KLASIFIKASI TIPE KANKER BERDASARKAN *SIGNATURE TUMOR DNA* MENGGUNAKAN METODE XGBOOST

(Skripsi)

Oleh

RACHELIA DITA ANGGRAINI
NPM 1657051001

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021

ABSTRACT

CLASSIFICATION OF CANCER TYPES BASED ON TUMOR SIGNATURE DNA USING XGBOOST METHOD

By

Rachelia Dita Anggraini

Cancer diagnosis using liquid biopsy focuses on tumor DNA signatures in the form of circulating tumor DNA (ctDNA) derived from primary tumors obtained from patient blood samples. Somatic mutations and the copy number alterations in ctDNA can be used as key markers for early detection and monitoring of cancer. The DNA mutation data when combined with bioinformatics and biomedical data analysis using machine learning algorithms can help simplify the classification of cancer. XGBoost or extreme gradient boosting is an ensemble learning method that combines several models to produce one optimal prediction model. This study aims to determine the results of the XGBoost algorithm classification using mutations that occur in DNA tumors in the patient's body. This study performs a classification using data from the research of Soh et al., 2017 using the XGBoost algorithm with two types of data sharing, namely 10-fold cross-validation and holdout cross-validation with training data distribution of test data of 60:40, 70:30, 75:25, 80:20, 90:10, 95:5 and 97:3. Each classification experiment carried out in this study uses two types of data, namely data with overall features and the top 900 features as a result of feature selection.

The highest accuracy result was obtained by the XGBoost holdout cross-validation experiment with 90:10 900 features with an accuracy of 89.91%, while the lowest accuracy value was obtained by the XGBoost 10-fold cross-validation 900 feature experiments, which was 70.01%. Based on the experiments that have been carried out, the accuracy value tends to be higher in the 900-feature experiment than the overall feature, and the accuracy of the experiment using the holdout cross-validation division gets a better accuracy value than the 10-fold cross-validation experiment. For holdout cross-validation, the greater the percentage of training data, the higher the accuracy obtained.

Keywords: classification, machine learning, bioinformatics, cancer, mutation, XGBoost

ABSTRAK

KLASIFIKASI TIPE KANKER BERDASARKAN *SIGNATURE TUMOR DNA* MENGGUNAKAN METODE XGBOOST

Oleh

Rachelia Dita Anggraini

Diagnosis penyakit kanker menggunakan biopsi cair berfokus pada *signature tumor DNA* berupa *circulating tumor DNA* (*ctDNA*) yang berasal dari tumor primer yang didapatkan dari sampel darah penderita. Mutasi somatik dan perubahan nomor salinan pada ctDNA dapat digunakan sebagai penanda utama untuk deteksi dan pemantauan dini dari penyakit kanker. Data mutasi DNA tersebut apabila dikombinasikan dengan bidang bioinformatika dan analisis data biomedis menggunakan algoritma *machine learning* dapat membantu mempermudah klasifikasi penyakit kanker. XGBoost atau *extreme gradient boosting* adalah metode pembelajaran *ensemble* yang menggabungkan beberapa model untuk menghasilkan satu model prediksi optimal. Penelitian ini bertujuan untuk mengetahui hasil klasifikasi algoritma XGBoost dengan menggunakan mutasi yang terjadi pada tumor DNA dalam tubuh penderita. Penelitian ini melakukan klasifikasi menggunakan data penelitian Soh et al., 2017 menggunakan algoritma XGBoost dengan dua jenis pembagian data yaitu *10-fold cross-validation* dan *holdout cross-validation* dengan pembagian data latih data uji sebesar 60:40, 70:30, 75:25, 80:20, 90:10, 95:5, dan 97:3. Masing-masing percobaan klasifikasi yang dilakukan dalam penelitian ini menggunakan dua jenis data yaitu data dengan keseluruhan fitur dan 900 fitur teratas hasil seleksi fitur.

Hasil akurasi paling tinggi didapatkan percobaan klasifikasi XGBoost *holdout cross-validation* 90:10 900 fitur dengan akurasi sebesar 89,91%, sedangkan nilai akurasi paling kecil diperoleh percobaan XGBoost *10-fold cross-validation* 900 fitur yaitu sebesar 70,01%. Berdasarkan percobaan yang sudah dilakukan, nilai akurasi cenderung lebih tinggi pada percobaan 900 fitur dibandingkan keseluruhan fitur dan akurasi percobaan menggunakan pembagian *holdout cross-validation* mendapat nilai akurasi lebih baik dibandingkan percobaan *10-fold cross-validation*. Untuk *holdout cross-validation* semakin besar persentase data latih semakin tinggi pula akurasi yang diperoleh.

Kata kunci : klasifikasi, *machine learning*, bioinformatika, kanker, mutasi, XGBoost

KLASIFIKASI TIPE KANKER BERDASARKAN *SIGNATURE TUMOR DNA* MENGGUNAKAN METODE XGBOOST

Oleh

RACHELIA DITA ANGGRAINI

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA ILMU KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2021

Judul Skripsi	: **KLASIFIKASI TIPE KANKER BERDASARKAN *SIGNATURE TUMOR DNA* MENGGUNAKAN METODE XGBOOST**
Nama Mahasiswa	: Rachelia Dita Anggraini
Nomor Pokok Mahasiswa	: 1657051001
Program Studi	: Ilmu Komputer
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Favorisen R. Lumbanraja, Ph.D.
NIP 19830110 200812 1 002

Yohana Tri Utami, M.Kom.
NIP 19900110 201903 2 010

2. Ketua Jurusan Ilmu Komputer

Didik Kurniawan, S.Si., M.T.
NIP 19800419 200501 1 004

MENGESAHKAN

1. Tim Penguji

Ketua	: **Favorisen R. Lumbanraja, Ph.D.**	.....
Penguji I Sekretaris	: **Yohana Tri Utami, M.Kom.**	.....
Penguji II Bukan Pembimbing	: **Dr. Eng. Admi Syarif**	.....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T.
NIP 19740705 200003 1 001

Tanggal Lulus Ujian Skripsi : **22 Juli 2021**

PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya yang berjudul "Klasifikasi Tipe Kanker Berdasarkan *Signature Tumor DNA* Menggunakan Metode XGBoost" merupakan karya saya sendiri dan bukan karya orang lain. Semua tulisan yang tertuang di skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila dikemudian hari terbukti skripsi saya merupakan hasil penjiplakan atau dibuat orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar yang saya terima.

Bandar Lampung, 22 Juli 2021

Rachelia Dita Anggraini
NPM. 1657051001

RIWAYAT HIDUP

Penulis dilahirkan di Bandar Lampung, Lampung pada tanggal 4 Januari 1998, sebagai anak pertama dari dua bersaudara dari ayah yang bernama Dwi Pristiwanto dan ibu bernama Aris Djuli Mariyanti. Penulis menyelesaikan Pendidikan formal pertama kali di TK Dharma Wanita Persatuan Universitas Lampung Bandar Lampung pada tahun 2004, kemudian melanjutkan pendidikan dasar di SD Negeri 1 Rajabasa Raya Bandar Lampung dan selesai pada tahun 2010. Menempuh pendidikan menengah pertama di SMP Negeri 22 Bandar Lampung pada tahun 2013, kemudian melanjutkan pendidikan menengah atas di SMA Fransiskus Bandar Lampung yang diselesaikan pada tahun 2016.

Pada tahun 2016 penulis terdaftar sebagai mahasiswa di Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung. Selama menjadi mahasiswa beberapa kegiatan yang dilakukan penulis antara lain melaksanakan kerja praktik di PT. Pelindo II Cabang Panjang Bandar Lampung pada bulan Januari 2019 dan melaksanakan Kuliah Kerja Nyata (KKN) di Desa Karya Tani, Kecamatan Labuhan Maringgai, Kabupaten Lampung Timur pada bulan Juni 2019.

MOTTO DAN PERSEMBAHAN

"Kuatkan dan teguhkanlah hatimu, janganlah takut dan jangan gemetar karena mereka, sebab TUHAN, Allahmu, Dialah yang berjalan menyertai engkau; Ia tidak akan membiarkan engkau dan tidak akan meninggalkan engkau."

(Ulangan 31:6)

"Karena masa depan sungguh ada, dan harapanmu tidak akan hilang."

(Amsal 23:18)

Skripsi ini saya persembahkan kepada:

Tuhan Yesus Kristus yang karena kasih karunia-Nya telah memberikan kesehatan dan kekuatan sehingga penulis dapat menyelesaikan skripsi ini.

Terima kasih kepada kedua orang tua, Bapak dan Ibu, yang selalu memberikan kasih sayang, dukungan, semangat, dan doa.

Terima kasih kepada bapak dan ibu dosen di Jurusan Ilmu Komputer yang senantiasa membagikan ilmu dan memberikan nasihat yang memotivasi.

Terima kasih kepada teman-teman Ilmu Komputer 2016 yang juga selalu mendukung dan berjuang bersama dalam meraih cita-cita.

SANWACANA

Puji syukur kepada Tuhan yang Maha Esa, karena atas berkat karunia-Nya dan penyertaan-Nya sehingga skripsi yang berjudul "Klasifikasi Tipe Kanker Berdasarkan *Signature Tumor DNA* Menggunakan Metode XGBoost" ini dapat diselesaikan. Skripsi ini merupakan salah satu syarat untuk memperoleh gelar sarjana Ilmu Komputer di Universitas Lampung.

Dalam penyusunan skripsi ini penulis menyadari banyak mendapat bimbingan, dukungan, dan motivasi dari berbagai pihak yang dengan tulus telah membantu penulis untuk menyelesaikan skripsi ini. Penulis sangat bersyukur dan mengucapkan terima kasih sedalam-dalamnya kepada:

1. Bapak Dekan Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T., selaku Dekan FMIPA Universitas Lampung.
2. Bapak Favorisen R. Lumbanraja, Ph.D., selaku dosen pembimbing utama atas kesediaannya dan kesabarannya untuk memberikan dukungan, bimbingan, kritik, dan saran dalam proses penyelesaian skripsi.
3. Ibu Yohana Tri Utami, S.Kom., M.Kom., selaku dosen pembimbing kedua atas kesediaannya dalam memberikan bimbingan, nasihat, juga kritik, dan saran selama proses pengerjaan skripsi.
4. Bapak Dr. Eng. Admi Syarif, selaku dosen pembahas skripsi yang telah memberikan saran dan masukan guna penyempurnaan penulisan skripsi.

5. Bapak Didik Kurniawan, S.Si., MT., selaku Ketua Jurusan Ilmu Komputer FMIPA Universitas Lampung.
6. Ibu Astria Hijriani, S.Kom., M.Kom., selaku Sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
7. Bapak dan Ibu Dosen Jurusan Ilmu Komputer Universitas Lampung yang telah memberikan ilmu dan pengetahuan hidup selama penulis menjadi mahasiswa.
8. Ibu Ade Nora Maela, Pak Zainudin, dan Mas Nofal yang telah memudahkan segala urusan administrasi penulis di Jurusan Ilmu Komputer.
9. Kedua orang tua, Bapak dan Ibu, serta Galuh yang selalu memberikan dukungan, motivasi, dan doa.
10. Teman-teman di Jurusan Ilmu Komputer Angkatan 2016 dan teman-teman Kriskat, yang telah menemani keseharian penulis selama menjadi mahasiswa dengan tawa, canda, dan inspirasi.
11. Semua pihak yang secara langsung maupun tidak langsung yang telah membantu menyelesaikan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna karena terbatasnya kemampuan serta pengetahuan yang dimiliki penulis. Namun, penulis berharap skripsi ini dapat membawa manfaat bagi orang-orang yang membacanya. Kiranya Tuhan YME selalu melimpahkan kasih dan berkat kepada seluruh pihak yang telah membantu selama proses penulisan skripsi ini.

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

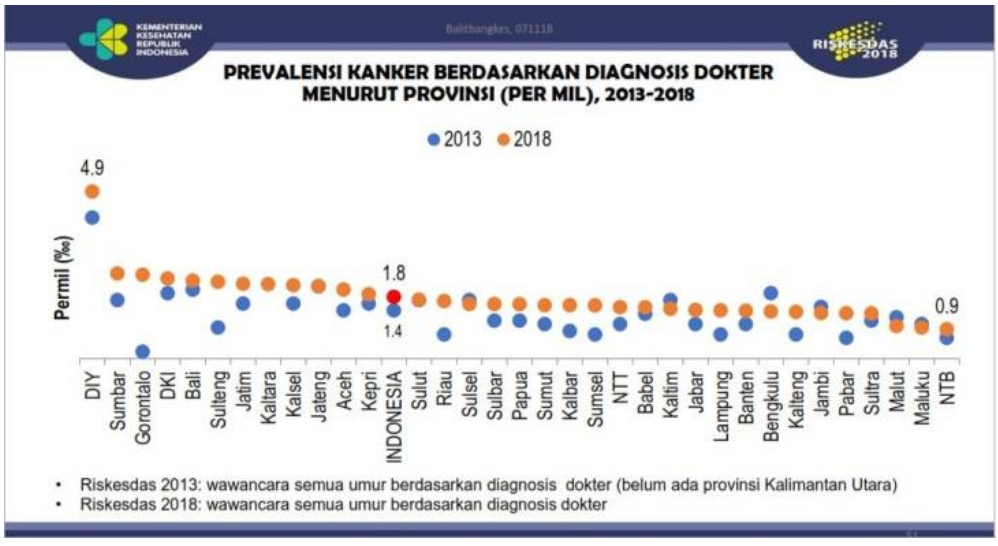
DAFTAR KODE PROGRAM

Kode Program **Halaman**

I. PENDAHULUAN

1.1. Latar Belakang Masalah

Penyakit kanker adalah salah satu penyebab kematian utama di seluruh dunia. Kanker dapat didefinisikan sebagai sekelompok penyakit yang disebabkan oleh adanya perubahan genetik maupun non-genetik ditandai dengan adanya pertumbuhan sel-sel abnormal secara tak terkendali yang mengabaikan aturan normal pembelahan sel (Nordberg et al., 2013). Berdasarkan data *Riset Kesehatan Dasar* (*Riskesdas*) tahun 2018 didapatkan prevalensi, jumlah keseluruhan penyakit yang terjadi pada suatu waktu tertentu di suatu wilayah, di Indonesia mengalami kenaikan. Gambar 1 menunjukkan pada tahun 2013 tercatat prevalensi tumor atau kanker hanya 1,4 per 1.000 penduduk Indonesia, akan tetapi pada tahun 2018 nilai prevalensi naik menjadi 1,8 per 1.000 penduduk.



Gambar 1. Prevalensi Kanker Menurut Provinsi (Per Mil) tahun 2013 sampai dengan tahun 2018 (Kementerian Kesehatan RI Badan Penelitian dan Pengembangan, 2018).

Permasalahan pada tingkat jaringan dan varietas gen pada penyakit kanker menjadi tantangan tersendiri untuk melakukan diagnosis secara spesifik (Hassanpour & Dehghani, 2017). Prosedur-prosedur pemeriksaan fisik dan penunjang untuk mendiagnosis adanya tumor ganas atau kanker seperti penggunaan sinar X dan teknologi komputer pada *Computerized Tomography scan* (CT *scan*) hanya dapat memperkirakan kehadiran kanker dalam tubuh. Diagnosis kanker menggunakan biopsi jaringan dapat secara lebih pasti mengetahui jenis tumor, jenis sel kanker, dan stadium sehingga pengobatan yang harus dijalani penderita dapat segera ditentukan. Meski efektivitas biopsi jaringan dinilai baik. Namun, risiko seperti infeksi, pendarahan, bahkan sampai dengan penyebaran kanker ke area sekitar masih mengintai penderita kanker. Selain itu, total biaya yang harus dikeluarkan untuk biopsi jaringan juga besar.

Biopsi cair atau *liquid biopsy* memberikan perspektif dan dimensi baru pada bidang medis. Biopsi cair yang berfokus pada proses pengambilan darah penderita secara rutin mampu mendeteksi secara cepat serta mengatasi keterbatasan jaringan (Mathai et al., 2019). Selain itu menurut Neumann et al., pada 2018, biopsi cair juga dinilai lebih *simple*, lebih cepat, dan lebih efisien dalam hal pemantauan status penyakit atau respons terhadap pengobatan. *Circulating tumor* DNA atau singkatnya ctDNA adalah fragmen DNA dari sel tumor yang berasal dari tumor primer dan dapat ditemukan dalam plasma darah penderita kanker. Mutasi somatik dari *circulating tumor* DNA adalah salah satu jenis penanda utama untuk deteksi dan pemantauan dini dari penyakit kanker (Xue et al., 2019). Phillippy et al., tahun 2007 menjelaskan bahwa s*ignature* DNA merupakan urutan rangkaian nukleotida yang dapat digunakan untuk mendeteksi keberadaan suatu organisme serta berguna untuk membedakan spesies organisme tersebut dengan spesies lain.

Berdasarkan data mutasi DNA yang diperoleh dari biopsi cair tersebut jika dikombinasikan dengan bidang bioinformatika dan analisis data biomedis menggunakan algoritma *machine learning* dapat membantu penderita kanker dengan mempermudah klasifikasi penyakit kanker yang diderita (Wulandari & Muflikhah, 2018). Penelitian yang dilakukan oleh Soh et al., pada 2017 selain

menggunakan mutasi somatik dalam mengidentifikasi jenis kanker penderita, perubahan nomor salinan juga digunakan untuk mengetahui keterkaitannya terhadap peningkatan klasifikasi kanker. Hasil penelitian tersebut menunjukkan bahwa klasifikasi jenis kanker menggunakan mutasi somatik dan perubahan nomor salinan terbukti meningkatkan akurasi.

Penelitian menggunakan *machine learning* untuk mengklasifikasi kanker dilakukan oleh Soh et al., pada 2017 yang mengidentifikasi sebanyak 28 jenis tipe kanker berdasarkan mutasi pada tumor DNA pasien dengan membandingkan tiga jenis algoritma yaitu *random forest, L1-regularised logistic regression*, dan *linear Support Vector Machine* (SVM). Berdasarkan penelitian tersebut dihasilkan bahwa linear SVM menghasilkan akurasi keseluruhan yang lebih tinggi dengan memakai sebanyak 900 gen berisi gabungan mutasi somatik dan perubahan nomor salinan, yaitu sebesar 88,4 ± 0,2 %. Rata-rata *precision* dari 28 tipe kanker adalah 88 ± 2 %, sedangkan rata-rata *recall* sebesar 84 ± 2 %.

Sedangkan penelitian pada tahun 2019 yang dilakukan oleh Taninaga et al., memprediksi kemungkinan pasien rumah sakit memiliki risiko besar atau kecil untuk memiliki kanker perut menggunakan dua algoritma yaitu *logistic regression* dan XGBoost (*Extreme Gradient Boosting*). Hasilnya akurasi XGBoost mengungguli *logistic regression* untuk melakukan pengawasan komprehensif pasien menggunakan data karakteristik biologis, hasil tes darah, infeksi bakteri *Helicobacter pylori*, hasil endoskopi, dan lain sebagainya.

Penelitian ini didasari oleh penelitian Soh et al., pada 2017 yang bertujuan mencari hasil klasifikasi paling baik dengan memanfaatkan mutasi yang diperoleh dari *liquid biopsy* dengan mengutamakan efisiensi dari penggunaan data menggunakan teknik pembagian data latih dan uji yang berbeda. Sehingga pada penelitian ini algoritma XGBoost dan dua teknik pembagian data latih dan data uji akan digunakan untuk melakukan klasifikasi terhadap jenis penyakit kanker berdasarkan mutasi yang terjadi pada tumor DNA yang ada dalam tubuh penderita.

1.2. Rumusan Masalah

Adapun rumusan masalah dalam penelitian ini adalah bagaimana hasil kinerja dari algoritma XGBoost dalam mengklasifikasi kanker berdasarkan perubahan pada mutasi somatik dan perubahan nomor salinan sampel tumor?

1.3. Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini, yaitu:

1. Menerapkan algoritma XGBoost menggunakan dua teknik pembagian data yaitu *holdout cross-validation* dan *10-fold cross validation* untuk mengklasifikasikan jenis kanker serta mengukur performa dari algoritma seperti *overall accuracy*, *precision*, dan *recall*.

2. Mengetahui dan mendapatkan hasil dari tiap percobaan yang dilakukan dalam penelitian.

1.4. Manfaat Penelitian

Manfaat dari penelitian ini yaitu mengetahui hasil dari pengukuran performa algoritma XGBoost untuk klasifikasi penyakit kanker berdasarkan *signature* tumor DNA dalam tubuh manusia.

1.5. Batasan Masalah

Adapun batasan masalah pada penelitian ini di antaranya sebagai berikut:

1. Proses klasifikasi penyakit kanker akan dilakukan berdasarkan tumor DNA menggunakan set data yang terdiri atas 28 jenis kanker yang diperoleh dari penelitian yang dilakukan oleh Soh et al., tahun 2017 https://doi.org/10.3929/ethz-b-000206154.

2. Penelitian ini menguji algoritma XGBoost dengan menggunakan data berisi 6640 sampel tumor dengan fitur berupa gabungan dari gen mutasi somatik dan perubahan nomor salinan sebanyak 13150 gen serta 900 gen yang paling berperan penting dalam klasifikasi.

3. Hasil *overall accuracy*, rata-rata *precision*, dan rata-rata *recall* yang diperoleh akan dibandingkan dari setiap hasil percobaan yang dilakukan menggunakan *10-fold cross-validation* dan *holdout cross-validation*.

4. Pada penelitian ini tahap *pre-processing* dan *feature extraction* tidak dilakukan kembali karena set data yang akan digunakan sudah melewati tahapan-tahapan tersebut pada penelitian sebelumnya.

II. TINJAUAN PUSTAKA

2.1. Penelitian Sebelumnya

Penelitian ini memerlukan hasil dari penelitian-penelitian terdahulu dengan topik serupa sebagai bahan kajian dan perbandingan. Berikut adalah beberapa penelitian terkait dengan penggunaan metode *machine learning* dalam bidang bioinformatika untuk mendeteksi penyakit kanker:

a. Penelitian yang dilakukan oleh Ismaeel & Mikhail tahun 2016 mendeteksi kanker berdasarkan mutasi gen p53. Penelitian ini mengklasifikasikan jenis kanker, mendiagnosis mutasi, dan memprediksi mutasi penderita kanker memakai data uji berupa sekuens gen TP53 manusia yang didapatkan dari *Catalogue of Somatic Mutation in Cancer* (COSMIC) dan menggunakan *BioEdit package* pada *National Center for Biotechnology Information* (NCBI). Data latih berasal dari http://p53.free.fr/Database /p53_MUT_MA yang berisi 53 kolom dan 1448 baris. Algoritma yang digunakan adalah *Back Propagation Network* (BPN) serta program *Alyuda NeuroIntelligence*. Hasil penelitian ini menunjukkan nilai performa 0,000006 dengan *training rate* (R) sebesar 0,9987.

b. Penelitian lainnya dilakukan oleh Wulandari & Muflikhah pada tahun 2018 mengenai klasifikasi jenis kanker berdasarkan struktur protein menggunakan algoritma *naïve bayes*. Pengujian dilakukan menggunakan set data sekuen protein TP53 isoform 1 manusia berisi gabungan string sebanyak 393 karakter yang didapatkan dari www.uniport.org dan p53.fr. Data berjumlah 848 set data dengan empat kelas yaitu *non-cancer, breast cancer*, *lung cancer*, dan *colorectal cancer* dengan pengujian

menggunakan lima set data yakni 320, 400, 480, 588, dan 848 data. Hasil akurasi maksimal yang didapatkan dengan menggunakan algoritma *naïve bayes* ini sebesar 79,17%, yaitu pada pengujian 848 data dengan persentase data uji 60%.

c. Soh et al., dalam penelitiannya pada tahun 2017 mengenai *Predicting Cancer Type from Tumour DNA Signatures* berfokus kepada prediksi 28 tipe kanker berdasarkan mutasi pada tumor DNA penderita dengan membandingkan beberapa algoritma *machine* learning. Algoritma yang digunakan dalam penelitian ini di antaranya adalah *random forest, linear* SVM, dan *L1-regularised logistic regression*. Hasil yang didapatkan yaitu *linear* SVM menghasilkan akurasi paling baik dibandingkan algoritma lain yaitu akurasi keseluruhan sebesar 88,4 ± 0,2 % yang dilatih dengan 900 gen yang terdiri dari gabungan gen mutasi somatik serta perubahan nomor salinan. Nilai *precision* yang didapatkan secara umum tinggi untuk keseluruhan tipe kanker. Rata-rata *precision* dari 28 tipe kanker adalah 88 ± 2 %, sedangkan rata-rata *recall* sebesar 84 ± 2 %.

d. Penelitian yang dilakukan oleh Taninaga et al., pada 2019 juga membahas mengenai penggunaan metode *machine learning* untuk memprediksi penyakit kanker perut dalam tubuh pasien menggunakan data-data pemeriksaan kesehatan pasien seperti data karakteristik biologis, hasil tes darah, infeksi bakteri *Helicobacter pylori*, hasil endoskopi, dan lain sebagainya pada sebuah rumah sakit di Jepang. Dalam penelitian tersebut, Taninaga bersama dengan rekannya membandingkan algoritma XGBoost dan *logistic regression* dan didapatkan kesimpulan bahwa algoritma XGBoost lebih unggul dari *logistic regression*.

2.2. Kanker

Kanker (tumor ganas) dapat didefinisikan sebagai suatu penyakit yang dapat memengaruhi organ-organ lain di dalam tubuh dengan adanya pertumbuhan sel

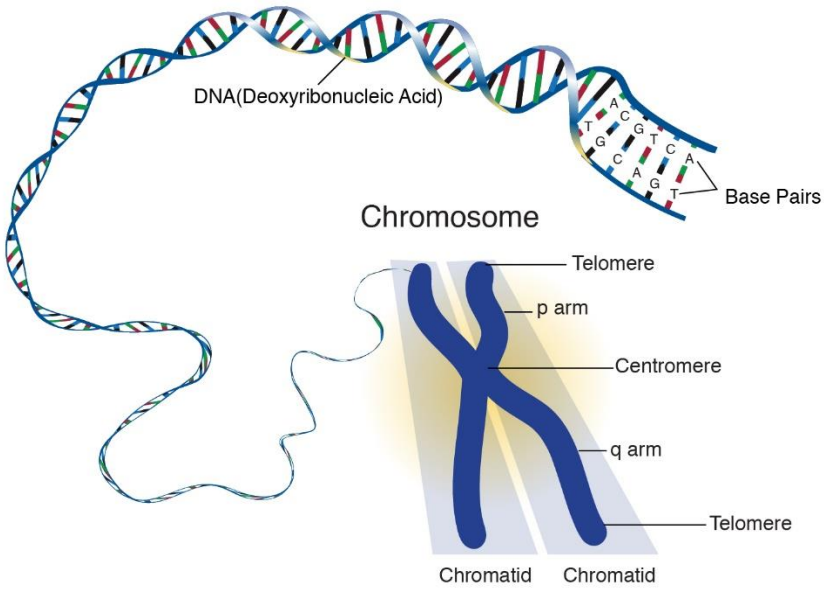
abnormal secara tidak terkendali dan penyebaran yang tidak terkontrol (Hejmadi, 2013). Menurut data kanker global terkini yang dirilis oleh *World Health Organization* (WHO), tercatat bahwa beban penyakit kanker meningkat pada tahun 2018 menjadi 18,1 juta kasus baru dan terdapat sekitar 9,6 juta kematian akibat kanker (*World Health Organization*, 2018). Sebanyak satu dari lima pria dan satu dari enam wanita di seluruh dunia telah terserang kanker. Sedangkan satu dari delapan pria dan satu dari sebelas wanita meninggal akibat kanker.

Berdasarkan data dari WHO disebutkan bahwa beberapa tipe kanker yang banyak berpengaruh di tahun 2018 ini di antaranya adalah kanker paru-paru, kanker payudara dan kanker kolorektal. Ketiga jenis kanker tersebut merupakan tiga besar teratas dalam hal insiden dan menduduki peringkat lima teratas dalam hal mortalitas, yaitu kanker paru-paru yang menduduki peringkat pertama mortalitas, kanker kolorektal peringkat kedua, dan peringkat kelima adalah kanker payudara. Menurut WHO, ketiga jenis kanker ini merupakan penyebab dari sepertiga insiden kanker dan naiknya jumlah kematian akibat kanker di seluruh dunia. Beberapa faktor risiko yang dapat menyebabkan munculnya penyakit kanker pada tubuh manusia berasal dari faktor genetik dan non-genetik (*American Cancer Society*, 2019). Contoh faktor genetik yaitu mutasi gen yang diwariskan lewat garis keturunan dan kualitas kekebalan tubuh seseorang. Sedangkan, faktor non-genetik misalnya adalah gaya hidup, obesitas, penggunaan tembakau, dan lain sebagainya.

2.3. Kromosom

Kromosom merupakan struktur mirip benang yang mengemas molekul *deoxyribonucleic acid* atau DNA pada inti setiap sel. Sebuah kromosom adalah struktur DNA *double helix* dengan protein yang paling pekat. Kromosom yang khas menyimpan ribuan gen. Setiap kromosom terdiri dari DNA yang melilit erat beberapa kali di sekitar protein yang disebut *histones* yang mendukung strukturnya. Menurut Panawala pada tahun 2017, perbedaan utama antara kromosom dan gen yaitu bahwa kromosom adalah struktur DNA yang paling

padat dengan protein, sedangkan gen adalah segmen DNA yang terletak pada kromosom. Sebuah kromosom tunggal terdiri dari banyak gen sedangkan gen adalah lokus, atau letak gen pada kromosom. Bentuk dari kromosom dapat dilihat seperti pada Gambar 2 berikut ini.

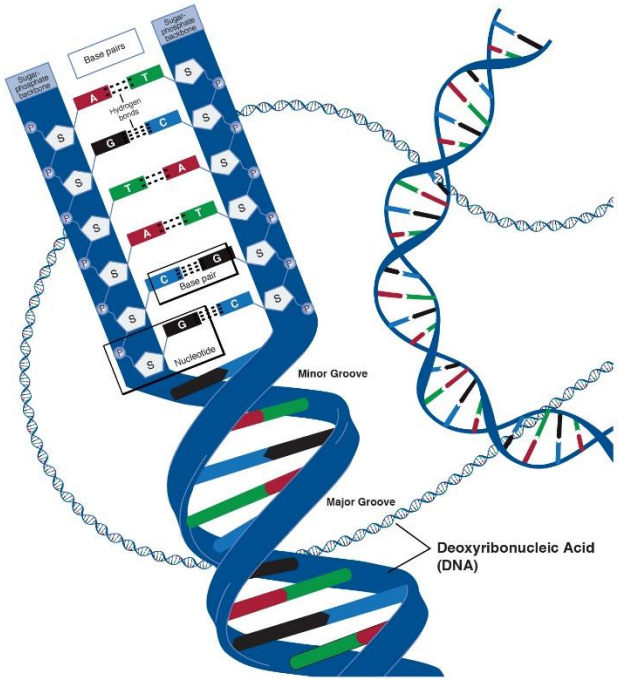


Gambar 2. Gambaran bentuk kromosom dalam tubuh manusia (Green, 2021).

2.4. DNA

DNA atau *deoxyribonucleic acid* merupakan komponen dalam tubuh manusia dan hampir semua organisme lainnya yang menjadi materi pembawa sifat hereditas. DNA manusia terdiri dari kurang lebih 3 miliar basis, dan lebih dari 99% basis tersebut sama pada semua orang (*Lister Hill National Center for Biomedical Communications,* 2019). Kebanyakan DNA ditemukan di inti sel disebut dengan *nuclear DNA*, tetapi DNA dapat ditemukan juga dengan jumlah yang sedikit pada mitokondria sehingga dikenal juga dengan sebutan *mitochondrial* DNA atau mtDNA.

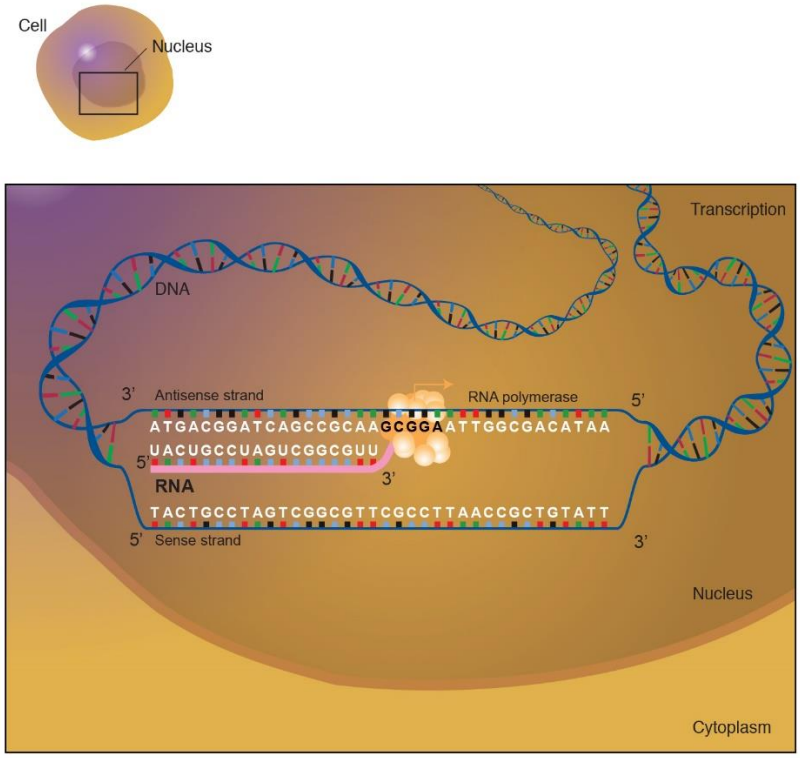
Fungsi genetik DNA dapat dipahami sebagai sinergisme dari dua sifat sebuah pita yang berisi informasi yang mengkode urutan protein dan molekul RNA serta polimer yang ada sebagai string heliks ganda yang memungkinkan pengemasan, aksesibilitas, dan replikasi penyimpanan informasi (Travers & Muskhelishvili, 2015). Hanya sekitar 1 % DNA yang terdiri dari gen penyandi protein, 99 persen lainnya adalah non-kode. Informasi pada DNA disimpan sebagai kode dengan empat basis kimia: *adenine* (A), *guanine* (G), *cytosine* (C), dan *thymine* (T). Sedangkan menurut Phillippy et al., pada tahun 2007 s*ignature* DNA merupakan urutan rangkaian nukleotida yang dapat digunakan untuk mendeteksi keberadaan suatu organisme serta berguna untuk membedakan organisme tersebut dengan spesies yang lain. Adapun bentuk DNA digambarkan seperti yang tampak pada Gambar 3.



Gambar 3. Bentuk DNA dalam tubuh manusia (Austin, 2021).

2.5. RNA

RNA atau *Ribonucleic Acid* merupakan senyawa kompleks dengan berat molekul tinggi yang berguna untuk mensintesis protein dan menggantikan asam deoksiribonukleat atau DNA sebagai pembawa kode genetik pada beberapa virus. RNA terdiri dari nukleotida ribosa (gula ribosa) yang dilekatkan oleh ikatan dan membentuk untaian dengan panjang bervariasi. Bedanya dengan DNA, informasi yang disimpan dalam RNA adalah empat basis kimia: *adenine* (A), *guanine* (G), *cytosine* (C), dan *uracil* (U). Gambar 4 menggambarkan bagaimana bentuk dari RNA dan perbedaannya dengan DNA.



Gambar 4. Perbedaan DNA dan RNA (Biesecker, 2021).

Contoh-contoh molekul RNA khusus yang dihasilkan dari DNA non-kode meliputi RNA transfer (tRNA) dan RNA ribosom (rRNA), yang membantu merakit blok pembangun protein (asam amino) ke dalam rantai yang membentuk protein (Hasler & Meister, 2016). MicroRNAs (miRNAs), yang

merupakan panjang pendek RNA yang menghambat proses produksi protein; dan RNA non-kode yang panjang (lncRNA), yang merupakan RNA yang lebih panjang yang memiliki peran beragam dalam mengatur aktivitas gen.

2.6. Gen

Dalam jurnal yang ditulis oleh Susman pada tahun 2001, menjelaskan bahwa kata gen pertama kali mulai dipopulerkan tahun 1909 oleh W. Johannsen berdasarkan konsep yang ditemukan oleh *Gregor Mendel* pada tahun 1860-an mengenai pewarisan karakteristik yang berbeda pada varietas kacang polong kebun yang saat itu benar-benar berkembang biak. Pada manusia, jumlah gen yang diperkirakan ada pada angka 20.000 hingga 25.000 gen dan tiap gen tersebut terdiri dari basis DNA sebanyak lebih dari 2 juta basis (*Lister Hill National Center for Biomedical Communications,* 2019). Gen sendiri merupakan unit fisik dan fungsional dasar pembawa sifat hereditas makhluk hidup yang digambarkan sebagai molekul organik tunggal yang didalamnya memuat DNA yang mengkode RNA yang nantinya akan berperan terhadap sintesis protein dalam tubuh.

2.7. Circulating tumor DNA (ctDNA)

Circulating tumor DNA (ctDNA) merupakan DNA dengan untai tunggal atau untai ganda (berbentuk *double helix*) yang berisikan mutasi DNA tumor asli yang dilepaskan oleh sel tumor utama ke dalam plasma darah. Tipe dari ctDNA dalam plasma dapat diidentifikasi secara lebih spesifik dengan melihat adanya perubahan molekuler yang sebelumnya terjadi di dalam jaringan tumor. Perubahan tersebut di antaranya seperti nukleotida tunggal, jumlah salinan, varian struktural, dan metilasi (Ferreira et al., 2016). ctDNA juga bergantung pada deteksi mutasi somatik yang tidak terdapat pada DNA normal. Mutasi somatik adalah perubahan dalam DNA yang terjadi setelah pembuahan. Perubahan-perubahan ini dapat (namun tidak selalu) menyebabkan kanker atau penyakit lainnya. ctDNA memiliki mutasi karakteristik yang berasal dari tumor

primer sehingga digunakan dalam manajemen kanker menggunakan biopsi cair. Hasil penelitian Bettegowa et al., dalam Cheng et al., tahun 2016 juga menyatakan bahwa dengan menggunakan ctDNA untuk mendeteksi kanker stadium I, II, III, sampai dengan stadium IV hasil persentase yang didapatkan adalah 47%, 55%, 69%, dan 82%. Hal ini menandakan bahwa semakin besar progres kanker maka semakin meningkat pula level ctDNA dalam plasma darah penderitanya.

2.8. Biopsi

Biopsi merupakan uji atau tes yang bertujuan untuk mendiagnosis keberadaan kanker dalam tubuh dengan menggunakan sampel bagian tubuh penderita seperti jaringan atau plasma darah. Menurut Crowley et al., pada tahun 2013 biopsi memungkinkan untuk mengungkapkan rincian profil genetik tumor yang mempermudah prediksi perkembangan penyakit dan bagaimana respons yang ditimbulkan terhadap terapi yang dijalani. Terdapat beberapa jenis biopsi untuk mendiagnosis kanker, beberapa di antaranya yaitu biopsi jaringan dan biopsi cair. Menurut Jr & Bardelli pada jurnalnya tahun 2014, disebutkan bahwa biopsi cair dengan mengakses pembuluh darah memberikan keuntungan seperti mendapatkan sumber DNA yaitu *circulating tumor* DNA yang dapat memberikan informasi genetik sama seperti biopsi jaringan. Selain itu juga metode pada biopsi cair lebih minim dalam hal invasif dan risiko komplikasi.

2.9. *Machine Learning*

Machine Learning adalah kategori kecerdasan buatan yang memungkinkan komputer untuk berpikir dan belajar sendiri yang mana simulasinya terkait dengan komputasi statistik untuk membuat prediksi melalui komputer (Alzubi et al., 2018). Pembelajaran mesin biasanya digunakan dalam perencanaan, diagnosis, prediksi, pengenalan, kontrol robot, serta tugas-tugas lainnya yang terkait dengan kecerdasan buatan atau sering disebut dengan *artificial intelligence* (Mohammed et al., 2016). Tipe *machine learning* sendiri terdiri

dari *supervised learning* yang membuat sebuah model dengan mempelajari data latih berlabel, sedangkan *unsupervised learning* mempelajari kemiripan dari data latih tak berlabel.

2.9.1. *Unsupervised Learning*

Unsupervised learning merupakan pembelajaran dalam *machine learning* yang menggunakan algoritma untuk mengidentifikasi pola tersembunyi dari data input tak berlabel. *Unsupervised learning* sangat berguna untuk mengetahui variasi dan struktur grup dari data dengan tipe tak berlabel, dan *unsupervised learning* dapat berguna untuk pra-pemrosesan *supervised learning* (Ghahramani, 2004; Sathya & Abraham, 2013). Saat data baru digunakan atau diperkenalkan, fitur sebelumnya digunakan untuk mengenali kelas dari data tersebut (Dey, 2016). Pengelompokan (*clustering*) dan pengurangan fitur (*feature reduction*) sering menggunakan *unsupervised learning* dalam penerapannya. Beberapa algoritma yang digunakan dalam pengelompokan dan *feature reduction* yaitu *K-means* dan *principal component analysis* (Dey, 2016).

2.9.2. *Supervised Learning*

Supervised learning merupakan algoritma yang memerlukan bantuan eksternal, dimana didasarkan pada melatih sampel data dari sumber data dengan klasifikasi yang benar, untuk menghasilkan hipotesis umum atau model dari distribusi label kelas dalam hal prediksi atau klasifikasi (Muhammad & Yan, 2015; Y et al., 2017) . Pada algoritma ini set data masukan dibagi menjadi data latih dan data uji. Data latih memiliki variabel keluaran yang perlu diprediksi atau diklasifikasikan. Algoritma-algoritma yang sering digunakan dalam *supervised learning* ini misalnya *decision tree, naïve bayes, support vector machine*, dan *extreme gradient boosting* (XGBoost).

2.9.2.1. Klasifikasi

Kesavaraj & Sukumaran dalam Soofi & Awan pada tahun 2017 menyatakan bahwa klasifikasi adalah pendekatan *data mining* (pembelajaran mesin) yang berguna untuk memperkirakan keanggotaan grup atau kelas kategorikal untuk data. Klasifikasi memetakan fungsi (*f*) dari variabel masukan (*X*) menjadi variabel keluaran diskrit (*y*). Hasil dari variabel keluaran sering disebut juga dengan label atau kategori. Banyak cara dapat digunakan untuk memperkirakan kemampuan dari model klasifikasi yang dibuat, akan tetapi yang paling umum adalah menghitung akurasi klasifikasi tersebut. Akurasi klasifikasi merupakan persentase dari contoh yang diklasifikasikan dengan benar dari semua prediksi yang dihasilkan. Algoritma yang mampu untuk mempelajari model klasifikasi disebut sebagai algoritma klasifikasi.

2.9.2.2. Regresi

Menurut Halili & Rustemi pada tahun 2016 regresi dapat digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen dan variabel dependen. Dalam *data mining*, variabel independen adalah atribut yang sudah diketahui dan variabel respons adalah apa yang ingin diprediksi.

2.10. *Extreme Gradient Boosting* (XGBoost)

XGBoost atau *extreme gradient boosting*, yang adalah implementasi dari kerangka *gradient boosting* yang efisien dan terukur (Chen & He, 2021). XGBoost adalah salah satu implementasi paling populer dan efisien dari algoritma *Gradient Boosted Trees,* sebuah metode *supervised learning* yang didasarkan pada perkiraan fungsi dengan mengoptimalkan fungsi kerugian spesifik dan menerapkan beberapa teknik regularisasi.

XGBoost adalah metode pembelajaran *ensemble*. Metode pembelajaran *ensemble* merupakan teknik pembelajaran mesin yang menggabungkan beberapa model dasar untuk menghasilkan satu model prediksi yang optimal. Pembelajaran *ensemble* menawarkan solusi sistematis untuk menggabungkan kekuatan prediktif banyak *learner*. Hasilnya adalah model tunggal yang memberikan *output* agregat dari beberapa model. *Boosting* mengacu pada algoritma yang mampu mengubah *weak learner* menjadi *strong learner* dengan prinsip utama menyesuaikan urutan model *weak learner*, yang hanya sedikit lebih baik daripada menebak secara acak seperti pohon keputusan sederhana, untuk versi data yang berbobot. Lebih banyak bobot diberikan pada contoh yang salah diklasifikasi oleh putaran sebelumnya. Prediksinya kemudian digabungkan melalui suara terbobot mayoritas atau *weighted majority vote* (klasifikasi) ataupun jumlah tertimbang atau *weighted sum* (regresi) untuk menghasilkan prediksi akhir. Adapun langkah-langkah dalam menerapkan algoritma XGBoost ini di antaranya sebagai berikut:

a. Langkah pertama yaitu menentukan nilai *initial prediction*. Nilai *initial prediction* memiliki nilai *default* 0,5 baik untuk regresi atau klasifikasi. Nilai ini bisa mencakup hal-hal seperti probabilitas dari pengamatan pada data latih yang diobservasi. Angka tersebut menandakan bahwa ada 50% kemungkinan yang sedang diobservasi bernilai positif (*true*).

b. Selanjutnya setelah nilai *initial prediction* ditentukan, akan diperoleh nilai *residual*. Nilai *residual* merupakan perbedaan yang ada atau *gap* antara sampel yang sedang diobservasi dengan yang diprediksi yang dapat dilihat pada Persamaan (1).

$$Residual = Observed - Predicted \quad (1)$$

c. Sama halnya dengan XGBoost regresi, untuk menyesuaikan bentuk XGBoost *tree* ke nilai-nilai *residual* digunakan rumus *Similarity Score*

yang sedikit berbeda untuk regresi. Rumus *Similarity Score* untuk klasifikasi tertera seperti dalam Persamaan (2).

$$Similarity\ score = \frac{\left(\sum Residual_i\right)^2}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)] + \lambda} \ldots\ldots (2)$$

Keseluruhan *residual* dimasukkan ke dalam satu *leaf* yang sama dan dihitung nilai *Similarity Score* dari *leaf* tersebut.

d. Untuk mengetahui apakah kinerja mengelompokkan *residual* yang serupa akan lebih baik apabila dilakukan pemisahan menjadi dua grup terpisah digunakan perbandingan nilai *Gain* dari tiap kemungkinan pemisahan yang ada. Nilai *Gain* menentukan seperti apa *branch* dari tree yang sedang dikerjakan. *Gain* dirumuskan seperti dalam Persamaan (3).

$$Gain = Left_{similarity} + Right_{similarity} - Root_{similarity} \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Dari semua kemungkinan nilai *Gain* yang didapatkan setelah pemisahan tiap sampel yang diobservasi, nilai *Gain* tertinggi dipilih menjadi *branch* yang memisahkan *residual*.

e. Selanjutnya dilakukan pengecekan apakah masih ada *residual-residual* di dalam *leaf* yang masih dapat dipisahkan dan dibentuk menjadi *branch*. Proses *splitting* atau pemisahan dilakukan sesuai dengan batas kedalaman tree yang ditentukan. Kedalaman dari *Tree* XGBoost standarnya memungkinkan hingga 6 level kedalaman dan kedalaman levelnya bebas ditentukan.

f. Jumlah minimum dari *residual* dalam tiap-tiap *leaf* dan ditentukan dengan menghitung nilai *Cover*. Secara *default*, nilai minimum dari *Cover* adalah 1 dan apabila nilai *Cover* suatu *leaf* kurang dari 1 maka XGBoost tidak memperbolehkan *leaf* tersebut. Cover dirumuskan seperti dalam Persamaan (4).

$$Cover = \sum[\ Previous\ Probability_i \times (1 - Previous\ Probability_i)] \quad \text{.....(4)}$$

g. Untuk melakukan pemangkasan atau *pruning* dari *tree*, yang perlu dilakukan adalah menghitung selisih antara *Gain* dari cabang atau *branch* paling bawah dari *tree* dengan nilai γ (gamma) yang sudah ditetapkan. Apabila selisih yang didapatkan bernilai positif maka tidak akan dilakukan pemangkasan pada *branch* tersebut. Namun, jika bernilai negatif maka pemangkasan dilakukan dan berhenti hingga mencapai *branch* lain yang bernilai positif.

h. Dalam melakukan pemangkasan atau *pruning* ini dapat terjadi pemangkasan ekstrim yaitu kondisi dimana *tree* habis terpangkas dan menyisakan nilai *initial prediction*. Untuk mencegah terjadinya *extreme pruning*, *Regularization Parameter*, atau λ (lambda) pada mengurangi nilai *Similarity Scores* yang berdampak ke menurunnya nilai *Gain*. Nilai λ (lambda) lebih dari 0 mengurangi sensitivitas dari *tree* ke *individual observation* dengan *pruning* dan menggabungkan mereka dengan observasi lain.

i. Jika *tree* sudah terbentuk, langkah selanjutnya adalah menghitung nilai *Output Value* dari tiap-tiap node pada *tree* yang dirumuskan seperti Persamaan (5).

$$Output\ Value = \frac{(\sum Residual_i)}{\sum[Previous\ Probability_i \times (1 - Previous\ Probability_i)] + \lambda} \quad \text{.....(5)}$$

j. Selanjutnya prediksi baru dapat dilakukan. Seperti metode *boosting* lainnya, XGBoost untuk klasifikasi membuat prediksi baru dengan mengawalinya dengan *initial prediction*. Seperti pada *Gradient Boost Classification*, perlu untuk mengubah nilai probabilitas pada *initial prediction* menjadi nilai *log(odds). Odds* atau peluang (secara teknis peluang keberhasilan) didefinisikan sebagai probabilitas keberhasilan

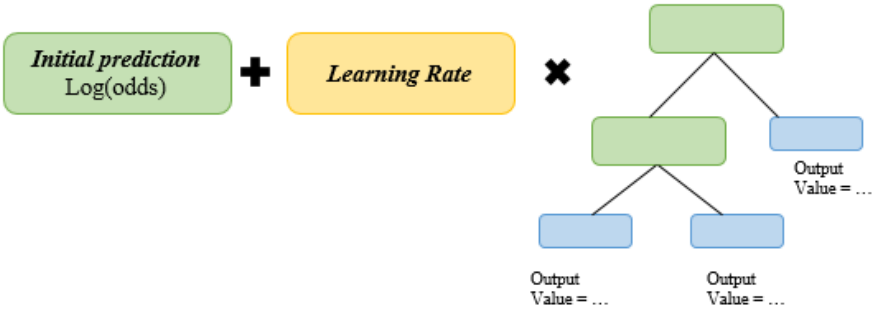
dibagi dengan probabilitas kegagalan dan dirumuskan dengan Persamaan (6).

$$odds = \frac{Probability}{1-Probability} \quad \text{.....(6)}$$

Sedangkan *log(odds)* adalah logaritma peluang yang dirumuskan seperti dalam Persamaan (7).

$$\log(odds) = log\left(\frac{Probability}{1-Probability}\right) \quad \text{.....(7)}$$

Kemudian mulai membuat prediksi baru atau *new prediction* dengan menggabungkan nilai-nilai seperti pada Gambar 5.



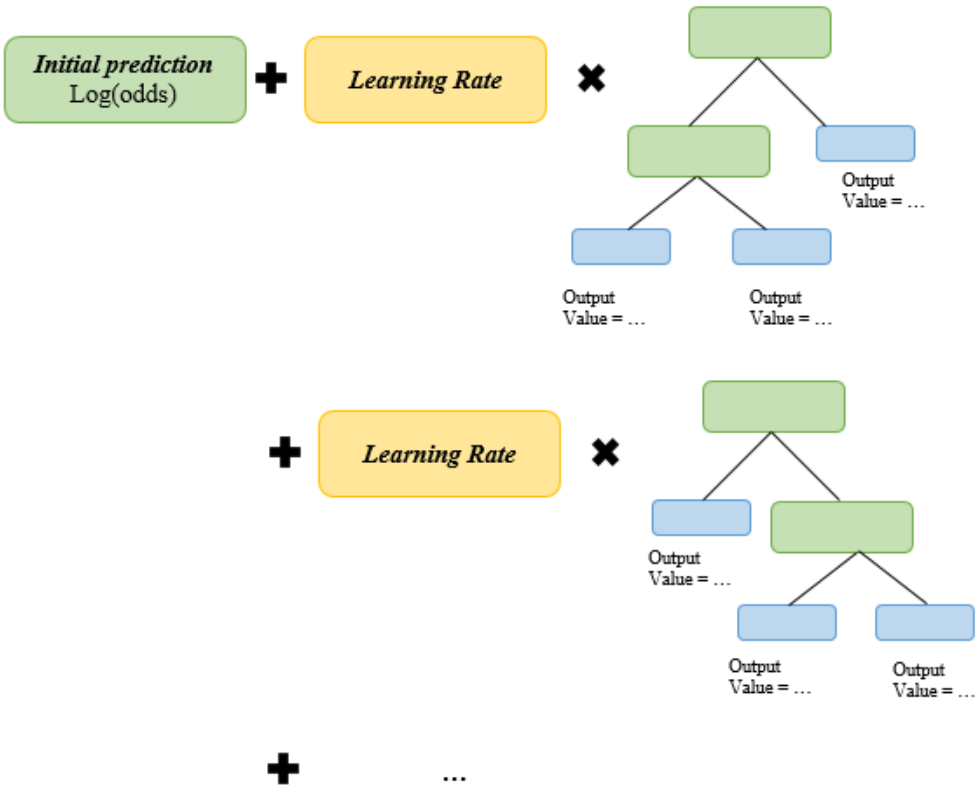
Gambar 5. *New prediction* dalam XGBoost

Learning rate atau eta dalam XGBoost secara *default* bernilai 0,3 dan nilainya bebas ditentukan. Jika nilai *new prediction* sudah didapat, selanjutnya adalah mengubah nilai yang diperoleh ke dalam probabilitas dengan *Logistic Function* seperti yang dirumuskan dalam Persamaan (8).

$$\text{Probability} = \frac{e^{\log(odds)}}{1+e^{\log(odds)}} \quad \text{.....(8)}$$

Apabila nilai *new prediction* menghasilkan nilai probabilitas yang membuat nilai *residual* makin kecil dari sebelumnya maka prediksinya mengarah ke arah yang benar.

k. Untuk membuat tree selanjutnya, acuan berubah pada nilai *residual* baru atau dengan kata lain *tree* selanjutnya menyesuaikan nilai *residual* yang baru dihasilkan. Alur algoritma ditunjukkan pada Gambar 6. Proses diulang seperti di awal dan *tree* terus menerus dibuat sampai nilai-nilai *residual* menjadi semakin sangat kecil atau pembuatan *tree* sudah mencapai maksimum.



Gambar 6. Penggambaran algoritma XGBoost

2.11. Seleksi Fitur (*Feature Selection*)

Wang et al., dalam penelitiannya tahun 2016 menjelaskan bahwa *feature selection* atau seleksi fitur merupakan cara yang efektif dan efisien dalam mempersiapkan data dengan dimensi yang tinggi untuk *data mining* dan *machine learning*. Seleksi fitur adalah salah satu konsep inti dalam pembelajaran mesin yang sangat memengaruhi kinerja model. Hal ini karena fitur data yang digunakan untuk melatih model berpengaruh besar terhadap

kinerja yang akan dicapai. Fitur yang tidak relevan atau sedikit sekali relevan dapat berdampak negatif terhadap kinerja model, yaitu mengurangi akurasi model dan membuat model belajar berdasarkan fitur yang tidak relevan. Seleksi fitur merupakan proses, secara otomatis atau manual, memilih fitur-fitur yang paling berkontribusi terhadap keluaran. Sedangkan, manfaat yang diperoleh dari seleksi fitur adalah mengurangi waktu *learning*¸ meningkatkan akurasi, dan mengurangi *overfitting*. *Overfitting* yaitu model memiliki *error* (galat) yang rendah pada data latih tetapi memiliki galat yang sangat tinggi pada data uji.

2.11.1. *Feature Importance*

Feature Importance merupakan salah satu teknik seleksi fitur yang mudah digunakan dan memberikan hasil yang baik. Pada *feature importance*, terdapat skor untuk setiap fitur data, semakin tinggi skor yang dimiliki fitur maka semakin penting atau semakin relevan fitur tersebut terhadap hasil keluaran model. Teknik *feature importance* adalah kelas *built-in* yang ada di dalam klasifier *tree-based*. Dalam algoritma XGBoost, setelah *boosted trees* terbentuk, akan relatif mudah untuk mengambil skor penting untuk setiap atribut. Semakin banyak atribut tersebut digunakan untuk membuat *key decision* dengan pohon keputusan, semakin tinggi pula kepentingan relatifnya. Kepentingan ini dihitung secara eksplisit untuk tiap atribut dalam set data, berdasarkan hal ini atribut akan diurutkan dan dibandingkan. Kepentingan dihitung untuk pohon keputusan berdasarkan jumlah dari tiap *split point* atribut meningkatkan ukuran performa, ditimbang dengan jumlah observasi yang bertanggung jawab atas node. Kepentingan fitur kemudian dirata-rata di semua pohon keputusan dalam model. Untuk algoritma XGBoost sendiri, urutan kepentingan fitur paling penting berdasarkan hal berikut.

2.11.1.1. *Gain*

Gain memberi indikasi informasi mengenai bagaimana fitur tersebut penting dalam membuat cabang dari pohon keputusan menjadi lebih murni.

2.11.1.2. *Cover*

Cover mengukur jumlah relatif dari observasi yang terkait dengan fitur.

2.11.1.3. *Frequency*

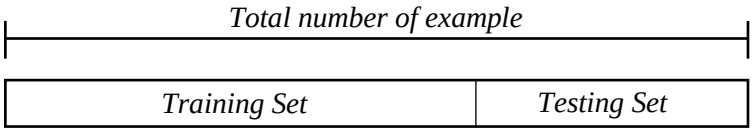
Frequency menghitung berapa kali fitur tersebut digunakan di semua pohon keputusan yang dihasilkan.

2.12. *Cross-Validation*

Cross-Validation merupakan metode statistik untuk menguji keefektifan model pembelajaran mesin. Menurut Refaeilzadeh et al., pada tahun 2008 *cross-validation* mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen: satu digunakan untuk mempelajari atau melatih suatu model dan yang lainnya digunakan untuk memvalidasi model.

2.12.1. *Holdout Cross-Validation*

Menurut Koul et al., tahun 2018 *holdout validation* adalah bentuk paling sederhana dari *cross-validation* dan sering dianggap sebagai metode validasi. Prosedur *holdout* membagi data hanya sekali menjadi dua bagian yaitu set pelatihan dimana set ini digunakan untuk melatih model sedangkan set uji untuk mengevaluasi kinerja model. Jumlah data yang digunakan untuk data latih secara acak dan bagian data yang tersisa umumnya 1/3 bagian ditetapkan sebagai data uji. Keuntungan metode ini adalah beban komputasinya yang lebih rendah. *Holdout cross-validation* ditunjukkan seperti Gambar 7.



Gambar 7. Penggambaran *Holdout cross-validation*

2.12.2. *k-fold Cross-Validation*

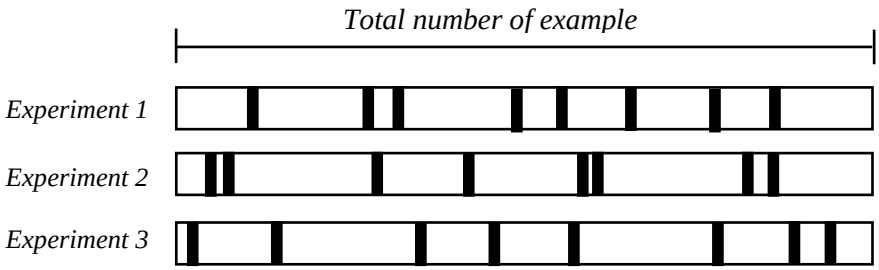
Koul et al., tahun 2018 menyatakan metode ini menghindari keacakan yang berasal dari perkiraan yang dihasilkan dengan memisahkan data hanya sekali. Pada k-*fold cross-validation* seperti pada Gambar 8, data pertama kali dipartisi menjadi segmen atau lipatan yang sama (atau hampir sama). Selanjutnya k iterasi pelatihan dan validasi dilakukan sedemikian rupa sehingga dalam setiap iterasi lipatan data yang berbeda digunakan untuk validasi atau *testing* sedangkan k - 1 (k minus 1) lipatan lainnya digunakan untuk pembelajaran atau *training*. Akibatnya, akurasi yang berbeda dihasilkan dari prosedur. Dalam *data mining* dan *machine learning* 10-*fold cross-validation* (k = 10) adalah yang paling umum digunakan.

Gambar 8. Gambaran *k-Fold Cross Validation*

2.12.3. *Leave One Out Cross-Validation*

Leave-one-out cross-validation (LOOCV) merupakan teknik validasi silang yang lengkap. Dimana set data dengan n baris, baris ke-1 digunakan untuk validasi, dan baris lainnya yaitu n-1 digunakan untuk melatih model. Kemudian pada iterasi berikutnya, baris ke-2 digunakan

untuk validasi dan sisanya untuk melatih model. Proses diulangi hingga n langkah atau sesuai dengan jumlah operasi yang diinginkan. Menurut Berrar dalam jurnalnya tahun 2018 kesalahan pengujian dalam LOOCV kira-kira merupakan perkiraan yang tidak bias dari kesalahan prediksi yang sebenarnya, tetapi memiliki varians yang tinggi, karena n set pelatihan secara praktis sama. Biaya komputasi LOOCV juga bisa sangat tinggi untuk n besar, terutama jika pemilihan fitur harus dilakukan. Teknik *Leave-one-out cross-validation* ditunjukkan pada Gambar 9.



Gambar 9. Gambaran *Leave One Out Cross-Validation*

2.13. Confusion Matrix

Menurut Novakovic et al., tahun 2017 *Confusion matrix* memuat informasi mengenai aktual klasifikasi dan prediksi klasifikasi yang dilakukan oleh sistem klasifikasi. Kinerja atau performa sistem klasifikasi tersebut biasanya dievaluasi menggunakan data dalam matriks. Tabel 1 menunjukkan *confusion matrix* untuk dua kelas klasifier.

Tabel 1. Tabel *confusion matrix*

		Actual Values	
		Positive	***Negative***
Predicted Values	***Positive***	**TP** (*True Positive*)	**FP** (*False Positive*)
	Negative	**FN** (*False Negative*)	**TN** (*True Negative*)

Entri dalam *confusion matrix* memiliki arti sebagai berikut:

- 'TP' adalah jumlah prediksi yang tepat bahwa sebuah *instance* adalah positif atau *True Positive*,
- 'FP' adalah jumlah prediksi yang salah bahwa sebuah *instance* adalah positif atau *False Positive*,
- 'FN' adalah jumlah prediksi yang salah bahwa sebuah *instance* adalah negatif atau *False Negative*,
- 'TN' adalah jumlah prediksi yang tepat bahwa sebuah *instance* adalah negatif atau *True Negative*.

2.13.1. *Accuracy*

Novakovic et al., tahun 2017 dalam jurnalnya mengatakan bahwa *Accuracy* mengevaluasi model dengan melihat rasio jumlah dari contoh kasus yang diklasifikasi secara tepat dengan total banyaknya contoh yang diklasifikasi dan dirumuskan dengan Persamaan (9).

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Number\ of\ Cases} \quad \text{.....(9)}$$

Atau dapat juga ditulis seperti dalam Persamaan (10).

$$Accuracy = \frac{Number\ of\ Correctly\ Classified\ Examples}{Total\ Number\ of\ Cases} \quad \text{.....(10)}$$

2.13.2. *Precision*

Precision atau *confidence* membahas mengenai proporsi dari kasus *Predicted Positive* yang benar *Real Positive* (Powers, 2011). Perhitungan *precision* dirumuskan dalam Persamaan (11).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \text{.....(11)}$$

Atau dapat ditulis sebagai berikut seperti dalam Persamaan (12).

$$Precision = \frac{True\ Positive}{Total\ Predicted\ Positive} \quad \text{.....(12)}$$

2.13.3. *Recall*

Sebaliknya, *Recall* atau *Sensitivity* adalah proporsi dimana kasus *Real Positive* yang dengan benar diprediksi sebagai positif atau *Predicted Positive* (Powers, 2011). Recall dirumuskan dalam Persamaan (13).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \text{.....(13)}$$

Atau dapat ditulis seperti dalam Persamaan (14).

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \quad \text{.....(14)}$$

III. METODE PENELITIAN

3.1. Tempat dan Waktu

3.1.1. Tempat

Penelitian dilakukan di Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung (Unila) yang beralamat lengkap di Jalan Prof. Dr. Ir. Sumantri Brojonegoro No. 1, Gedong Meneng, Kecamatan Rajabasa, Kota Bandar Lampung, Lampung 35141.

3.1.2. Waktu

Penelitian mulai dilaksanakan pada semester genap tahun ajaran 2019/2020 tepatnya pada awal bulan Februari 2020 dan direncanakan selesai sampai dengan bulan Februari tahun ajaran 2020/2021. Tabel 2 berisi tahapan yang harus dilalui dalam penelitian, tahapan tersebut di antaranya yaitu tahap penulisan laporan bab satu sampai dengan bab tiga selama empat bulan, tahap *import* data sekaligus *data preparation* selama satu minggu, tahap pembagian data dengan *holdout* dan 10-*fold cross-validation* selama tujuh minggu, tahap klasifikasi selama empat minggu, tahap *feature selection* selama satu bulan, tahap evaluasi performa *classifier* yang berlangsung selama satu bulan, dan tahapan penulisan laporan bab empat dan lima yang berjalan sejak awal bulan November 2020 sampai dengan bulan Februari 2021.

Tabel 2. Tahapan penelitian dan waktu pengerjaan

No.	Langkah	Feb				Mar				Apr				Mei				Jun				Jul				Ags				Sep				Okt				Nov				Des				Jan				Feb			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	**Studi Literatur**																																																	
2	***Import Data*** Menggunakan data gabungan mutasi somatik dan perubahan nomor salinan gen sebanyak 6640 sampel & 13151 fitur																																																		
3	***Data Preparation*** Mengubah nilai kelas label menjadi integer mulai dari nol																																																		
4	***10-fold cross validation*** Membagi data sebanyak k lipatan dengan nilai k = 10 ***holdout cross-vaidation*** Membagi data menjadi 90% data latih dan 10% data uji																																																		
5	**Klasifikasi** Melakukan klasifikasi menggunakan *package XGBoost*																																																		
6	***Feature selection*** Mengetahui fitur atau gen paling baik untuk mengklasifikasi dengan menggunakan *feature importance* berdasarkan tingkat kepentingannya																																																		
7	**Evaluasi performa** Mendapatkan nilai *overall accuracy, precision* , dan *recall* dari tiap percobaan yang dilakukan																																																		
8	**Penulisan laporan bab 4-5**																																																		

3.2. Data dan Alat

3.2.1. Data

Penelitian ini menggunakan data mutasi somatik beserta perubahan jumlah salinan gen pada 6640 sampel tumor dan terdiri atas 28 jenis kanker yang digunakan dalam penelitian kanker (Soh et al., 2017) yang dapat diunduh melalui https://doi.org/10.3929/ethz-b-000206154.

	ZSWIM6	ZSWIM8	ZW10	ZWINT	ZXDA	ZXDB	ZYG11A	ZYG11B	ZZEF1	classLabels
TCGA.AC.A3TM.01	0	0	0	0	0	0	0	0	0	2
TCGA.AC.A3W7.01	0	0	0	0	0	0	0	0	0	2
TCGA.AC.A5EH.01	0	0	0	0	0	0	0	0	0	2
TCGA.AC.A5EI.01	0	0	0	0	0	0	0	0	0	2
TCGA.AC.A5XU.01	0	0	0	0	0	0	0	0	0	2
TCGA.AR.A2LJ.01	0	0	0	0	0	0	0	0	0	2
TCGA.B6.A3ZX.01	0	0	0	0	0	0	0	0	0	2
TCGA.B6.A400.01	0	0	0	0	0	0	0	0	0	2
TCGA.B6.A401.01	0	0	0	0	0	0	0	0	0	2
TCGA.B6.A409.01	0	0	0	0	0	0	0	0	0	2
TCGA.BH.A5IZ.01	0	0	0	0	0	0	0	0	0	2
TCGA.BH.A5J0.01	0	0	0	0	0	0	0	0	0	2
TCGA.HN.A2NL.01	0	0	0	0	0	0	0	0	0	2
TCGA.HN.A2OB.01	0	0	0	0	0	0	0	0	0	2
TCGA.PE.A5DC.01	0	0	0	0	0	0	0	0	0	2
TCGA.PE.A5DD.01	0	0	0	0	0	0	0	0	0	2
TCGA.PE.A5DE.01	0	0	0	0	0	0	0	0	0	2
TCGA.A6.2670.01	0	0	0	0	0	0	0	0	0	3
TCGA.A6.2677.01	0	0	0	0	0	0	0	0	0	3
TCGA.AA.3524.01	0	0	0	0	0	0	0	0	0	3
TCGA.AA.3529.01	0	0	0	0	0	0	0	0	0	3
TCGA.AA.3532.01	0	0	0	0	0	0	0	0	0	3

Gambar 10. Gambaran set data 6640×7673 mutasi somatik

Matriks mutasi somatik ditunjukkan oleh Gambar 10 dan matriks perubahan nomor salinan ditunjukkan oleh Gambar 11. Gambar 10 berisi matriks mutasi somatik kanker dengan ukuran 6640×7673. Tiap baris pada matriks tersebut merupakan sampel tumor, sedangkan kolom pada matriks adalah gen mutasi somatik. Angka (1) pada posisi ke-(*i*,*j*) pada matriks menandakan bahwa gen ke-*j* membawa mutasi somatik dalam sampel tumor ke-*i*, sedangkan angka (0) menandakan yang sebaliknya, yaitu gen ke-*j* tidak membawa mutasi somatik dalam sampel tumor ke-*i*.

	ZSCAN1.p	ZSCAN1.n	ZSCAN18.p	ZSCAN18.n	ZSCAN22.p	ZSCAN22.n	ZSCAN4.n	ZSWIM7.p	ZSWIM7.n	classLabels
TCGA.AC.A3TM.01	0	0	0	0	0	0	0	0	0	2
TCGA.AC.A3W7.01	0	0	0	0	0	0	0	0	1	2
TCGA.AC.A5EH.01	0	0	0	0	0	0	0	0	1	2
TCGA.AC.A5EI.01	1	0	1	0	1	0	0	0	0	2
TCGA.AC.A5XU.01	1	0	1	0	1	0	0	0	1	2
TCGA.AR.A2LJ.01	0	0	0	0	0	0	0	0	1	2
TCGA.B6.A3ZX.01	0	0	0	0	0	0	0	0	0	2
TCGA.B6.A400.01	0	1	0	1	0	1	1	0	0	2
TCGA.B6.A401.01	1	0	1	0	1	0	0	1	0	2
TCGA.B6.A409.01	0	1	0	1	0	1	1	0	1	2
TCGA.BH.A5IZ.01	0	1	0	1	0	1	1	0	1	2
TCGA.BH.A5J0.01	0	1	0	1	0	1	1	0	1	2
TCGA.HN.A2NL.01	0	1	0	1	0	1	1	1	0	2
TCGA.HN.A2OB.01	0	0	0	0	0	0	0	0	1	2
TCGA.PE.A5DC.01	1	0	1	0	1	0	0	0	1	2
TCGA.PE.A5DD.01	0	0	0	0	0	0	0	0	1	2
TCGA.PE.A5DE.01	0	0	0	0	0	0	0	0	1	2
TCGA.A6.2670.01	1	0	1	0	1	0	0	0	1	3
TCGA.A6.2677.01	0	0	0	0	0	0	0	0	1	3
TCGA.AA.3524.01	0	0	0	0	0	0	0	0	1	3
TCGA.AA.3529.01	1	0	1	0	1	0	0	0	0	3
TCGA.AA.3532.01	0	0	0	0	0	0	0	0	1	3

Gambar 11. Gambaran set data 6640×5477 (*copy number alteration*)

Gambar 11 menggambarkan matriks perubahan nomor salinan (*copy number alteration*) pada kanker dengan ukuran 6640×5477. Tiap baris pada matriks tersebut merupakan sampel tumor, sedangkan kolom pada matriks adalah gen yang mengalami perubahan nomor salinan. Akhiran pada tiap kolom menandakan tipe alterasi atau perubahan yang dialami gen. Akhiran (p) mewakili positif yang berarti gen mengalami amplifikasi atau peningkatan jumlah salinan gen tanpa peningkatan proporsional pada gen lain. Sedangkan akhiran (n) mewakili negatif yang berarti gen mengalami delesi atau terdapat sesuatu yang hilang dari materi genetik. Angka (1) pada posisi ke-(*i*,*j*) pada matriks menandakan bahwa gen ke-*j* mengalami perubahan nomor salinan dalam sampel tumor ke-*i*. Sedangkan angka (0) menandakan yang sebaliknya. Tabel 3 berisi rincian 28 jenis kanker yang dimuat dalam data beserta kelas dan jumlah sampel tiap kelasnya.

Tabel 3. Tabel 28 jenis kanker yang dimuat dalam data

Tipe Kanker	Kelas Label	Banyak Sampel
Bladder urothelial carcinoma	1	127
Breast invasive carcinoma	2	973
Colorectal adenocarcinoma	3	212
Glioblastoma	4	280
Head and neck squamous cell carcinoma	5	279
Kidney renal clear cell carcinoma	6	418
Acute myeloid leukaemia	7	190
Lung adenocarcinoma	8	230
Lung squamous cell carcinoma	9	178
Ovarian serous cystadenocarcinoma	10	316
Uterine corpus endometrial carcinoma	11	240
Adenoid cystic carcinoma	12	55
Brain lower grade glioma	13	279
Cervical squamous cell carcinoma and endocervical adenocarcinoma	14	191
Kidney renal papillary cell carcinoma	15	161
Liver hepatocellular carcinoma	16	231
Pancreatic adenocarcinoma	17	145
Prostate adenocarcinoma	18	331
Skin cutaneous melanoma	19	278
Stomach adenocarcinoma	20	287
Papillary thyroid carcinoma	21	399
Adrenocortical carcinoma	22	88
Kidney chromophobe	23	65
Pheochromocytoma and paraganglioma	24	161
Sarcoma	25	240
Testicular germ cell cancer	26	149
Uterine carcinosarcoma	27	56
Uveal melanoma	28	80
Total		**6.639**

3.2.2. Alat

Adapun alat yang digunakan dalam penelitian ini di antaranya terdiri dari perangkat keras dan perangkat lunak dengan rincian sebagai berikut:

3.2.2.1. Perangkat keras

Laptop dengan spesifikasi *processor* AMD Ryzen 5 4500U, 8GB DDR4, 512GB SSD M.2 PCle NVMe, VGA AMD Radeon Graphics

3.2.2.2. Perangkat lunak

Perangkat lunak yang digunakan selama penelitian di antaranya:

a. Sistem operasi *Microsoft Windows 10 Home Single Language Version 10.0.19041 Build 19041*,
b. *Microsoft Word 2019* untuk pembuatan draft laporan,
c. *Microsoft Excel 2019* untuk menampilkan data *.csv*,
d. *R for Windows* versi 4.0.2 sebagai bahasa pemrograman,
e. *R Studio* versi 1.4.1103 sebagai *software* pengolahan data menggunakan bahasa pemrograman R. Di dalam R terdapat *package-package* yang digunakan untuk mempermudah pengerjaan di antaranya sebagai berikut:

- *Package* xgboost versi 1.3.1.1, merupakan *package* buatan Chen & Guestrin pada tahun 2016 yang dapat secara otomatis melakukan berbagai fungsi objektif seperti regresi, klasifikasi, dan pemeringkatan yang merupakan implementasi dari *framework gradien boosting*.

- *Package* caret versi 6.0-86 merupakan kependekan dari *Classification And Regression Training* buatan Kuhn tahun 2020 yang memuat fungsi untuk merampingkan proses pelatihan model untuk masalah klasifikasi dan regresi yang kompleks.

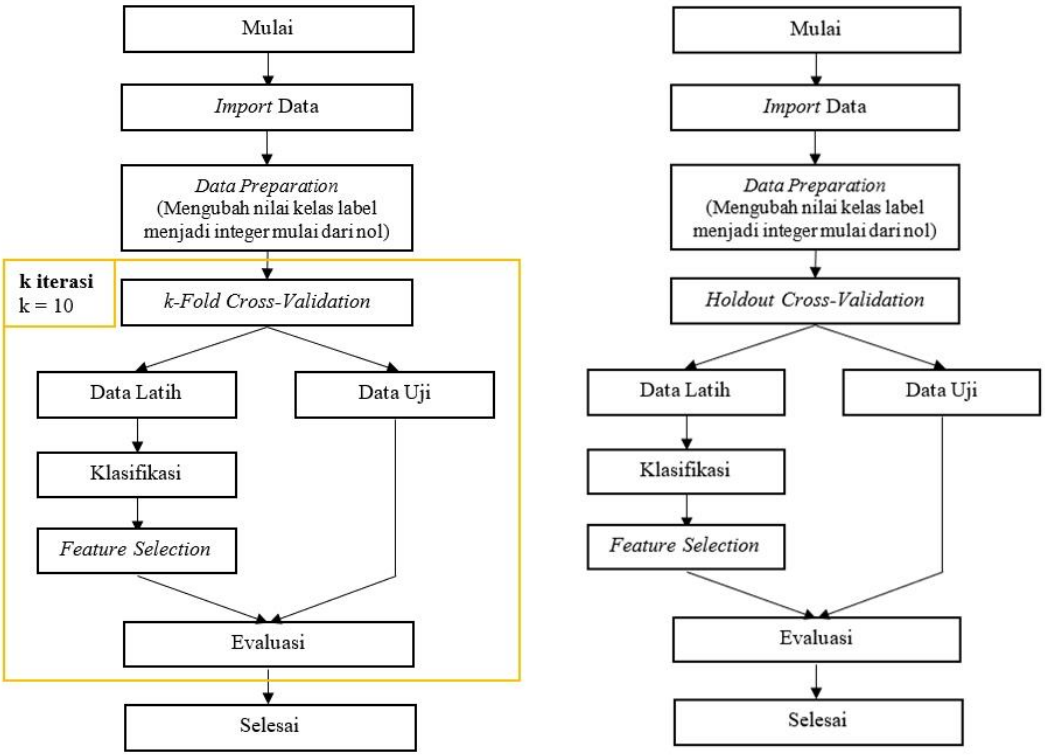
- *Package* e1071 versi 1.7-4 seperti dalam Hornik et al., 2020 berguna untuk analisis kelas, langkah komputasi yang lebih pendek, dan lainnya,

- *Package* dplyr versi 1.0.2 berfungsi sebagai alat yang cepat dan konsisten untuk bekerja menggunakan objek seperti *data frame*, baik di dalam memory maupun di luar memory (Wickham, 2021),

- *Package* Ckmeans.1d.dp versi 4.3.3 yang dibuat oleh (H. Wang & Song, 2011) dengan fungsi menghasilkan histogram yang adaptif dengan pola dalam data. Paket ini menyediakan seperangkat alat canggih untuk analisis data univariat dengan jaminan optimalisasi, efisiensi, dan reproduktivitas.

3.3. Metode

Tahapan metode yang dilakukan dalam penelitian ini digambarkan dalam diagram alur yang ditunjukkan oleh Gambar 12.



Gambar 12. Alur kerja atau tahapan penelitian

Adapun tahapan dari metode pada penelitian ini di antaranya sebagai berikut:

3.3.1. *Import data*

Data yang digunakan dalam penelitian ini didapatkan dari penelitian sebelumnya yaitu penelitian Soh et al., tahun 2017, dimana data tersebut diperoleh dari *cBioPortal for Cancer Genomics* yang terdiri dari mutasi somatik dan perubahan jumlah salinan pada 6640 sampel tumor yang termasuk ke dalam 28 jenis kanker.

3.3.2. *Data Preparation*

Pada tahap ini dari penelitian sebelumnya telah dilakukan penghapusan kolom yang hanya berisi nol karena sesuai dengan gen yang tidak membawa perubahan dalam sampel tumor, penghapusan kolom yang digandakan dari matriks jumlah salinan, dan lain sebagainya. Sampai dihasilkannya matriks mutasi titik somatik 6640×7673 dan matriks jumlah salinan 6640×5477 yang digabungkan menjadi matriks berukuran 6640×13151. Pada matriks tersebut sebanyak 7673 kolom pertama menggambarkan mutasi somatik yang digunakan untuk menilai klasifikasi kanker dengan hanya memakai gen bermutasi titik somatik, sedangkan 5477 kolom terakhir pada matriks gabungan menggambarkan informasi mengenai perubahan nomor salinan gen. Kemudian untuk menerapkan metode XGBoost, kelas label harus diubah tipe datanya terlebih dahulu menjadi angka dan angka dimulai dari nol.

3.3.3. Pembagian Data

Untuk percobaan *10-fold cross-validation,* data diacak kemudian ditetapkan pembagian data uji dan data latih dilakukan dengan persentase 90% sebagai data latih dan sebanyak 10% sisanya digunakan sebagai data testing. Kemudian model dilatih menggunakan lipatan k - 1 (k minus 1) dimana k = 10, ada sebanyak 9 lipatan digunakan untuk melatih dan lipatan yang tersisa dipakai untuk menguji. Proses diulang sebanyak 10 kali sampai setiap lipatan k berfungsi sebagai set uji. Sedangkan untuk *holdout cross-validation,* data diacak kemudian tanpa perulangan sebanyak 90% data digunakan sebagai data latih dan 10% sebagai data uji. Selain itu digunakan juga pembagian data 80:20, 75:25, 70:30, dan 60:40 untuk mengetahui hasilnya.

3.3.4. Klasifikasi

Tahapan klasifikasi pada penelitian ini menggunakan XGBoost yang adalah salah satu implementasi dari algoritma *Gradient Boosted Trees*, sebuah metode *supervised learning* yang termasuk dalam pembelajaran ensemble dan didasarkan pada penggabungan beberapa model dasar untuk menghasilkan satu model prediksi yang optimal.

3.3.5. *Feature Selection*

Penelitian ini menggunakan pemilihan fitur bawaan pada *tree-based model* XGBoost yakni *feature importance* XGBoost untuk mengetahui peringkat kepentingan dari fitur-fitur yang ada sehingga model dapat belajar dengan menggunakan fitur-fitur yang tepat dan berperan penting terhadap hasil klasifikasi atau keluaran. Setelah pemeringkatan kepentingan fitur didapatkan akan dilakukan penyeleksian sebanyak 900 fitur teratas yang kemudian akan digunakan dalam percobaan selanjutnya dengan tujuan untuk mengetahui hasil yang didapatkan agar dibandingkan dengan klasifikasi menggunakan keseluruhan fitur.

3.3.6. Evaluasi Performa

Tahapan terakhir adalah menilai performa *classifier* menggunakan *confusion matrix.* Keakuratan atau *accuracy* menjadi penilaian kinerja *classifier*. Selain mendapatkan nilai akurasi, nilai-nilai lain seperti *precision* dan *recall* juga digunakan untuk mendapat informasi mengenai seberapa baik klasifikasi dilakukan.

V. SIMPULAN DAN SARAN

5.1. Simpulan

Dari hasil penelitian yang sudah dilakukan, dapat diambil simpulan sebagai berikut:

1. Penelitian ini memperoleh hasil klasifikasi dengan mencoba menggunakan metode XGBoost dengan dua teknik pembagian data yaitu *10-fold cross-validation* dan *holdout cross-validation* dengan pembagian 90% data latih 10% data uji memakai data 28 jenis kanker yang digunakan dalam penelitian Soh et al., 2017. Adapun rincian data tersebut memuat informasi sebanyak 6640 sampel tumor dengan jumlah keseluruhan sebanyak 13150 fitur yang terdiri atas 5477 perubahan nomor salinan dan 7673 mutasi somatik. Dalam penelitian ini juga dilakukan klasifikasi memakai 900 fitur teratas hasil dari seleksi fitur menggunakan fungsi seleksi fitur bawaan XGBoost yaitu *xgb.importance* untuk melihat hasil yang didapatkan.

2. Hasil klasifikasi paling baik didapatkan dari percobaan menggunakan XGBoost *holdout cross-validation* memakai 900 fitur teratas hasil seleksi fitur, yakni sebesar 89,91%. Nilai akurasi terbaik kedua didapatkan pada percobaan klasifikasi menggunakan XGBoost *holdout cross-validation* 13150 fitur yaitu 88,86%. Kemudian nilai akurasi terkecil dihasilkan pada percobaan klasifikasi XGBoost *10-fold cross-validation* yaitu sebesar 69,79±1%. Berdasarkan hasil yang diperoleh dari tiap percobaan yang telah dilakukan, dapat dilihat bahwa nilai akurasi pada percobaan XGBoost *holdout cross-validation* lebih tinggi dibandingkan XGBoost *10-fold cross-validation*. Selain itu hasil akurasi dari percobaan yang

dilakukan menggunakan 900 fitur teratas yang diperoleh setelah seleksi fitur mendapatkan hasil yang sedikit lebih tinggi dibandingkan dengan menggunakan keseluruhan fitur. Selain itu, pada percobaan XGBoost *holdout cross-validation* semakin tinggi persentase data latih dibandingkan data uji, nilai akurasi yang didapatkan lebih tinggi.

5.2. Saran

Adapun saran yang dapat diberikan untuk penelitian selanjutnya mengenai Klasifikasi Tipe Kanker Berdasarkan *Signature* Tumor DNA Menggunakan Metode XGBoost adalah sebagai berikut:

1. Mencoba menggunakan metode lain dalam penelitian klasifikasi selanjutnya sehingga hasil klasifikasi dapat diketahui dan dapat dibandingkan dengan hasil dari metode lainnya.

2. Mencoba menggunakan data tipe kanker lain untuk melakukan klasifikasi multikelas menggunakan XGBoost untuk mengetahui kinerja atau performa XGBoost.

DAFTAR PUSTAKA

DAFTAR PUSTAKA

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1). https://doi.org/10.1088/1742-6596/1142/1/012012

American Cancer Society. (2019). Facts & Figures 2019. *American Cancer Society*, 1–76. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf

Austin, C. P. (2021). *Deoxyribonucleic Acid (DNA)*. Diakses pada 20 Maret 2021. https://www.genome.gov/sites/default/files/tg/en/illustration/dna_deoxyribonucleic_acid_adv.jpg

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*(January 2018), 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Biesecker, L. G. (2021). *Ribonucleic acid (RNA)*. Diakses pada 20 Maret 2021. https://www.genome.gov/sites/default/files/tg/en/illustration/rna_ribonucleic_acid.jpg

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., & He, T. (2021). *xgboost: eXtreme Gradient Boosting* (pp. 1–3). https://doi.org/doi:10.1145/2939672.2939785

Cheng, F., Su, L., & Qian, C. (2016). Circulating tumor DNA: A promising biomarker in the liquid biopsy of cancer. In *Oncotarget* (Vol. 7, Issue 30, pp. 48832–48841). https://doi.org/10.18632/oncotarget.9453

Crowley, E., Di Nicolantonio, F., Loupakis, F., & Bardelli, A. (2013). Liquid biopsy: Monitoring cancer-genetics in the blood. *Nature Reviews Clinical Oncology*, *10*(8), 472–484. https://doi.org/10.1038/nrclinonc.2013.110

Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, *7*(3), 1174–1179. www.ijcsit.com

Ferreira, M. M., Ramani, V. C., & Jeffrey, S. S. (2016). Circulating tumor cell technologies. *Molecular Oncology*, *10*(3), 374–394. https://doi.org/https://doi.org/10.1016/j.molonc.2016.01.007

Ghahramani, Z. (2004). *Unsupervised Learning*. 72–112.

Green, E. D. (2021). *Chromosome*. Diakses pada 20 Maret 2021. https://www.genome.gov/sites/default/files/tg/en/illustration/chromosome.jpg

Halili, F., & Rustemi, A. (2016). International Journal of Computer Science and Mobile Computing. *International Journal of Computer Science and Mobile Computing*, *5*(8), 2017–2215.

Hasler, D., & Meister, G. (2016). From tRNA to miRNA: RNA-folding contributes to correct entry into noncoding RNA pathways. *FEBS PRESS*. https://doi.org/10.1002/1873-3468.12294

Hassanpour, S. H., & Dehghani, M. (2017). Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice*, *4*(4), 127–129. https://doi.org/10.1016/j.jcrpr.2017.07.001

Hejmadi, M. (2013). *Introduction to Cancer Biology* (Vol. 2, Issue 4). Ventus Publishing. https://doi.org/10.1517/14656566.2.4.613

Hornik, K., Weingessel, A., Leisch, F., & Davidmeyerr-projectorg, M. D. M. (2020). *Package 'e1071.'*

Ismaeel, A. G., & Mikhail, D. Y. (2016). Effective Data Mining Technique for Classification Cancers via Mutations in Gene using Neural Network.

International Journal of Advanced Computer Science and Applications, *7*(7). https://doi.org/10.14569/ijacsa.2016.070710

Jr, L. A. D., & Bardelli, A. (2014). Liquid Biopsies: Genotyping Circulating Tumor DNA. *JOURNAL OF CLINICAL ONCOLOGY BIOLOGY OF NEOPLASIA*, *32*, 579–587. https://doi.org/10.1200/JCO.2012.45.2011

Kementerian Kesehatan RI Badan Penelitian dan Pengembangan. (2018). Hasil Utama Riset Kesehatan Dasar. *Kementerian Kesehatan Republik Indonesia*, 1–100. http://www.depkes.go.id/resources/download/info-terkini/hasil-riskesdas-2018.pdf

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, *9*(JUL), 1–4. https://doi.org/10.3389/fpsyg.2018.01117

Kuhn, M. (2020). Package ‘ caret ’ R topics documented: In *CRAN Repository* (pp. 1–223).

Lister Hill National Center for Biomedical Communications. (2019). *Help Me Understand Genetics Cells and DNA* (p. 19). https://lhncbc.nlm.nih.gov/LHC-research/research-areas.html

Mathai, R., Vidya, R., Reddy, B., Thomas, L., Udupa, K., Kolesar, J., & Rao, M. (2019). Potential Utility of Liquid Biopsy as a Diagnostic and Prognostic Tool for the Assessment of Solid Tumors: Implications in the Precision Oncology. *Journal of Clinical Medicine*, *8*(3), 373. https://doi.org/10.3390/jcm8030373

Mohammed, M., Khan, M. B., & Bashie, E. B. M. (2016). Machine learning: Algorithms and applications. In *Machine Learning: Algorithms and Applications* (Issue July). https://doi.org/10.1201/9781315371658

Muhammad, I., & Yan, Z. (2015). Supervised Machine Learning Approaches: a Survey. *ICTACT Journal on Soft Computing*, *05*(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Neumann, M. H. D., Bender, S., Krahn, T., & Schlange, T. (2018). ctDNA and CTCs in Liquid Biopsy – Current Status and Where We Need to Progress.

Computational and Structural Biotechnology Journal, *16*, 190–195. https://doi.org/10.1016/j.csbj.2018.05.002

Nordberg, P., Monnet, D. L., Cars, O., Lodato, E. M., & Kaplan, W. (2013). *Priority Medicines for Europe and the World "A Public Health Approach to Innovation" Background Paper 6.1 Antimicrobial resistance*. *April*, 122.

Novakovic, J., Veljovi, A., Iiic, S., Papic, Z., & Tomovic, M. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39–46.

Panawala, L. (2017). *Difference Between Chromosome and Gene Trusted Mutation Content Main Difference – Chromosome vs*. *February*.

Phillippy, A. M., Mason, J. A., Ayanbule, K., Sommer, D. D., Taviani, E., Huq, A., Colwell, R. R., Knight, I. T., & Salzberg, S. L. (2007). Comprehensive DNA signature discovery and validation. *PLoS Computational Biology*, *3*(5), 0887–0894. https://doi.org/10.1371/journal.pcbi.0030098

Powers, D. M. W. (2011). *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. 37–63. http://arxiv.org/abs/2010.16061

Refaeilzadeh, P., Tang, L., & Liu, H. (2008). Cross-Validation. *Preparing for Construction in the 21st Century*, 148–149. https://doi.org/10.1017/s0081130000004822

Sathya, R., & Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*, *2*(2), 34–38. https://doi.org/10.14569/ijarai.2013.020206

Soh, K. P., Szczurek, E., Sakoparnig, T., & Beerenwinkel, N. (2017). Predicting cancer type from tumour DNA signatures. *Genome Medicine*, *9*(1), 1–11. https://doi.org/10.1186/s13073-017-0493-2

Soofi, A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, *13*, 459–465. https://doi.org/10.6000/1927-5129.2017.13.76

Susman, M. (2001). Genes: Definition and Structure. In *ENCYCLOPEDIA OF LIFE SCIENCES* (p. 7). Nature Publishing Group.

Taninaga, J., Nishiyama, Y., Fujibayashi, K., & Gunji, T. (2019). Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data : A case- control study. *Scientific Reports*, *August*, 1–9. https://doi.org/10.1038/s41598-019-48769-y

Travers, A., & Muskhelishvili, G. (2015). DNA structure and function. *FEBS Journal*, *282*(12), 2279–2295. https://doi.org/10.1111/febs.13307

Wang, H., & Song, M. (2011). Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *R Journal*, *3*(2), 29–33. https://doi.org/10.32614/rj-2011-015

Wang, S., Tang, J., & Liu, H. (2016). Encyclopedia of Machine Learning and Data Mining. *Encyclopedia of Machine Learning and Data Mining*, *October 2017*. https://doi.org/10.1007/978-1-4899-7502-7

Wickham, H. (2021). *Package ‘ dplyr .’*

World Health Organization. (2018). International agency for research on cancer. *Asian Pacific Journal of Cancer Prevention*, *4*(1), 3–4.

Wulandari, T., & Muflikhah, L. (2018). Klasifikasi Jenis Kanker Berdasarkan Struktur Protein Menggunakan Algoritma Naive Bayes. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *2*(10), 3738–3743.

Xue, V. W., Wong, C. S. C., & Cho, W. C. S. (2019). Early detection and monitoring of cancer in liquid biopsy: advances and challenges. *Expert Review of Molecular Diagnostics*, *19*(4), 273–276. https://doi.org/10.1080/14737159.2019.1583104

Y, O. F., T, A. J. E., O, A., O., H. J., O, O., & J., A. (2017). Supervised Machine Learning Algorithms Classification and Comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*, 128–138.