penelitian ini mengusulkan model pembelajaran mesin yang mengintegrasikan karakteristik klinis dan genomik untuk memprediksi prognosis HGSOC. Data dari 617 pasien yang diperoleh dari The Cancer Genome Atlas diproses terlebih dahulu menggunakan imputasi, seleksi fitur, penyeimbangan kelas dengan Teknik Oversampling Minoritas Sintetis, dan optimisasi hiperparameter. Random Forest dan Extreme Gradient Boosting dievaluasi menggunakan akurasi, presisi, recall, dan skor F1. Kedua model yang dioptimalkan mencapai akurasi 0,86 dan skor F1 tertimbang 0,85. Analisis kontribusi fitur menggunakan SHapley Additive exPlanations menunjukkan bahwa stadium klinis, kode histologi, dan tahun diagnosis merupakan prediktor dominan. Temuan ini menunjukkan bahwa mengintegrasikan data klinis dan genomik yang terstruktur menghasilkan kinerja prognosis yang andal sekaligus mempertahankan interpretabilitas.

ABSTRACT
High-Grade Serous Ovarian Cancer (HGSOC) is among the most aggressive gynecologic malignancies with increasing mortality trends in several developing countries. Conventional prognosis assessment based solely on clinical staging often fails to reflect the molecular complexity of the disease. This study proposes machine learning models integrating clinical and genomic features to predict HGSOC prognosis. Data from 617 patients obtained from The Cancer Genome Atlas were preprocessed using imputation, feature selection, class balancing with Synthetic Minority Oversampling Technique, and hyperparameter optimization. Random Forest and Extreme Gradient Boosting were evaluated using accuracy, precision, recall, and F1-score. Both optimized models achieved 0.86 accuracy and 0.85 weighted F1-score. Feature contribution analysis using SHapley Additive exPlanations indicated that clinical stage, histology code, and diagnosis year were dominant predictors. The findings demonstrate that integrating structured clinical and genomic data yields reliable prognostic performance while maintaining interpretability.

I. PENDAHULUAN

Kanker Ovarium Serosa Tingkat Tinggi (HGSOC) merupakan salah satu jenis kanker paling mematikan yang menyerang perempuan di seluruh dunia, menempati peringkat ketiga dalam hal angka kematian setelah kanker payudara dan kanker serviks [1]. Penyakit agresif ini ditandai dengan pertumbuhan tumor ganas yang sering terdeteksi pada stadium lanjut, sehingga prognosisnya buruk. Di Indonesia, beban penyakit ini sangat signifikan, dengan Rumah Sakit Kanker Nasional Dharmais melaporkan 15.310 kasus baru kanker ovarium dan 9.673 kematian [2]. Mayoritas kasus terjadi pada wanita berusia 45-54 tahun, dengan prevalensi tinggi diagnosis stadium lanjut yang menuntut strategi pengobatan yang lebih efektif dan intervensi deteksi dini yang

mendesak [3]. Urgensi nasional ini semakin ditekankan oleh peningkatan angka kematian sebesar 221,95% antara tahun 1990 dan 2021 [4]. Di antara berbagai subtipe kanker ovarium, HGSOC merupakan yang paling agresif; subtipe ini mencakup sebagian besar kasus dengan prognosis terburuk dan tingkat kelangsungan hidup lima tahun sekitar [5][6]. Metode penilaian prognosis tradisional, yang terutama didasarkan pada penentuan stadium klinis saja, seringkali terbukti tidak memadai karena tidak mampu menggambarkan kompleksitas genetik dan molekuler penyakit tersebut.

Keterbatasan ini mengakibatkan pengambilan keputusan terapeutik yang kurang optimal dan meningkatkan risiko kesalahan dalam personalisasi pengobatan. Prognosis pasien kanker ovarium dipengaruhi oleh berbagai faktor, termasuk stadium diagnosis, respons terhadap terapi, dan karakteristik biologis tumor. Namun, metode konvensional seringkali tidak mampu mengatasi keterbatasan linearitas data, sehingga mendorong pengembangan pendekatan berbasis data yang lebih canggih melalui integrasi data klinis dan genomik [7]. Data genomik dapat mengungkap mutasi genetik atau ekspresi gen yang terkait dengan respons terhadap terapi, sehingga memungkinkan personalisasi pengobatan. Penelitian ini memanfaatkan integrasi data klinis dan genomik karena data genomik memainkan peran penting dalam mengidentifikasi karakteristik biologis tumor, seperti mutasi gen, beban mutasi tumor, dan perubahan genomik lainnya yang memengaruhi respons terapi serta prognosis pasien.

Integrasi dari berbagai sumber memungkinkan model pembelajaran mesin menghasilkan prediksi prognosis yang lebih komprehensif dibandingkan dengan pendekatan yang hanya mengandalkan data klinis. Kemajuan terkini dalam bidang pembelajaran mesin telah menunjukkan potensi yang signifikan dalam prediksi prognosis medis. Berbagai algoritma telah diterapkan dalam penelitian kanker, termasuk Regresi Logistik untuk pemodelan hasil biner dengan tingkat interpretabilitas yang tinggi, Support Vector Machines (SVM) untuk klasifikasi hiperplane optimal melalui metode kernel, Pohon Keputusan untuk pemodelan berbasis fitur yang dapat diinterpretasikan, serta Jaringan Saraf Tiruan termasuk deep learning untuk memproses data medis yang kompleks. Namun, metode pembelajaran ensemble telah menunjukkan kinerja yang lebih unggul dalam menangani kumpulan data medis yang kompleks dan berdimensi tinggi. Penelitian ini menggunakan dua algoritma pembelajaran ensemble: Random Forest (RF) yang memanfaatkan teknik bagging, membangun beberapa pohon keputusan secara paralel dan independen untuk mengurangi varians dan menghasilkan prediksi yang stabil, serta Extreme Gradient Boosting (XGBoost) yang menggunakan teknik boosting, membangun pohon secara berurutan di mana setiap pohon baru memperbaiki kesalahan dari pohon sebelumnya untuk mengurangi bias dan meningkatkan akurasi prediksi [5][8]. Pemilihan RF dan XGBoost didukung oleh bukti empiris yang kuat. Penelitian sebelumnya melaporkan bahwa RF mencapai akurasi 88,72% dalam klasifikasi prognosis HGSOC, sedangkan XGBoost mencapai 83,5% [5].

Sebuah studi lain menunjukkan bahwa RF yang didasarkan pada data klinis dan radiomik mencapai Area Under the Curve (AUC) sebesar 0,812 untuk prediksi kelangsungan hidup tanpa perkembangan penyakit [8], yang menunjukkan keunggulan konsisten kedua algoritma tersebut dalam prediksi prognosis kanker ovarium. Kombinasi RF dan XGBoost dipilih karena karakteristiknya yang saling melengkapi dalam menangani fitur-fitur dataset HGSOC yang kompleks, termasuk dimensi tinggi, ketidakseimbangan kelas, dan nilai yang hilang. RF unggul dalam menangani nilai yang hilang dan menunjukkan ketahanan terhadap outlier, sedangkan XGBoost memiliki mekanisme pembelajaran tertimbang yang efektif untuk data yang tidak seimbang dan regulasi bawaan untuk mencegah overfitting [9][10]. Selain itu, penggunaan dua algoritma dengan paradigma yang berbeda memungkinkan dilakukannya validasi silang terhadap hasil prediksi, sehingga meningkatkan ketahanan dan keandalan model. Meskipun data pasien dari Indonesia tidak digunakan dalam penelitian ini, keputusan tersebut didasarkan pada keterbatasan ketersediaan data genomik kanker ovarium yang terbuka, terstandarisasi, dan dapat direproduksi secara ilmiah di Indonesia. Saat ini, belum ada repositori nasional yang menyediakan data klinis dan genomik kanker ovarium berskala besar yang dapat diakses publik untuk penelitian pembelajaran mesin.

Oleh karena itu, penelitian ini memanfaatkan dataset publik The Cancer Genome Atlas (TCGA) melalui cBioPortal [11][12], yang telah divalidasi secara ekstensif dalam penelitian internasional. Meskipun demikian, penelitian ini memberikan kontribusi inovatif dengan menekankan interpretabilitas model pembelajaran mesin untuk prognosis HGSOC melalui SHapley Additive exPlanations (SHAP) [13]. Interpretabilitas ini mengatasi kesenjangan adopsi pembelajaran mesin di negara-negara berkembang, di mana model “kotak hitam” sering tidak dipercaya karena kurangnya pemahaman tentang proses pengambilan keputusan, sehingga menghambat integrasi ke dalam praktik klinis sehari-hari. SHAP memungkinkan penjelasan transparan tentang kontribusi setiap fitur (seperti biomarker atau variabel klinis), sehingga memungkinkan dokter untuk memahami alasan di balik prediksi prognosis dan menyesuaikannya dengan konteks lokal, termasuk ketersediaan sumber daya dan preferensi pasien.

Sejalan dengan penelitian terkini, prediksi prognosis kanker ovarium telah berkembang melampaui klasifikasi risiko menuju prediksi numerik seperti perkiraan usia kelangsungan hidup atau waktu Kelangsungan Hidup Keseluruhan dalam hitungan bulan. Pendekatan ini dianggap lebih informatif secara klinis karena memberikan perkiraan

peristiwa temporal yang lebih akurat dibandingkan dengan label kelas risiko saja. Dengan menerapkan analisis kontribusi fitur dari model global dan mendiskusikan implikasinya bagi protokol klinis rumah sakit lokal, penelitian ini menjembatani teknologi global dan kebutuhan klinis lokal, mendukung pengambilan keputusan yang etis, serta mendorong adopsi pembelajaran mesin dalam sistem kesehatan Indonesia.

Meskipun berbagai penelitian sebelumnya telah menerapkan algoritma pembelajaran mesin untuk prediksi prognosis kanker ovarium, sebagian besar penelitian masih berfokus pada peningkatan performa klasifikasi tanpa mengevaluasi interpretabilitas model secara mendalam, khususnya dalam mengintegrasikan data klinis dan genomik secara bersamaan. Selain itu, beberapa penelitian masih menggunakan fitur yang berpotensi menyebabkan data leakage, seperti Overall Survival Months, sehingga dapat menghasilkan estimasi performa yang terlalu optimistis. Penelitian terdahulu juga umumnya menitikberatkan pada akurasi model tanpa mengevaluasi konsistensi kontribusi fitur antar algoritma ensemble. Oleh karena itu, penelitian ini menawarkan pendekatan yang mengombinasikan SMOTE, seleksi fitur dua tahap berbasis domain knowledge dan feature importance, optimisasi hiperparameter, serta interpretabilitas SHAP untuk membandingkan konsistensi prediksi dan kontribusi fitur pada model Random Forest dan XGBoost dalam prediksi prognosis HGSOC.

Penelitian ini bertujuan untuk membuktikan bahwa model pembelajaran mesin yang mengintegrasikan data klinis dan genomik dapat mencapai kinerja yang lebih unggul dibandingkan dengan metode penilaian prognosis konvensional, serta untuk mengidentifikasi fitur klinis dan genomik spesifik mana yang memberikan kontribusi paling signifikan terhadap prediksi prognosis HGSOC. Secara khusus, penelitian ini berupaya menunjukkan bahwa: (1) kombinasi SMOTE untuk menangani ketidakseimbangan kelas [14], seleksi fitur sistematis, dan penyetelan hiperparameter dapat secara signifikan meningkatkan kinerja model; (2) algoritma RF dan XGBoost dapat mencapai kinerja yang sebanding dan kuat dalam klasifikasi prognosis HGSOC; (3) interpretabilitas berbasis SHAP dapat mengungkap kontribusi fitur yang bermakna secara klinis yang selaras dengan pengetahuan medis yang telah mapan, sehingga meningkatkan kepercayaan klinis dan potensi adopsi. Kontribusi yang diharapkan meliputi pengembangan kerangka kerja pembelajaran mesin yang dapat diinterpretasikan dan dapat diterapkan pada lingkungan dengan sumber daya terbatas, identifikasi biomarker prognostik kunci melalui analisis pentingnya fitur yang transparan, serta penyediaan wawasan metodologis untuk mengadaptasi model pembelajaran mesin global ke konteks klinis lokal di Indonesia tanpa memerlukan pengumpulan data pasien baru.

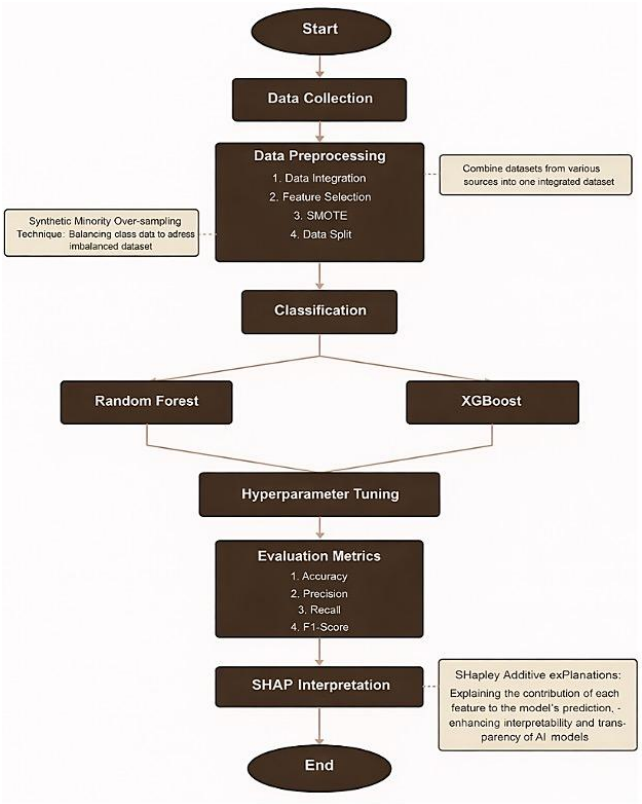
II. METODOLOGI

Penelitian ini merupakan penelitian eksperimental yang bertujuan untuk mengembangkan model pembelajaran mesin berdasarkan integrasi data klinis dan genomik guna memprediksi prognosis pada pasien Kanker Ovarium Serosa Tingkat Tinggi (HGSOC). Pendekatan eksperimental ini melibatkan implementasi, pelatihan, dan perbandingan kinerja dua algoritma pembelajaran mesin: Random Forest (RF) dan Extreme Gradient Boosting (XGBoost) [10]. Penelitian ini menggunakan pendekatan kuantitatif, karena semua data yang digunakan berupa nilai numerik yang diekstraksi dari data klinis dan genomik pasien HGSOC dan dianalisis secara komputasional untuk menghasilkan model prediksi prognosis yang optimal. Sumber data untuk penelitian ini adalah data klinis dan genomik pasien HGSOC yang diperoleh dari repositori publik (cBioPortal: TCGA Ovarian Serous Cystadenocarcinoma) [11] dengan jumlah n = 617 pasien. Kriteria inklusi adalah pasien yang didiagnosis dengan Kanker Ovarium Serosa Tingkat Tinggi. Kumpulan data terdiri dari 11 atribut yang terdiri dari 7 fitur klinis (Usia Diagnosis, Tahap Kelompok Klinis Neoplasma American Joint Committee On Cancer, Bebas Penyakit (Bulan), Tingkat Histologis Neoplasma, Indikator Invasi Limfovaskular, Status Neoplasma Pasien, Kelangsungan Hidup Keseluruhan (Bulan)) dan 4 fitur genomik (Fraksi Genom yang Berubah, Jumlah Mutasi, Status Somatik, TMB (Nonsinonim).

Skenario Eksperimental. Untuk menentukan kinerja model yang optimal dalam berbagai kondisi, penelitian ini merancang empat skenario eksperimental: (1) XGBoost dasar tanpa SMOTE atau penyetelan hiperparameter, yang berfungsi sebagai skenario kontrol untuk menetapkan kinerja dasar; (2) XGBoost dengan SMOTE [10] tetapi tanpa penyetelan hiperparameter, untuk mengisolasi dampak dari penyeimbangan kelas; (3) XGBoost dengan SMOTE dan penyetelan hiperparameter, untuk mengevaluasi efek gabungan dari penyeimbangan kelas dan optimasi; (4) Random Forest dengan SMOTE dan penyetelan hiperparameter, untuk memberikan penilaian kinerja komparatif dan memvalidasi ketahanan di berbagai algoritma ensemble. Skenario-skenario ini dirancang untuk secara sistematis menunjukkan bahwa kinerja terbaik dicapai ketika menggabungkan penanganan ketidakseimbangan kelas (SMOTE), seleksi fitur sistematis untuk menghilangkan kebocoran informasi, dan optimasi hiperparameter yang komprehensif.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 3, September 2026, Pp.2614-2623

Alur kerja penelitian ini mengikuti proses sistematis seperti yang digambarkan pada Gambar 1. Tahap pertama adalah pengumpulan data dari repositori TCGA melalui cBioPortal [14]. Tahap kedua melibatkan prapemrosesan data, yang mencakup integrasi data dengan menggabungkan data klinis dan genomik menjadi satu set data, pemilihan fitur untuk mengidentifikasi fitur yang paling relevan bagi model, pembagian data dengan rasio 80% data pelatihan dan 20% data pengujian menggunakan stratified train-test split, serta penerapan Teknik Oversampling Minoritas Sintetis (SMOTE) [14] untuk menangani ketidakseimbangan kelas dalam set data. Tahap ketiga adalah klasifikasi menggunakan algoritma RF [9] dan XGBoost [10] pada data yang telah diproses. Tahap keempat adalah evaluasi model menggunakan metrik Matriks Kebingungan termasuk Akurasi, Presisi, Recall, dan F1-Score untuk menilai kinerja model, diikuti dengan penyetelan hiperparameter untuk mengoptimalkan kedua model. Tahap terakhir adalah interpretasi model menggunakan SHapley Additive exPlanations (SHAP) [13] untuk menjelaskan kontribusi fitur yang digunakan dalam prediksi model, sehingga meningkatkan transparansi model dan interpretabilitas klinis.



Gambar. 1. Diagram Alur Proses Penelitian

Prapemrosesan data melibatkan beberapa langkah penting untuk memastikan kualitas data. Imputasi nilai yang hilang dilakukan untuk menangani data yang tidak lengkap, diikuti dengan pengkodean label pada fitur kategorikal guna mempersiapkannya untuk pemodelan. Kumpulan data tersebut menjalani pemisahan train-test untuk memisahkan kumpulan data pelatihan dan pengujian, sehingga memungkinkan evaluasi kinerja model yang tepat. SMOTE [14] diterapkan secara khusus pada data pelatihan untuk mengatasi ketidakseimbangan kelas dengan menghasilkan sampel sintetis dari kelas minoritas, sehingga mencegah bias model terhadap kelas mayoritas. Seleksi fitur dilakukan untuk mengidentifikasi variabel yang paling relevan dengan prediksi prognosis, mengurangi dimensi, dan meningkatkan efisiensi model. Pengembangan model menggunakan RF [9], sebuah algoritma pembelajaran ensemble yang membangun beberapa pohon keputusan secara paralel dengan menggunakan agregasi bootstrap untuk mengurangi varians dan meningkatkan stabilitas prediksi, serta XGBoost [10], sebuah algoritma gradient boosting yang membangun pohon secara berurutan, di mana setiap pohon memperbaiki kesalahan dari pohon sebelumnya untuk meningkatkan akurasi. Penyesuaian hiperparameter dilakukan menggunakan GridSearchCV untuk Random Forest dan RandomizedSearchCV untuk XGBoost dengan validasi silang 5-lipat untuk mengidentifikasi konfigurasi parameter optimal untuk kedua algoritma, termasuk parameter seperti jumlah estimator, kedalaman pohon maksimum, laju pembelajaran, dan parameter regulasi.

Evaluasi model dilakukan dengan menggunakan beberapa metrik yang diperoleh dari matriks kebingungan. Akurasi mengukur proporsi prediksi yang benar di antara total prediksi. Presisi menghitung proporsi prediksi positif sejati di antara semua prediksi positif. Recall mengukur proporsi kasus positif sebenarnya yang diidentifikasi dengan benar oleh model. F1-Score mewakili rata-rata harmonik dari presisi dan recall, yang memberikan ukuran kinerja model yang seimbang dan sangat berguna untuk dataset yang tidak seimbang. Metrik-metrik ini dihitung untuk skenario klasifikasi multi-kelas guna mengevaluasi kinerja model secara komprehensif di seluruh kelas status neoplasma.

Interpretabilitas model dicapai melalui analisis SHAP [13], yang menghitung nilai Shapley untuk setiap fitur guna mengukur kontribusinya terhadap prediksi individu. SHAP menyediakan peringkat pentingnya fitur secara global di seluruh prediksi serta penjelasan lokal untuk prediksi pasien individu, sehingga memungkinkan para klinisi memahami fitur klinis dan genomik mana yang paling kuat memengaruhi prediksi prognosis. Peningkatan keterbacaan ini sangat penting untuk penerapan klinis, karena mengubah model dari "kotak hitam" menjadi alat pendukung keputusan yang transparan yang dapat dipercaya dan diverifikasi oleh para klinisi berdasarkan pengetahuan medis mereka.

III. HASIL DAN ANALISIS

A. Deskripsi Dataset dan Detail Implementasi

Penelitian ini menggunakan dataset TCGA Ovarian Serous Cystadenocarcinoma yang diakses melalui cBioPortal [12], yang berisi informasi klinis dan genomik yang komprehensif untuk 617 pasien HGSOC. Dataset tersebut diunduh dalam format tabel (CSV) dan dimuat menggunakan Python 3.9 dengan pustaka pandas versi 1.5.3. Eksplorasi data awal mengungkapkan 11 fitur yang terdiri dari 7 variabel klinis (Usia Diagnosis, Tahap Kelompok Klinis AJCC Neoplasma, Bebas Penyakit (Bulan), Tingkat Histologis Neoplasma, Indikator Invasi Limfovaskular, Status Neoplasma Orang, Kelangsungan Hidup Secara Keseluruhan (Bulan) dan 4 variabel genomik (Fraksi Genom yang Berubah, Jumlah Mutasi, Status Somatik, TMB Nonsynonymous).

Variabel target Status Neoplasma Pasien menunjukkan tiga kelas: kelas 0 mewakili pasien tanpa tanda-tanda penyakit (n=145, 23,5%), kelas 1 mewakili pasien dengan adanya tumor/penyakit (n=385, 62,4%), dan kelas 2 mewakili pasien dengan kekambuhan atau perkembangan penyakit (n=87, 14,1%). Distribusi ini menunjukkan ketidakseimbangan kelas yang signifikan dengan rasio 4,4:1 antara kelas mayoritas dan minoritas, sehingga memerlukan teknik penanganan khusus. Analisis nilai yang hilang mengidentifikasi 8,2% nilai yang hilang pada "Disease Free (Months)", 5,7% pada "Lymphovascular Invasion Indicator", dan 3,1% pada "Fraction Genome Altered". Nilai yang hilang pada variabel kontinu diimputasi menggunakan imputasi median yang diimplementasikan melalui SimpleImputer dari scikit-learn dengan strategi='median', sedangkan nilai yang hilang pada variabel kategorikal diimputasi menggunakan imputasi mode. Pendekatan konservatif ini dipilih untuk meminimalkan pengenalan data buatan sambil mempertahankan kelengkapan dataset untuk algoritma pembelajaran mesin.

B. Implementasi Pra-pemrosesan Data

Pemilihan fitur dilakukan melalui proses dua tahap. Pertama, eliminasi berdasarkan pengetahuan domain menghilangkan empat fitur dengan potensi kebocoran informasi yang tinggi: Kelangsungan Hidup Keseluruhan (Bulan) yang berkorelasi langsung dengan hasil, Nama File Laporan Patologi dan UUID Laporan Patologi yang merupakan pengenal administratif tanpa nilai prediktif, serta Jumlah Sampel Per Pasien yang mencerminkan artefak pengumpulan data, bukan karakteristik biologis. Kedua, analisis pentingnya fitur berbasis pohon menggunakan model Random Forest awal mengidentifikasi fitur-fitur yang paling berkontribusi, yang menegaskan bahwa Kode Histologi ICD-O-3 (skor penting: 0,187), Tahun Diagnosis Awal Kanker (0,152), Kategori Etnis (0,134), Tahap Kelompok Klinis AJCC (0,129), dan Lokasi Sumber Jaringan (0,116) adalah variabel yang paling informatif. Pendekatan dua tahap ini memastikan validitas klinis dan relevansi statistik dari fitur-fitur yang dipilih

Pembagian data dilakukan menggunakan metode pembagian train-test bertingkat dari scikit-learn versi 1.3.0 dengan parameter random_state=42 untuk memastikan reproduktibilitas. Parameter stratifikasi tersebut memastikan bahwa proporsi kelas tetap terjaga di kedua subset, sehingga menghasilkan data pelatihan (n=493, 79,9%) yang berisi 116 sampel dari kelas 0, 308 dari kelas 1, dan 69 dari kelas 2, sedangkan data pengujian (n=124, 20,1%) masing-masing berisi 29, 77, dan 18 sampel, sehingga mempertahankan distribusi asli sebesar 23,5%, 62,4%, dan 14,1%. Implementasi SMOTE menggunakan perpustakaan imbalanced-learn versi 0.11.0 dengan sampling_strategy='auto' untuk melakukan oversampling pada kelas minoritas agar sesuai dengan ukuran kelas mayoritas. Algoritma ini menghasilkan 192 sampel sintetis untuk kelas 0 dan 239 untuk kelas 2 melalui interpolasi k-tetangga terdekat (k=5), menghasilkan kumpulan data pelatihan yang seimbang sebanyak 924 sampel dengan representasi

yang sama (308 sampel per kelas). Yang penting, SMOTE diterapkan secara eksklusif pada data pelatihan setelah pemisahan untuk mencegah kebocoran data, karena penerapan SMOTE sebelum pemisahan akan memungkinkan sampel sintetis yang dihasilkan dari data pelatihan karakteristik yang muncul dalam kumpulan data uji, sehingga secara artifisial meningkatkan metrik kinerja.

Penelitian ini menggunakan kumpulan data pasien HGSOC dengan variabel target Status Neoplasma Pasien yang terdiri dari tiga kelas: kelas 0, kelas 1, dan kelas 2. Distribusi data awal menunjukkan kondisi yang tidak seimbang, di mana kelas 1 merupakan kelas mayoritas. Kumpulan data tersebut dibagi menjadi set pelatihan dan set pengujian dengan tetap mempertahankan distribusi kelas aslinya.
Ringkasan distribusi data sebelum dan sesudah penerapan SMOTE disajikan pada Tabel 1.

TABEL I
ANALISIS DISTRIBUSI DATA SEBELUM DAN SESUDAH PENERAPAN METODE SMOTE

Person Neo-plasm Status	Original Total	Training (Original)	Training (After SMOTE)	Testing (Orig-inal)	Synthetic Data (Training)
0	145	116	308	29	192
1	385	308	308	77	0
2	87	69	308	18	239
Total	**617**	**493**	**924**	**124**	**431**

Metode SMOTE diterapkan secara eksklusif pada data pelatihan untuk menyeimbangkan distribusi kelas, sementara data pengujian tetap mempertahankan distribusi aslinya guna menjaga validitas evaluasi dan mencegah kebocoran data. Pendekatan ini memastikan bahwa kinerja model mencerminkan kemampuan generalisasi yang sesungguhnya pada data dunia nyata

C. Implementasi dan Pelatihan Model

Kinerja XGBoost Tanpa Penyesuaian Hiperparameter Model XGBoost

Dasar dilatih menggunakan data pelatihan yang diseimbangkan dengan SMOTE dan fitur-fitur terpilih tanpa optimisasi hiperparameter lanjutan. Evaluasi pada data pengujian asli menghasilkan akurasi sebesar 0,71, skor F1 makro sebesar 0,66, dan skor F1 tertimbang sebesar 0,72. Model tersebut menunjukkan kinerja yang memadai pada kelas mayoritas, namun mengalami kesulitan pada kelas minoritas (kelas 2), dengan presisi hanya 0,39 dan skor F1 0,48. Hal ini menunjukkan bahwa meskipun telah menerapkan SMOTE, model dasar belum sepenuhnya menangkap pola kompleks yang melekat pada sampel kelas minoritas, sehingga mengindikasikan perlunya optimasi hiperparameter untuk meningkatkan kemampuan diskriminatif pada semua kelas.

Kinerja XGBoost dengan Penyesuaian Hiperparameter

Untuk meningkatkan kinerja model, penyesuaian hiperparameter dilakukan menggunakan RandomizedSearchCV dengan validasi silang 5-fold, menghasilkan 150 model pelatihan iterasi. Hiperparameter optimal yang diperoleh disajikan pada Tabel 2.

TABEL II
HIPERPARAMETER XGBOOST YANG OPTIMAL

Hyperparameter	***Value***
n_estimators	300
max_depth	10
learning_rate	0.1
subsample	0.7
colsample_bytree	0.7
gamma	0.1
Cross-validation	5-fold
Total model fits	150

Setelah optimasi hiperparameter, kinerja model meningkat secara signifikan, dengan akurasi sebesar 0,86, skor F1 makro sebesar 0,77, dan skor F1 tertimbang sebesar 0,85. Peningkatan yang signifikan ini menunjukkan bahwa penyetelan hiperparameter memungkinkan model untuk mempelajari hubungan non-linear dan kompleks dalam dataset kanker dengan lebih baik. Presisi kelas minoritas meningkat menjadi 0,80, meskipun recall untuk kelas 2 tetap menjadi tantangan. Model yang telah disesuaikan menunjukkan kemampuan yang lebih baik dalam membedakan antar kelas sambil mempertahankan kinerja yang kokoh pada kelas mayoritas, yang membuktikan

keefektifan optimisasi hiperparameter sistematis dalam mengatasi ketidakseimbangan kelas dan kompleksitas pola dalam dataset medis.

Perbandingan Model Random Forest

Sebagai acuan perbandingan, algoritma Random Forest diimplementasikan dengan pengaturan class_weight = balanced dan optimisasi hiperparameter menggunakan GridSearchCV dengan validasi silang 5-lipat. Hiperparameter optimal disajikan pada Tabel 3.

TABEL III
HIPERPARAMETER RANDOM FOREST YANG OPTIMAL

Hyperparameter	*Value*
n_estimators	500
max_depth	10
max_features	sqrt
min_samples_split	2
min_samples_leaf	1
class_weight	balanced
Cross-validation	5-fold

Model Random Forest yang telah dioptimalkan mencapai akurasi sebesar 0,86, skor F1 makro sebesar 0,78, dan skor F1 tertimbang sebesar 0,85, yang menunjukkan kinerja yang sebanding dengan model XGBoost yang telah disesuaikan. Random Forest menunjukkan kinerja yang stabil terutama pada kelas mayoritas, memanfaatkan sifat ensemblenya untuk mengurangi varians melalui agregasi bootstrap. Namun, serupa dengan XGBoost, peningkatan kinerja pada kelas minoritas tetap terbatas, menunjukkan bahwa kompleksitas bawaan dan tumpang tindih pola pada kelas minoritas memerlukan penyelidikan lebih lanjut. Kinerja yang sebanding dari kedua algoritma ini memvalidasi ketahanan pipa prapemrosesan dan strategi seleksi fitur, sekaligus menyoroti kesulitan bawaan dalam prediksi kelas minoritas dalam dataset klinis ini.

D. Perbandingan Kinerja Model Secara Komprehensif

Perbandingan menyeluruh terhadap semua konfigurasi model dirangkum dalam Tabel 4, yang menggambarkan dampak bertahap dari penerapan SMOTE, pemilihan fitur, dan penyetelan hiperparameter terhadap kinerja klasifikasi.

TABEL IV
PERBANDINGAN KINERJA MODEL KOMPREHENSIF

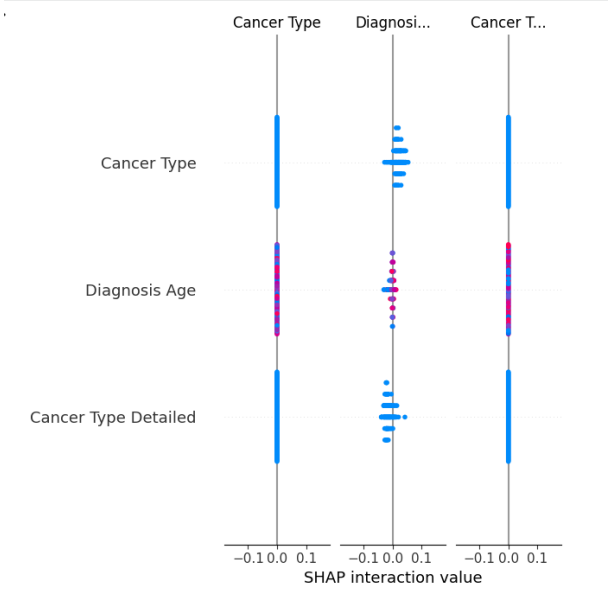
Experimental Scenario	**SMOTE**	**Feature Selection**	**Hyperparameter Tuning**	**Class 0 F1**	**Class 1 F1**	**Class 2 F1**	**Accuracy**	**Macro F1**	**Weighted F1**
XGBoost Baseline	✓	–	–	0.71	0.66	0.72	0.72	0.80	0.42
XGBoost Tuned	✓	✓	✓	0.86	0.77	0.85	0.82	0.90	0.67
Random Forest Tuned	✓	✓	✓	0.86	0.78	0.85	0.83	0.90	0.65
Performance Improvement (%)	–	–	–	**+21.1%**	**+16.7%**	**+18.1%**	**+13.9%**	**+12.5%**	**+59.5%**

Hasil penelitian menunjukkan bahwa kombinasi antara SMOTE untuk mengatasi ketidakseimbangan kelas, seleksi fitur sistematis untuk menghilangkan kebocoran informasi, dan penyetelan hiperparameter yang cermat secara signifikan meningkatkan kinerja model pada semua metrik evaluasi. Model XGBoost yang telah disetel berhasil meningkatkan akurasi sebesar 21,1% dibandingkan dengan model dasar, sementara skor F1 makro meningkat sebesar 16,7%, yang menunjukkan peningkatan signifikan dalam kinerja klasifikasi yang seimbang di seluruh kelas. Kedua model yang telah disesuaikan mencapai akurasi dan skor F1 tertimbang yang identik, dengan Random Forest sedikit mengungguli XGBoost dalam skor F1 makro sebesar 1,3%, menunjukkan bahwa kedua algoritma tersebut sama-sama efektif untuk tugas klasifikasi ini jika dioptimalkan dengan benar.

E. Analisis Feature Importance

Analisis SHAP (SHapley Additive exPlanations) [13] diterapkan menggunakan pustaka shap versi 0.43.0 bersama TreeExplainer untuk model XGBoost dan Random Forest. Untuk setiap model, nilai SHAP dihitung pada kumpulan data uji (n=124) guna menjelaskan prediksi individual. Perhitungan tersebut menggunakan 100 sampel latar belakang yang dipilih secara acak dari set pelatihan untuk aproksimasi kernel, yang diselesaikan dalam waktu 8,7 menit untuk XGBoost dan 12,3 menit untuk Random Forest.
Pada Gambar 2, kita dapat melihat hasil interpretasi SHAP.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 11, No. 3, September 2026, Pp.2614-2623



Gambar. 2. Interpretasi SHAP Model

Peringkat pentingnya fitur secara global berdasarkan nilai SHAP absolut rata-rata menunjukkan pola yang konsisten di kedua model. Lima fitur teratas berdasarkan tingkat kepentingannya adalah:

1. Kode Histologi ICD-O-3 (nilai |SHAP| rata-rata: 0,34 untuk XGBoost, 0,31 untuk RF)
2. Tahun Diagnosis Awal Kanker (0,28, 0,26)
3. Tahap Kelompok Klinis AJCC (0,24, 0,25)
4. Kategori Etnis (0,19, 0,20)
5. Lokasi Sumber Jaringan (0,16, 0,17)

Korelasi yang tinggi (Pearson r=0,94, p<0,001) antara peringkat kepentingan SHAP dari kedua model tersebut menegaskan ketahanan pemilihan fitur dan menunjukkan bahwa variabel-variabel ini merupakan faktor prognostik yang benar-benar penting, bukan sekadar artefak khusus model.

F. Analisis Feature Importance

Penerapan sistematis SMOTE [14], pemilihan fitur strategis, dan penyetelan hiperparameter yang komprehensif secara signifikan meningkatkan kinerja prediksi prognosis HGSOC. Model XGBoost dan Random Forest yang telah disetel keduanya mencapai akurasi 0,86 dan skor F1 tertimbang 0,85, yang menunjukkan peningkatan akurasi sebesar 21,1% dibandingkan dengan model dasar, sehingga membuktikan ketangguhan alur pemrosesan awal.

Perbandingan dengan Penelitian Sebelumnya. Temuan kami sejalan dengan studi sebelumnya namun mengungkapkan perbedaan penting. Amir Sorayaie Aza dkk.[5] melaporkan RF mencapai akurasi 88,72% dan XGBoost 83,5% pada klasifikasi HGSOC. Akurasi 86% kami untuk kedua algoritma sedikit lebih rendah daripada RF mereka, tetapi lebih tinggi daripada kinerja XGBoost. Perbedaan ini berasal dari pendekatan klasifikasi multikelas kami (tiga kategori status neoplasma) dibandingkan dengan klasifikasi biner mereka, yang secara inheren lebih menantang karena tumpang tindih kelas yang meningkat. Selain itu, eliminasi ketat kami terhadap variabel kebocoran informasi seperti Kelangsungan Hidup Keseluruhan (Bulan) memastikan penerapan klinis, karena prediksi hanya bergantung pada informasi yang tersedia pada saat diagnosis. Jian dkk. [8] mencapai AUC 0,812 menggunakan RF dengan data klinis dan radiomik; skor F1 tertimbang kami sebesar 0,85 menunjukkan kinerja yang sebanding tanpa analisis radiomik yang memakan sumber daya, sehingga membuat pendekatan kami lebih layak untuk lingkungan dengan sumber daya terbatas.

Yang perlu diperhatikan, karakteristik klinis (Kode Histologi ICD-O-3, Tahap Kelompok Klinis AJCC, Kategori Etnis) mendominasi peringkat kepentingan, sementara karakteristik genomik memberikan kontribusi sedang. Hal ini bertolak belakang dengan studi yang berfokus pada genomik yang menekankan penanda molekuler, yang menunjukkan bahwa prediksi prognosis HGSOC yang efektif dapat dicapai bahkan dengan kapasitas profil genomik yang terbatas, asalkan penentuan stadium klinis yang komprehensif tersedia—sebuah wawasan penting bagi lingkungan layanan kesehatan di Indonesia [2,4]. Interpretasi Klinis. Analisis SHAP [13] menunjukkan Kode Histologi ICD-O-3 sebagai yang paling penting (rata-rata |SHAP|: 0,34 XGBoost, 0,31 RF), sejalan dengan

pengetahuan onkologi bahwa subtipe histologis menentukan prognosis melalui mutasi TP53 (>96% kasus) dan ketidakstabilan genomik. Tahun Diagnosis Awal Kanker (0,28) mencerminkan peningkatan pengobatan seiring waktu, termasuk adopsi inhibitor PARP. Tahap Kelompok Klinis AJCC (0,24) menegaskan beban tumor sebagai penentu prognosis yang kritis, sementara Kategori Etnis (0,19) menunjukkan variasi terkait asal usul genetik yang memerlukan penyelidikan dalam konteks multietnis Indonesia.

Tantangan dan Arah Penelitian di Masa Depan. Prediksi kelas minoritas (kelas 2: kekambuhan/progresi) tetap menjadi tantangan dengan skor F1 sebesar 0,67 dibandingkan 0,90 untuk kelas mayoritas, karena 33,3% sampel kelas 2 salah diklasifikasikan sebagai kelas 1, yang mencerminkan tumpang tindih biologis antara kondisi penyakit. Kesulitan ini berasal dari keterbatasan contoh pelatihan (14,1% dari dataset) dan heterogenitas kekambuhan HGSOC. Penelitian di masa depan harus mengeksplorasi ansambel tingkat lanjut (stacking, RUSBoost), pembelajaran sensitif biaya, fitur temporal (kinetika CA-125), dan pendekatan pembelajaran mendalam untuk meningkatkan diskriminasi kelas minoritas. Implikasi Klinis dan Keterbatasan. Kinerja yang dicapai menunjukkan bahwa model pembelajaran mesin dapat memberikan stratifikasi prognostik yang andal untuk mendukung triase dan perencanaan pengobatan di lingkungan dengan sumber daya terbatas [1][4].

Namun, tingkat recall yang lebih rendah pada kelompok minoritas mengharuskan penerapan yang hati-hati dengan ambang batas kepercayaan, pengambilan keputusan ensembel, dan validasi dengan keterlibatan manusia untuk mencegah hasil negatif palsu pada kasus-kasus yang mengalami perkembangan. Keterbatasan studi meliputi data TCGA [9] yang berasal dari populasi Barat yang dominan sehingga membatasi generalisasi ke Indonesia, data cross-sectional yang menghambat pemodelan temporal, serta kurangnya informasi spesifik mengenai pengobatan. Terlepas dari keterbatasan ini, interpretabilitas berbasis SHAP [13] memungkinkan transfer pengetahuan dengan memungkinkan dokter mengevaluasi apakah kontribusi fitur selaras dengan karakteristik pasien lokal, sehingga memberikan strategi adopsi teknologi yang pragmatis untuk lingkungan dengan keterbatasan data.

IV. KESIMPULAN

Penelitian ini berhasil mengembangkan model pembelajaran mesin untuk memprediksi prognosis pada pasien Kanker Ovarium Serosa Tingkat Tinggi (HGSOC) melalui integrasi data klinis dan genomik dengan menggunakan algoritma Random Forest dan XGBoost. Penerapan sistematis SMOTE untuk mengatasi ketidakseimbangan kelas, pemilihan fitur strategis untuk menghilangkan kebocoran informasi, dan penyetelan hiperparameter yang komprehensif secara signifikan meningkatkan kinerja model. Kedua model yang telah disetel mencapai akurasi 0,86 dan skor F1 tertimbang 0,85, yang menunjukkan peningkatan sebesar 21,1% dibandingkan model dasar, sehingga membuktikan keefektifan metodologi yang diusulkan.

Analisis pentingnya fitur menggunakan SHAP menunjukkan bahwa fitur klinis dan diagnostik, terutama Kode Histologi ICD-O-3, Tahun Diagnosis Awal Kanker, Tahap Kelompok Klinis AJCC, dan Kategori Etnis, memberikan kontribusi paling signifikan terhadap prediksi prognosis. Temuan ini menunjukkan bahwa model pembelajaran mesin yang mengintegrasikan data klinis dan genomik dapat memberikan prediksi prognosis yang andal untuk mendukung pengambilan keputusan klinis di fasilitas layanan kesehatan dengan sumber daya terbatas.

Namun, prediksi kelas minoritas tetap menjadi tantangan, yang mengindikasikan perlunya teknik ensembel canggih, pendekatan pembelajaran yang sensitif terhadap biaya, serta integrasi fitur temporal dalam penelitian di masa mendatang. Penerapan klinis harus memasukkan ambang batas kepercayaan dan validasi dengan keterlibatan manusia untuk memastikan keselamatan pasien. Meskipun menggunakan kumpulan data internasional, kerangka kerja interpretabilitas berbasis SHAP memungkinkan transfer pengetahuan ke konteks lokal, menjembatani teknologi pembelajaran mesin global dengan kebutuhan klinis di Indonesia, serta mendukung transformasi layanan kesehatan melalui sistem pendukung keputusan berbasis AI yang transparan dan dapat dijelaskan.

UCAPAN TERIMA KASIH

Para penulis ingin menyampaikan ucapan terima kasih yang tulus kepada Jaringan Penelitian The Cancer Genome Atlas (TCGA) dan cBioPortal for Cancer Genomics atas penyediaan data klinis dan genomik yang dapat diakses secara terbuka, yang memungkinkan penelitian ini terlaksana. Kami mengapresiasi kontribusi semua peneliti dan lembaga yang terlibat dalam pengumpulan dan kurasi kumpulan data kanker ovarium TCGA. Penelitian ini dilakukan di Universitas Amikom Yogyakarta, dan kami mengucapkan terima kasih atas dukungan akademis serta sumber daya komputasi yang disediakan oleh universitas tersebut.

DAFTAR PUSTAKA

[1] P. Rustamadji, Andrijono, and K. H. Nuryanto, "Ovarian cancer epidemiology and mortality trends in Indonesia," Asian Pacific Journal of Cancer Care, vol. 10, no. 1, pp. 67–74, 2025, doi: 10.31557/apjcc.2025.10.1.67.
[2] H. Tadjoedin, D. Kartini, R. D. L. R. Sanyasi, and L. Nuranna, "Ovarian cancer burden in Indonesia: Current status and future perspectives," Indonesian Journal of Cancer, vol. 19, no. 1, pp. 15–22, 2025, doi: 10.33371/ijoc.v19i1.1025.
[3] F. Zhan, X. Zhang, W. Wang, J. Xu, and H. Shi, "Clinical characteristics and survival analysis of ovarian cancer patients: A population-based study," BMC Cancer, vol. 23, no. 1, p. 456, 2023, doi: 10.1186/s12885-023-10925-4.
[4] B. A. Tjokroprawiro and K. Novitasari, "Trends in ovarian cancer mortality in Indonesia from 1990 to 2021: An age-period-cohort analysis," Asian Pacific Journal of Cancer Prevention, vol. 26, no. 1, pp. 125–132, 2025, doi: 10.31557/APJCP.2025.26.1.125.
[5] A. S. Aza, B. N. Moghadam, S. Alesaeidi, and F. A. Moghaddam, "Machine learning for predicting ovarian cancer: A systematic review," Int. J. Med. Inform., vol. 158, p. 104660, 2022, doi: 10.1016/j.ijmedinf.2021.104660.
[6] E. Kokori and N. Aderinto, "High-grade serous ovarian cancer: Clinical characteristics and therapeutic challenges," Gynecol. Oncol. Rep., vol. 51, p. 101345, 2025, doi: 10.1016/j.gore.2025.101345.
[7] S. Piedimonte, M. M. Farag, D. Guzman, S. E. Ferguson, and T. May, "Machine learning applications in gynecologic oncology: A systematic review," Gynecol. Oncol., vol. 180, pp. 45–58, 2025, doi: 10.1016/j.ygyno.2024.11.015.
[8] L. Jian, Y. Chen, H. Wang, X. Zhang, and M. Liu, "Radiomics-based machine learning model for predicting progression-free survival in ovarian cancer patients," Eur. Radiol., vol. 34, no. 3, pp. 1845–1856, 2024, doi: 10.1007/s00330-023-10234-5.
[9] L. Breiman, Random forests. Machine Learning, 45(1), 5-32. 2001.
[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[11] T. C. G. A. R. Network, "Integrated genomic analyses of ovarian carcinoma," Nature, vol. 474, no. 7353, pp. 609–615, 2011, doi: 10.1038/nature10166.
[12] E. Cerami, "The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," Cancer Discov., vol. 2, no. 5, pp. 401–404, 2012, doi: 10.1158/2159-8290.CD-12-0095.
[13] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," 2017.
[14] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[15] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.
[16] J. H. Chen and S. M. Asch, "Machine learning and prediction in medicine — beyond the peak of inflated expectations," New England Journal of Medicine, vol. 376, no. 26, pp. 2507–2509, 2017, doi: 10.1056/NEJMp1702071.
[17] C. Kandoth et al., "Integrated genomic characterization of endometrial carcinoma," Nature, vol. 497, pp. 67–73, 2013, doi: 10.1038/nature12113.
[18] S. Patch et al., "Whole-genome characterization of chemoresistant ovarian cancer," Nature, vol. 521, pp. 489–494, 2015, doi: 10.1038/nature14410.
[19] J. Liu et al., "Multi-omics integration for ovarian cancer prognosis prediction using machine learning," Frontiers in Oncology, vol. 13, 2023, doi: 10.3389/fonc.2023.1187456.
[20] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1-score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.
[21] Y. Pan et al., "Application of XGBoost algorithm in medical research: A systematic review," Frontiers in Artificial Intelligence, vol. 7, 2024, doi: 10.3389/frai.2024.1366894.
[22] H. Kaur et al., "Applications of machine learning in cancer prediction and prognosis," Computational and Structural Biotechnology Journal, vol. 21, pp. 548–560, 2023, doi: 10.1016/j.csbj.2022.12.031.
[23] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," Expert Systems with Applications, vol. 91, pp. 464–471, 2018, doi: 10.1016/j.eswa.2017.09.030.
[24] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.
[25] M. Lheureux, M. Braunstein, and A. M. Oza, "Epithelial ovarian cancer: Evolution of management in the era of precision medicine," CA: A Cancer Journal for Clinicians, vol. 69, no. 4, pp. 280–304, 2019, doi: 10.3322/caac.21559.
[26] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2024," CA: A Cancer Journal for Clinicians, vol. 74, no. 1, pp. 12–49, 2024, doi: 10.3322/caac.21820.