KLASIFIKASI *MICROARRAY* PADA SEL KANKER PAYUDARA MENGGUNAKAN METODE *EXTREME GRADIENT BOOSTING*

Skripsi

Oleh

FRISKA DAESY ELVINA SIMBOLON
1617051141

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

ABSTRAK

KLASIFIKASI *MICROARRAY* PADA SEL KANKER PAYUDARA MENGGUNAKAN METODE *EXTREME GRADIENT BOOSTING*

Oleh

FRISKA DAESY ELVINA SIMBOLON

Kanker payudara adalah suatu penyakit yang ditandai dengan pembelahan sel yang tidak terkendali, yang disebabkan oleh perubahan gen (mutasi) sehingga membuat sel secara mendadak bisa berubah. Kanker disebut juga dengan tumor, yang dibagi menjadi dua golongan, yaitu tumor jinak, dan tumor ganas. Kanker payudara menjadi kanker yang paling mematikan di seluruh dunia, terutama yang diderita oleh kaum wanita. Klasifikasi penyakit kanker payudara sangat penting untuk menentukan pengobatan yang tepat dan memberikan perkiraan prognosis yang akurat untuk pasien. Metode yang digunakan dalam penelitian ini adalah *XGBoost*, yang merupakan teknik *ensemble learning* yang cukup populer dalam *machine learning.* Penelitian ini bertujuan untuk mengetahui hasil klasifikasi algoritma *XGBoost* dengan menggunakan data DNA microarray pada sel kanker payudara. yang berasal dari *National Center for Biotechnology Information (NCBI).* Pada tahap awal, model *XGBoost* dilatih dengan menggunakan parameter *default*, kemudian melakukan *hyperparameter* dengan menggunakan metode *GridSearch CV* untuk mencari parameter terbaik yang dapat meningkatkan performa model. Hasil penelitian menunjukkan bahwa pembagian data 10-*fold cross validation* dan telah dilakukan *hyperparameter*, performa model *XGBoost* meningkat signifikan. Akurasi model meningkat dari 50% menjadi 76%, presisi 50% menjadi 86%, *recall* dari 50% meningkat dari 60%, dan *f1-score* menjadi 71%. Hal ini menunjukkan bahwa *hyperparameter* dapat meningkatkan performa model *XGBoost* dalam klasifikasi *microarray* pada sel kanker payudara.

Kata kunci : Klasifikasi, Kanker Payudara, *XGBoost*, *Hyperparameter*

ABSTRACT

CLASSIFICATION OF BREAST CANCER CELL MICROARRAY USING EXTREME GRADIENT BOOSTING METHOD

By

FRISKA DAESY ELVINA SIMBOLON

Breast cancer is a disease characterized by uncontrolled cell division, caused by genetic mutations that can lead to sudden changes in the cells. Cancer, also known as a tumor, is divided into two categories: benign and malignant tumors. Breast cancer is the most fatal cancer worldwide, especially among women. The classification of breast cancer is crucial in determining appropriate treatment and providing accurate prognosis for patients. The method used in this study is XGBoost, which is a popular ensemble learning technique in machine learning. The study aims to investigate the classification results of the XGBoost algorithm using DNA microarray data on breast cancer cells obtained from the National Center for Biotechnology Information (NCBI). Initially, the XGBoost model was trained using default parameters, then hyperparameters were tuned using the GridSearch CV method to find the best parameters that could improve the model's performance. The results showed that after 10-fold cross-validation and hyperparameter tuning, the XGBoost model's performance significantly improved. The accuracy of the model increased from 50% to 76%, precision increased from 50% to 86%, recall increased from 50% to 60%, and the f1-score increased to 71%. This indicates that hyperparameter tuning can improve the XGBoost model's performance in the microarray classification of breast cancer cells.

Keywords : *Classification, Breast Cancer, XGBoost, Hyperparameter*

KLASIFIKASI *MICROARRAY* PADA SEL KANKER PAYUDARA MENGGUNAKAN METODE *EXTREME GRADIENT BOOSTING*

Oleh

FRISKA DAESY ELVINA SIMBOLON

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
2023

Judul Skripsi	: **KLASIFIKASI *MICROARRAY* PADA SEL KANKER PAYUDARA MENGGUNAKAN METODE *EXTREME GRADIENT BOOSTING***
Nama	: Friska Daesy Elvina Simbolon
Nomor Pokok Mahasiswa	: 1617051141
Program Studi	: Ilmu Komputer
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing,

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

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
Bukan Pembimbing : **Prof. Admi Syarif, Ph.D.**

Penguji
Bukan Pembimbing : **Aristoteles, S.Si., M.Si.**

2. Plt. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 19711001 200501 1 002

Tanggal Lulus Ujian Skripsi : **03 Maret 2023**

PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya yang berjudul "**Klasifikasi *Microarray* Pada Sel Kanker Payudara Menggunakan Metode *Extreme Gradient Boosting***" ini merupakan hasil karya sendiri dan bukan hasil karya orang lain. Semua tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi berupa pencabutan gelar akademik yang telah saya terima.

Bandar Lampung, 03 Maret 2023

Friska Daesy Elvina Simbolon
NPM. 1617051141

RIWAYAT HIDUP

Penulis dilahirkan pada 29 Desember 1997 di Kota Bandar Lampung, merupakan putri kedua dari pasangan Bapak R. Simbolon dan Ibu R. Situmorang. Penulis menyelesaikan pendidikan formal pertama kalinya di Taman Kanak-Kanak (TK) Xaverius Panjang Bandar Lampung pada tahun 2004, kemudian melanjutkan Pendidikan dasar di SD Xaverius 2 Bandar Lampung dan selesai pada tahun 2010. Menempuh pendidikan menengah pertama di SMP Xaverius 3 Bandar Lampung yang diselesaikan pada tahun 2013, serta menyelesaikan pendidikan menengah kejuruan dengan Jurusan Kimia Analisis pada tahun 2016 di SMK SMTI Bandar Lampung.

Pada tahun 2016, penulis terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung. Selama menjadi mahasiswa, penulis mengikuti beberapa kegiatan antara lain:

1. Anggota Abacus Himpunan Mahasiswa Jurusan Ilmu Komputer periode 2016/2017.
2. Anggota bidang Keilmuan Himpunan Mahasiswa Jurusan Ilmu Komputer (HIMAKOM) periode 2017.
3. Melaksanakan karya wisata ilmiah di Desa Margosari, Kecamatan Pagelaran Utara, Kabupaten Pringsewu pada bulan Januari 2017.
4. Pada bulan Desember 2018 penulis melaksanakan kerja praktik di PT. Pelindo II Cabang Panjang, Bandar Lampung.
5. Pada bulan Juli 2019 penulis melaksanakan KKN di Desa Sukarendah, Kecamatan Warunggunung, Kabupaten Lebak, Provinsi Banten.

PERSEMBAHAN

Skripsi ini saya persembahkan kepada:

Terima kasih kepada Bapak, yang selalu memberikan kasih sayang,

dukungan, semangat, doa, dan selalu mendidik anak-anaknya.

Terima kasih kepada Mama yang telah berbahagia di surga,

yang telah menjadi alasanku untuk selalu tetap kuat.

Terima kasih kepada Abang dan Kedua Adikku yang selalu

mendukung dan membantu di setiap saat.

Terima kasih kepada Bapak dan Ibu dosen di Jurusan Ilmu Komputer

Yang senantiasa memberikan 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.

MOTTO

“Segala perkara dapat kutanggung di dalam Dia yang memberi kekuatan kepadaku”
(Filipi 4:13)

“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)

“Sebab Aku ini mengetahui rancangan-rancangan apa yang ada pada-Ku mengenai kamu, demikianlah firman TUHAN, yaitu rancangan damai sejahtera dan bukan rancangan kecelakaan, untuk memberikan kepadamu hari depan yang penuh harapan”
(Yeremia 29:11)

“Lupakan semua penyesalanmu, terus melangkah dan jangan takut”
(D.O. Kyung-soo)

“Akan ada saatnya sesuatu itu menjadi sangat melelahkan, tetapi jangan menyerah. Bertahanlah sedikit lagi, dan itu akan segera berakhir”
(Zhang Yixing)

SANWACANA

Puji syukur kepada Tuhan Yesus Kristus, karena atas berkat dan penyertaan-Nya penulis dapat menyelesaikan skripsi yang berjudul "Klasifikasi *Microarray* Pada Sel Kanker Payudara Menggunakan Metode *Extreme Gradient Boosting."* Skripsi ini merupakan salah satu syarat untuk memperoleh gelar sarjana Ilmu Komputer di Universitas Lampung.

Penulis mengucapkan terimakasih kepada semua pihak yang telah membantu dan memiliki peran besar dalam penyusunan skripsi ini, yaitu:

1. Kedua orang tua, Bapak dan Mama, yang telah senantiasa memberikan doa, dukungan, kasih sayang yang sangat luar biasa, dan menjadi alasanku untuk tetap bertahan dan menjalani hidup sampai saat ini.
1. Abang dan kedua adikku yang selalu mendukung dan membantu dalam menyelesaikan skripsi.
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. Bapak Prof. Admi Syarif, Ph.D., selaku dosen penguji pertama atas kesediaannya telah memberikan saran dan masukan guna penyempurnaan penulisan skripsi.
4. Bapak Aristoteles, S.Si., M.Si., selaku dosen penguji kedua skripsi yang telah memberikan saran dan masukan guna penyempurnaan penulisan skripsi.
5. Bapak Didik Kurniawan, S.Si., M.T., selaku Ketua Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.
6. Bapak Dr. rer. nat. Akmal Junaidi, M.Sc., selaku Sekretaris Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

7. Bapak Plt. Dekan Dr. Eng. Heri Satria, S.Si., M.Si., selaku Plt. Dekan FMIPA Universitas Lampung.
8. Bapak dan Ibu Dosen Jurusan Ilmu Komputer, Universitas Lampung yang telah memberikan ilmu dan pengetahuan hidup selama penulis menjadi mahasiswa.
9. Ibu Ade Nora Maela, Pak Zainudin, Mas Sam dan Mas Nofal yang telah memudahkan segala urusan administrasi penulis di Jurusan Ilmu Komputer.
10. Teman-teman di jurusan Ilmu Komputer Angkatan 2016 yang menjadi teman satu Angkatan selama menjalankan masa studi di Universitas Lampung.
11. Teman-teman Grup Kriskat, Bopat, Bodat, Seminar Organizer, dan Naposo PSBI yang telah senantiasa menjadi teman yang menemani keseharian penulis selama masa perkuliahan dengan canda dan tawa.
12. Para Member EXO atas hiburan dan motivasi sebagai inspirasi melalui karya-karyanya saat penulisan skripsi hingga terselesaikan.
13. Semua pihak yang secara langsung maupun tidak langsung yang telah membantu menyelesaikan skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kata sempurna, akan tetapi semoga skripsi ini dapat membawa manfaat dan keberkahan bagi perkembangan ilmu pengetahuan terutama bagi semua civitas Ilmu Komputer, Universitas Lampung.

Bandar Lampung, 03 Maret 2023

Friska Daesy Elvina Simbolon
NPM. 1617051141

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR KODE PROGRAM

Kode Program **Halaman**

I. PENDAHULUAN

1.1 Latar Belakang

Kanker didefinisikan sebagai penyakit yang ditandai dengan pembelahan sel yang tidak terkendali, disebabkan oleh perubahan gen (mutasi) sehingga membuat sel secara mendadak bisa berubah. Sel-sel kanker memiliki kemampuan membelah dengan kecepatan yang berpuluh-puluh kali lipat dari sel normal, serta dapat menyerang pertumbuhan biologis lainnya, baik secara pertumbuhan langsung di jaringan yang bersebelahan (invasi) serta dengan perpindahan sel ke tempat yang jauh (metastasis). Perkembangan yang tidak terkendali dari sel-sel kanker yang dapat menyebabkan kerusakan DNA serta mutasi lanjut di gen vital yang mengontrol pembelahan sel. Kemampuan (mutasi) inilah yang akhirnya mengakibatkan sel-sel normal menjadi sel-sel kanker (Bustan, 2007). Kanker atau sel tumor ganas menyebar ke jaringan atau bagian tubuh lainnya, yang menyebabkan kematian.

Sebagian dari masyarakat menyebut bahwa kanker sebagai tumor. Tumor merupakan sebuah benjolan yang tidak normal atau abnormal. Tumor dibagi menjadi dua golongan, yaitu tumor jinak *(benign)* dan tumor ganas *(malignant)* (Handayani, et al., 2017). Pada tahun 2022, sekitar 8,2 juta kematian yang disebabkan oleh penyakit kanker serta menjadi salah satu penyebab angka kematian tertinggi di berbagai dunia setelah penyakit jantung. Tren penyakit ini akan meningkat dua kali lipat bahkan lebih mematikan pada tahun 2030. Setiap tahun kasus dan kematian karena kanker akan meningkat satu persen (Sholihin, 2017)

Menurut data *Global Burden Cancer* (Globocan, 2020) kanker terbesar yang banyak menimbulkan korban tiap tahunnya adalah kanker payudara, kanker paru-paru, kanker usus besar, kanker prostat, dan lain-lain. Tercatat 19,3 juta kasus baru kanker yang diderita semua kaum pada tahun 2022. Angka ini meningkat

dibandingkan pada tahun 2018 sebanyak 18,1 juta kasus baru yang diagnosis kanker dengan jumlah angka kematian sebesar 9,6 juta jiwa. Di Indonesia, kasus kematian akibat kanker pada tahun 2022 mencapai 235 ribu jiwa dari 397 ribu kasus baru. Penyakit kanker yang terbanyak di dunia bahkan Indonesia yang diderita pada perempuan adalah kanker payudara dan kanker serviks, sedangkan kanker paru-paru dan kanker prostat pada laki-laki.

Karsinoma merupakan kanker yang awalnya terjadi sel epitel *(squamosa).* Karsinoma disebut juga sebagai kanker kulit dikarenakan banyak terjadi di kulit. Kanker payudara adalah salah satu kanker yang termasuk ke dalam jenis karsinoma dikarenakan terjadinya mutasi sel epitel yang berbentuk silindris dan jaringan payudara (Sholihin, 2017).Kanker payudara umumnya terjadi pada wanita, akan tetapi kanker ini bisa terjadi pada pria. Gejala awal pada kanker payudara sebagian dari penderita tidak menyadari hingga memasuki stadium lanjut. Akibat dari terlambatnya penanganan penderita kanker payudara yang sangat fatal dan menyebabkan kematian, sehingga perlu dilakukan diagnosa kanker sejak dini agar penyakit kanker payudara dapat ditangani dengan baik.

Saat ini, perkembangan teknologi informasi merupakan suatu hal yang tidak dapat dihindari. Seiring dengan perkembangan teknologi informasi telah muncul berbagai kajian ilmu lain salah satunya bioinformatika. Bioinformatika adalah aplikasi dari alat komputasi dan analisis yang bertujuan untuk menyelesaikan masalah-masalah atau data-data biologi dengan menggunakan DNA dan asam amino, serta informasi-informasi yang terkait didalamnya (Bayat, 2002).

Hal ini yang mendukung dengan adanya perkembangan teknologi bernama *microarray*. *Microarray* adalah teknologi yang mempunyai kemampuan untuk memantau ribuan ekspresi gen dalam jumlah yang sangat besar dan berbeda secara bersamaan dalam satu percobaan. Data microarray dapat digunakan untuk mendeteksi dan mengklasifikasikan suatu jaringan penyakit di manusia (Trevino, et al., 2007). Teknologi ini dapat membantu peneliti dalam mempelajari berbagai penyakit, dan sangat berpengaruh besar dalam menentukan gen yang menjadi penyebab kanker (Diani, et al., 2017). Dalam bidang kedokteran, teknologi banyak

digunakan untuk proses pendekatan statistik terhadap kanker yang cepat dan efisien. Teknik komputasi di dalam *machine learning* dapat digunakan untuk menganalisis pada pemilihan gen atau protein yang mempunyai sifat yang serupa dan mengklasifikasikan tipe dari sampel ekspresi gen pada data *microarray* (Yang, et al., 2010). *Machine learning* juga dapat membantu membantu menyelesaikan permasalahan pada data dengan dimensi tinggi seperti data *microarray*.

Permasalahan dalam *machine learning* dalam diagnosa kanker menggunakan data *microarray* salah satunya yaitu klasifikasi. Klasifikasi adalah dapat mengklasifikasikan apakah pasien terkena kanker atau tidak melalui pola ekspresi gen dari sejumlah pasien yang telah mengidap kanker dari data *microarray*. Terdapat berbagai macam metode klasifikasi pada *machine learning* salah satunya adalah *Extreme Gradient Boosting (XGBoost)*. *XGBoost* adalah algoritma regresi dan klasifikasi metode *ensemble* yang merupakan suatu varian dari algoritma *Tree Gradient Boosting*, serta telah dikembangkan dengan optimasi 10 kali lebih cepat dibandingkan *Gradient Boosting* lainnya (Chen & Guestrin, 2016).

Contoh, penelitian pada jurnal yang berjudul Pengembangan Perangkat Lunak Prediktor Kanker Payudara Menggunakan Metode *Elastic SCAD SVM* dan Data *Microarray* (Firmansyah, et al., 2012). Penelitian tersebut dilakukan untuk mendeteksi kanker payudara menggunakan metode *elastic SVM* dan data dna *microarray*, mendapatkan nilai akurasi untuk *SVM* sebesar 94,25%, sedangkan untuk *scad SVM* mendapatkan nilai akurasi yang lebih baik 1.15% bernilai 95,4%. Penelitian lainnya, berjudul Komparasi Kinerja Algoritma *XGBoost* dan Algoritma *Support Vector Machine (SVM)* Untuk Diagnosa Penyakit Kanker Payudara yang dilakukan pada (Ravly, 2022) Penelitian tersebut menggunakan perbandingan metode antara *SVM* dan *XGBoost* untuk klasifikasi penyakit kanker payudara didapatkan nilai akurasi *XGBoost* memiliki akurasi terbaik sebesar 96,95% dengan nilai AUC sebesar 0.99, sedangkan *SVM* memiliki nilai akurasi terendah bernilai 90,24% dengan nilai AUC 0,98.

Berdasarkan uraian latar belakang tersebut, maka diperlukan sesuatu untuk mengklasifikasi penyakit kanker payudara dengan menggunakan metode *XGBoost.*.

Klasifikasi ini menggunakan data yang berupa data *microarray*, yang kemudian akan dilakukan optimasi dengan *hyperparameter* untuk melihat tingkat keakuratan dan kompleksitas komputasi. Penelitian ini diharapkan dapat membantu manusia, terkhususnya kepada bidang kesehatan, untuk melakukan prediksi penyakit kanker payudara secara cepat dan akurat.

1.2 Rumusan Masalah

Berdasarkan latar belakang, maka rumusan masalah penelitian ini adalah bagaimana hasil kinerja dari metode *extreme gradient boosting (XGBoost)* dalam mengklasifikasikan penyakit kanker payudara menggunakan data *microarray*.

1.3 Batasan Masalah

Batasan masalah pada penelitian ini, antara lain:

1. Penelitian ini akan dilakukan dengan metode *extreme gradient boosting (XGBoost)* untuk klasifikasi penyakit kanker payudara.
2. Data yang digunakan adalah data *microarray* kanker payudara berjumlah 42 pasien. Data ini bersumber pada *National Center for Biotechnology Information (NCBI)* yang diakses pada website https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283953.
3. Hasil pengujian kinerja model yang akan dibuat akan ditunjukkan melalui beberapa parameter, antara lain *accuracy, precision, recall,* dan *f1 score*.

1.4 Tujuan

Tujuan dari penelitian ini, antara lain:

1. Mengukur dan mengevaluasi kinerja metode *extreme gradient boosting (XGBoost)* dalam klasifikasi *microarray* pada sel kanker payudara.
2. Mengetahui *hyperparameter* yang optimal pada metode *extreme gradient boosting* untuk menghasilkan model dengan kinerja yang maksimal.

1.5 Manfaat

Manfaat dari penelitian ini, antara lain:

1. Dapat mendeteksi penyakit kanker payudara lebih dini dan akurat.

2. Dapat mengetahui hasil dan evaluasi kinerja klasifikasi pada penyakit kanker payudara menggunakan metode *extreme gradient boosting (XGBoost).*
3. Dapat menemukan *hyperparameter* yang optimal untuk mengukur kinerja algoritma *extreme gradient boosting (XGBoost)* pada klasifikasi *microarray* pada sel kanker payudara.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Terdapat beberapa penelitian terdahulu yang berhubungan dengan penelitian ini, ditampilkan pada Tabel 1.

Tabel 1. Penelitian Terdahulu Terkait Penelitian Ini.

No	Penelitian	Metode	Data	Hasil
1.	Komparasi Kinerja Algoritma *XGBoost* dan Algoritma *Support Vektor Machine (SVM)* Untuk Diagnosa Penyakit Kanker Payudara (Andryan, et al., 2022)	• *Support Vektor Machine (SVM)* • *Extreme Gradient Boosting (XGBoost)*	*UCI Machine Learning Breast Cancer Diagnostic* (569 *record* dengan 30 atribut)	*SVM* • Akurasi : 97.58% • Presisi : 93.67% • *Recall* : 91.34% *XGBoost* • Akurasi : 96.26% • Presisi : 96.95% • *Recall* : 93.57%
2.	Prediksi Penyakit Jantung *Cardiovascular* Menggunakan Model Algoritma Klasifikasi (Nugraha, 2021)	• *Random Forest* • *Support Vector Machine* • *Gradient Boosting Machine* • *XGBoost* • *Light GBM*	*Kaggle Datasets* (299 pasien dengan 12 fitur)	*RF*: • Akurasi : 77% • *f1-score* : 83% • *AUC* : 0.72 *SVM*: • Akurasi : 80% • *f1-score* : 86% • *AUC* : 0.73 *GBM*: • Akurasi :78% • *f1-score*: 85% • *AUC* : 0.73 *XGBoost*: • Akurasi : 80% • *f1-score* : 86% • *AUC* : 0.75 *Light GBM*: • Akurasi : 68% • *f1-score*: 81% • *AUC* : 0.50

No	Penelitian	Metode	Data	Hasil
3.	*Hyperparameter Tuning* pada Algoritma Klasifikasi dengan *GridSearch* (Nugraha & Sasongko, 2022)	• *Support Vector Machine (SVM)* • *Random Forest* • *Logistic Regression* • *Gaussian Naïve Baye* • *Decision Tree* • *Extreme Gradient Boosting (XGBoost)* • *K-Nearest Neighbors (k-NN)*	*UCI Pima Indians Diabetes Database* (768 sampel dengan 9 atribut)	*SVM* : 0,763 *RF* : 0,771 *Logistic Regression* : 0,769 *Gaussian Naive Bayes* : 0,7699 *Decision Tree* : 0,701 *XGBoost* : 0,772 *KNN* : 0,763
4.	*On Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier* (Nuklianggraita, et al., 2020)	• *Least Absolute Shrinkage and Selection Operator (LASSO)* • *Minimum Redundancy Maximum Relevance (MRMR)* • *Random Forest*	5 data terdiri dari • *Central Nervous System* (60 *record* dan 7129 fitur) • *Colon Cancer* (62 *record* dan 2000 fitur) • *Lung Cancer* (181 *record* dan 12533 fitur) • *Prostate Tumor* (136 *record* dan 12600 fitur)	*1. Without Dimension Reduction:* *Central Nervous System :* • Akurasi : 75% • Presisi : 0% • *Recall* : 0% *Colon Cancer :* • Akurasi : 84% • Presisi : 60% • *Recall* : 100% *Lung Cancer :* • Akurasi : 100% • Presisi : 100% • *Recall* : 100% *Prostate Tumor* • Akurasi : 85.71% • Presisi : 94% • *Recall* : 85% *Ovarian Cancer* • Akurasi : 100% • Presisi : 100% • *Recall* : 100%

No	Penelitian	Metode	Data	Hasil
			• *Ovarian Cancer* (253 *record* dan 15154)	2. *LASSO + Random Forest :* *Central Nervous System :* • Akurasi : 83% • Presisi : 33% • *Recall* : *Colon Cancer :* • Akurasi : 92% • Presisi : 80% • *Recall* : *Lung Cancer :* • Akurasi : 100% • Presisi : 100% • *Recall* : *Prostate Tumor* • Akurasi : 93% • Presisi : 90% • *Recall* : *Ovarian Cancer* • Akurasi : 100% • Presisi : 100% • *Recall* : 3. *MRMR + Random Forest :* *Central Nervous System :* • Akurasi : 75% • Presisi : 0% • *Recall* : 0% *Colon Cancer* • Akurasi : 69% • Presisi : 20% • *Recall* : 100% *Lung Cancer :* • Akurasi : 84% • Presisi : 14% • *Recall* : 100%

No	Penelitian	Metode	Data	Hasil
				Prostate Tumor : • Akurasi : 64% • Presisi : 0% • *Recall* : 0% *Ovarian Cancer : 67* • Akurasi : 67% • Presisi : 19% • *Recall* : 100%

Berdasarkan Tabel 1, menjelaskan bahwa penelitian ini menggunakan metode *extreme gradient boosting (XGBoost)*. *Tree boosting* adalah sebuah mesin yang sangat efektif dan banyak digunakan dalam *machine learning*. *XGBoost* disebut juga dengan istilah *a scalable end to end tree boosting system*. *XGBoost* meningkatkan skala yang melampaui miliaran contoh yang menggunakan sumber daya jauh lebih sedikit daripada sistem yang ada, dan memberikan wawasan tentang pola akses *cache*, data *compression* dan *sharding* untuk membangun *scalable tree boosting system.*

Penelitian pertama dilakukan oleh Andryan, et al. (2022). Penelitian ini bertujuan untuk mengklasifikasikan kanker payudara menggunakan metode *extreme gradient boosting* dan *support vector machine.* Data berasal dari *Kaggle dataset Wisconsin Breast Cancer Diagnostic* yang berisikan 569 sampel dengan jumlah atribut sebanyak 30, serta digolongkan kedalam 2 kelas yaitu kelas *malignant* dan *benign*. Data dibagi menjadi 80% data latih, dan 20% data uji, kemudian menggunakan *GridSearchCV* untuk melakukan *hyperparameter* untuk mendapatkan model terbaik dari semua kombinasi yang ada seperti *C* (0.1, 10, 100), *gamma* (1.0, 0.001, 0.001), *kernel (rbf, poly, sigmoid)* pada metode *SVM*, dan *max_depth* (1,2,3), *n_estimator* (100, 150, 200, 250), *learning_rate* (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9) pada metode *XGBoost*. Dari kombinasi diatas, diperoleh hasil dari masing-masing metode yaitu metode *SVM* dengan akurasi 94.28%, presisi 93.67%, *recall* 91.34%, dan *rou_auc* 97.58%. Metode *XGBoost* dengan akurasi 96.26%, presisi 96.95%, *recall* 93.57%, dan *roc_auc* bernilai 98.64%

Penelitian kedua dilakukan oleh Nugraha (2021). Penelitian ini mengklasifikasikan beberapa model untuk memprediksi penyakit *cardiovascular* seperti *random forest, SVM, gradient boosting machine, XGBoost,* dan *light GBM*. Data yang digunakan pada penelitian ini berasal dari *Kaggle Datasets Heart Failure Prediction* yang terdiri dari 299 sampel dengan 12 fitur. Dari model yang telah dibangun selanjutnya akan dilakukan perbandingan hasil menggunakan *confusion matrix*, dan menentukan model yang terbaik. Dari hasil eksperimen menunjukkan bahwa model yang telah diujikan model *SVM* dan *XGBoost* memperoleh nilai akurasi sebesar 0.80*, gradient boosting* sebesar 0.78, *random forest* 0.77, dan *light GBM* bernilai 0.68.*f1-score* masing-masing algoritma memperoleh nilai sebesar 0.86 untuk model *SVM* dan *XGBoost*, 0.85 untuk model *gradient boosting*, 0.83 untuk model *random forest*, dan 0.81 untuk model prediksi *light GMB*. Sedangkan untuk pengukuran model menggunakan *area under curve (AUC)* menunjukkan bahwa *XGBoost* memperoleh nilai sebesar 0.75, *SVM* dan *gradient boosting* bernilai 0.73, model *random forest* sebesar 0.72, dan *LightGBM* memperoleh nilai sebesar 0.50.

Penelitian selanjutnya yang dilakukan oleh Nugraha & Sasongko (2022). Penelitian ini menggunakan *hyperparameter gridsearchcv* menggunakan metode *SVM, random forest, logistic, naïve bayes, decision tree, XGBoost,* dan *KNN*. Data pada penelitian berasal dari *Kaggle dataset repository (UCI Pima Indians Diabetes Database)* yang berisikan 9 atribut dengan 768 sampel. Hasil dari penelitian didapatkan hasil berbagai metode, yaitu *SVM* sebesar 0.763, *random forest* bernilai 0.771, 0.769 untuk model *logistic regression*, *gaussian naïve bayes* 0,7699, *decision tree* sebesar 0,701, *KNN* 0, 763, dan *XGBoost* mendapatkan nilai akurasi terbaik daripada metode lainnya sebesar 0,772.

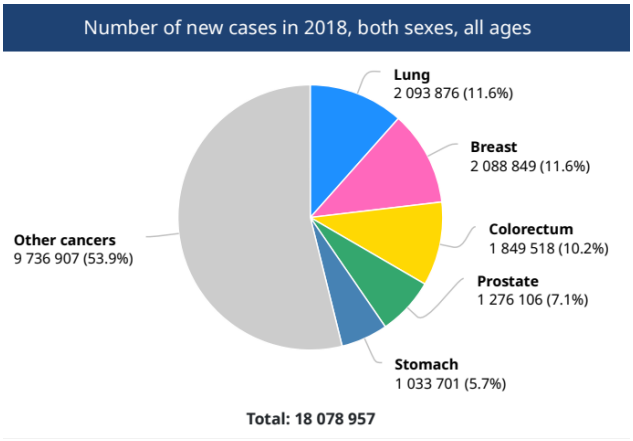
Penelitian terakhir dilakukan oleh Nuklianggraita, et al. (2020). Penelitian ini bertujuan untuk mengklasifikasi data *microarray* dari deteksi kanker menggunakan *random forest classifier*. Dataset yang digunakan ada lima data yang terdiri dari *central nervous system* (60 *record*, dan 7129 fitur), *colon cancer* (62 *record*, dan 2000 fitur), *lung cancer* (181 *record*, dan 12533 fitur), *prostate tumor* (136 *record*, dan 2600 fitur), dan *ovarian cancer* (253 *record*, dan 15154 fitur) yang berasal dari website http://leo.ugr.es/elvira/DBCRepository. Penelitian ini menggunakan *least*

absolute shrinkage and selection operator (lasso) dan *minimum redundancy maximum relevance (mrmr)* dengan pengaplikasian metode *random forest*. Hasil dari penelitian ini didapatkan *without dimension reduction* untuk nilai akurasi dari masing-masing data *central nervous system, colon cancer, lung cancer, prostate tumor,* dan *ovarian cancer* sebesar 75%, 84%, 100%, 85.71%, dan 100%. Hasil untuk *lasso* dengan *random forest* didapatkan hasil nilai akurasi 83% untuk *central nervous system*, 92% *colon cancer,* 100% *lung cancer, prostate tumor sebesar* 93%, dan 100% bernilai *ovarian cancer*. Sedangkan untuk *mrmr* dengan *random forest* nilai akurasi secara berturut 75%, 69%, 84%, 64%, dan 67%.

2.2 Kanker Payudara

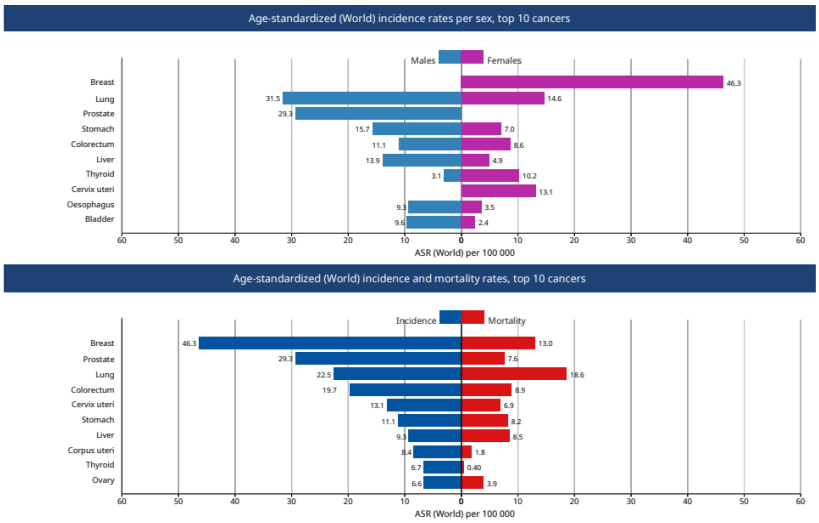
Setiap orang memiliki payudara, laki-laki maupun perempuan. Pada payudara laki-laki mengalami rudimenter yang tidak berfungsi, sedangkan payudara perempuan mengalami pertumbuhan, serta sangat berfungsi dalam hal reproduksi maupun kecantikan. (Bustan, 2007). Pada payudara terdiri atas jaringan kelenjar susu, jaringan lemak, serta jaringan ikat. Selama masa kehamilan, kelenjar susu pada payudara perempuan akan memproduksi dan mengeluarkan susu sebagai makanan untuk bayi. Akan tetapi, jika sel-sel di dalam kelenjar susu membelah diri dan akan berkembang secara tidak terkendali akan menjadi tumor jinak maupun ganas di dalam payudara perempuan.

Setiap risiko kanker pada perempuan memiliki probabilitas yang lebih tinggi maupun lebih rendah, tergantung pada beberapa faktor yang meliputi, riwayat keluarga, genetik, usia saat menstruasi pertama, serta faktor lainnya. Ketika perempuan dengan usia muda terkena kanker payudara, maka terdapat kecenderungan perkembangan kanker menjadi lebih agresif dibandingkan usia yang lebih tua. (Rasjidi, 2009). Gejala awal dari penyakit kanker payudara yaitu, munculnya benjolan asing di daerah payudara yang melekat pada kulit jika diraba, adanya perubahan warna kulit payudara, serta puting susu mengkerut didalam. Jika telah memasuki stadium lanjut, maka puting susu akan mengeluarkan cairan berbau busuk yang menyerupai nanah (Sholihin, 2017).



Gambar 1. Jumlah Kasus Baru pada Penduduk Sedunia tahun 2018 (Globocan, 2018).

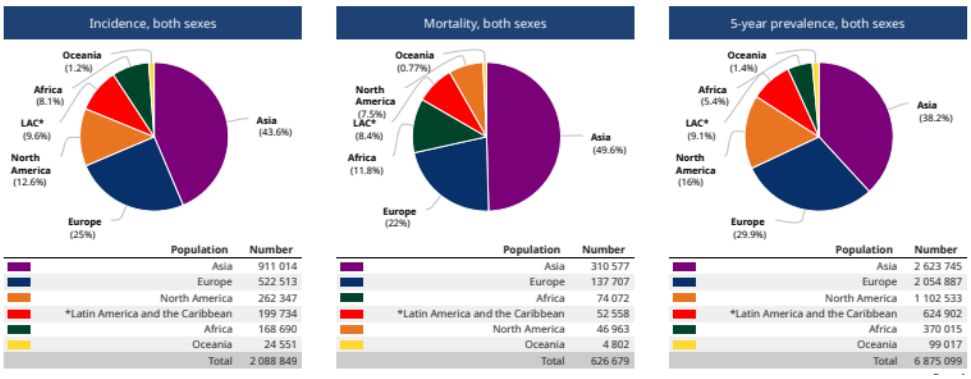
Pada Gambar 1, menunjukkan bahwa jumlah kasus baru kanker pada penduduk sedunia tahun 2018 terdapat 18.078.957 kasus baru, yang terdiri dari kanker paru-paru sebanyak 11.6%, kanker payudara 11.6%, kanker kolorektal 10.2%, kanker prostat 7.1%, dan kanker perut sebanyak 5.7% serta kanker yang lainnya 53.9%.



Gambar 2. Jumlah Insiden dan Kematian Akibat Kanker Pada Penduduk Sedunia (Globocan, 2018).

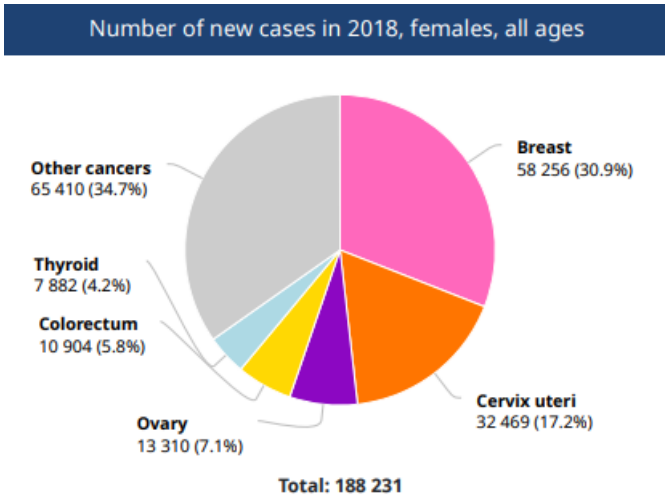
Dalam Gambar 2, menjelaskan bahwa jumlah insiden penyakit kanker berdasarkan jenis kelamin, yaitu kanker payudara memiliki persentase paling tertinggi daripada kanker yang lainnya sebesar 46.3% yang menyerang kaum perempuan di dunia, sedangkan untuk kaum pria sebanyak 31.5% untuk kanker paru-paru. Kanker

payudara menempati urutan pertama sebagai kematian keseluruhannya dengan persentase 46.3%, dan menjadi kanker yang mematikan di dunia yang sering diderita pada kaum wanita.



Gambar 3. Data Insiden, Kematian, dan Prevalensi Kanker Payudara Per Benua (Globocan, 2018).

Gambar 3, mengklasifikasikan data insiden, kematian, dan prevalensi kanker payudara per benua menyatakan bahwa penduduk terutama pada kaum perempuan di Benua Asia menempati kanker payudara tertinggi daripada yang lainnya.



Gambar 4. Jumlah Kasus Baru Kanker di Indonesia Pada Perempuan pada tahun 2018 (Globacan, 2019).

Pada Gambar 4, menunjukkan bahwa kanker yang dialami pada perempuan yang tertinggi di Indonesia adalah kanker payudara sebesar 30.9%, kemudian sebesar 17.2%, kanker ovarium sebanyak 7.1%.

2.3 Ekspresi Gen

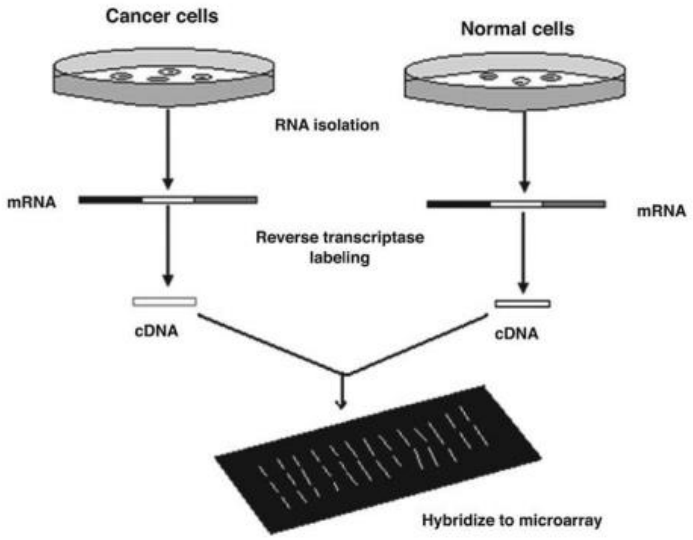
Deoxyribo Nucleic Acid (DNA) merupakan sebuah molekul atau asam nukleat yang membawa informasi genetik dari generasi ke generasi selanjutnya, dan sangat stabil serta dapat bertahan lama. DNA dapat ditemukan pada semua sel biologis yang ber nukleat atau tidak ber nukleat tetapi memiliki plastida dan mitokondria. Menurut Pereira, et al., 2008 dalam (Kalqutny, et al., 2020) menyatakan bahwa *deoxyribonucleic acid (DNA)* dapat memberikan informasi yang lebih banyak dibandingkan protein dikarenakan adanya degenerasi kode genetik (urutan basa *nukleotida* yang berbeda dapat menghasilkan polipeptida yang sama serta adanya untaian *non coding* seperti *adenine* (A), *guanine* (G), *sitosin* (C), dan *timin* (T). Suatu sel mengandung beberapa jenis RNA seperti *messenger* RNA (mRNA), RNA transfer (tRNA), dan RNA ribosom (rRNA) (Bolstad, 2004).

Ekspresi gen di dalam sel memerlukan dua proses yaitu transkripsi dan translasi. Menurut (A. Kumar, 2010) tingkat utama dari proses ekspresi gen adalah transkripsi yaitu proses sintesis mRNA dan translasi mengarah ke sintesis protein, molekul fungsional gen. *Proteome* adalah semua pelengkap protein dalam suatu organisme, sedangkan *proteomic* yaitu sampel biologis untuk kadar protein. Membandingkan tingkat mRNA dari seluruh set gen dalam sampel biologis dengan cDNA yang berlabel pewarna *fluorescent* (Cy3 dan Cy5) yang dilakukan oleh ekspresi gen *microarray*. Melalui *microarray* ini akan dapat digunakan untuk mengetahui gen mana yang diekspresikan dengan level serta kondisi tertentu.

2.4 Data *Microarray*

Gonzalo & Sánchez, 2018 menjelaskan bahwa *microarray* adalah sebuah mesin yang dikembangkan pada akhir 1990-an untuk mengukur ekspresi gen. *Microarray* berbentuk seperti suatu *chip* atau *slide* mikroskop yang berisi rangkaian sampel jaringan, protein, RNA, dan DNA (Iiris Hovatta, et al., 2014). *Microarray* merupakan hasil dari sebuah kombinasi bidang teknologi dan ilmu pengetahuan yang dikembangkan dari beberapa bagian penelitian seperti seperti mekanik, pembuatan mikro, kimia, perilaku DNA, mikrofluida, enzim, optik, dan bioinformatika (Dufva, 2009). *Microarray* merupakan sebuah teknologi yang digunakan untuk menyimpan ribuan bahkan jutaan ekspresi gen *(prone)* yang

diambil dalam beberapa sel sekaligus dalam sekali percobaan. Teknologi ini dapat membantu peneliti dalam mempelajari gen yang menjadi penyebab suatu penyakit (Diani, et al., 2017). Untuk mendapatkan data *microarray*, dapat dilakukan dengan ditunjukkan pada Gambar 5.

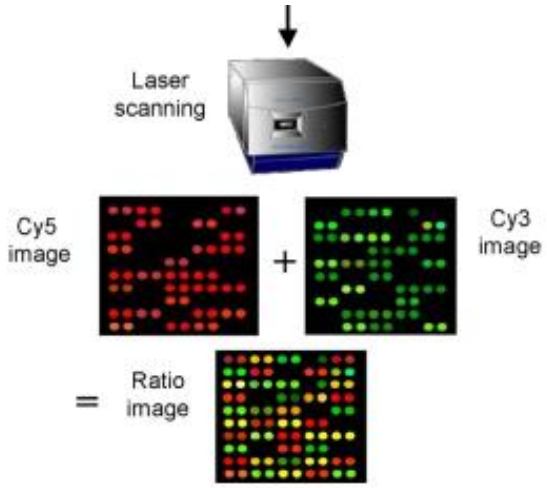


Gambar 5. Tahapan *Microarray* (Bhatia & Dahiya, 2015).

Pada Gambar 5, terlihat tahapan *microarray* untuk mendapatkan nilai ekspresi gen sebagai berikut:

1. Mendapatkan dua sampel mRNA dari jaringan pada sampel dalam dua kondisi yang berbeda yaitu sel pasien terkena kanker dan sel pasien normal.
2. Mengkonversi mRNA menjadi cDNA menggunakan enzim *reverse transcriptase.*
3. Dua sampel diberi label menggunakan dua pewarna *fluorescent* yang berbeda. Warna hijau dengan pewarna Cy3 untuk sel kanker, dan warna merah menggunakan pewarna Cy5 untuk sel normal.
4. Sampel mengalami hibridisasi, yaitu cDNA saling mengikat terhadap DNA, kemudian *microarray* disinari laser dan hasil warna setiap spot diukur.
5. Terdapat empat reaksi spot warna yaitu pada *chip microarray* yaitu spot yang bersinar merah terang mengekspresikan kondisi sel kanker. Spot hijau mengartikan bahwa kondisi sel normal, spot kuning gen dinyatakan dalam kedua kondisi sampel (kanker dan normal), dan spot hitam tidak ada ekspresi gen dalam kondisi baik.

Untuk melakukan analisis lebih lanjut titik-titik warna yang akan diinterpretasikan ke dalam bentuk nilai angka seperti pada Gambar 6.



Gambar 6. *Spot Microarray* (Lamartine, 2006).

Langkah analisis yang dilakukan pada Gambar 6, dengan mengkonversi setiap spot ke dalam bentuk angka berdasarkan perbandingan rasio intensitas warna merah dengan warna hijau dengan Persamaan (1).

$$Rasio = \frac{intensitas\ warna\ merah}{intensitas\ warna\ hijau} \quad (1)$$

Apabila nilai rasio lebih dari satu, maka diidentifikasikan bahwa gen tersebut tersimulasi untuk memproduksi lebih banyak mRNA akibat sel kanker, tetapi jika nilai rasio kurang dari satu, mengindikasikan bahwa gen tersebut tersimulasi untuk memproduksi lebih sedikit mRNA akibat sel kanker, dan apabila nilai rasio sama dengan satu maka tidak terjadi apa-apa.

2.5 *Machine Learning*

Machine learning adalah bidang ilmu yang mempelajari pola maupun teori pembelajaran komputasi dalam *artificial intelligence*. *Machine learning* merupakan proses pembelajaran, serta membangun algoritma yang dapat belajar dan membuat prediksi pada dataset (Simon, et al., 2016). *Machine learning* sangat cerdas dikarenakan mampu memiliki kemampuan generalisasi terhadap data bar yang belum dipelajari sebelumnya misal untuk memprediksi, mengklasifikasi, rangking, dan lain-lain (Abdillah, et al., 2016).

Machine learning mengenali pola dari caranya mempelajari kasus dari data yang telah ada untuk dianalisa dan menggunakan pola yang ada untuk memecahkan masalah dari suatu pertanyaan-pertanyaan. Program yang dibuat menggunakan metode *machine learning* diantaranya: *Online Fraud Detection*, Diagnosa Medis, Mobil Kendali Otomatis, Segmentasi Pelanggan dan Mesin Pencarian. *Machine learning* dibagi menjadi dua jenis, yaitu *supervised learning* dan *unsupervised learning*.

2.5.1 *Supervised Learning*

Supervised learning adalah suatu sistem dilatih dengan data yang telah diberi label, kemudian label tersebut mengkategorikan setiap titik data dalam satu atau beberapa kelompok. Sistem mempelajari bagaimana data tersebut dikenal sebagai data *training* terstruktur, dan menggunakan data training tersebut untuk memprediksi data test atau data uji. Kategori algoritma dalam *supervised learning* diantaranya adalah *regression* dan *classification.*

2.5.2 *Unsupervised Learning*

Unsupervised learning yaitu pembelajaran tanpa pengawasan artinya pembelajaran tanpa label. Hal ini bertujuan untuk mendeteksi karakteristik yang membuat titik suatu data kurang lebih mirip satu sama lain, misalnya dengan membuat cluster dan menetapkan data pada *cluster* tersebut. Kategori algoritma dalam *unsupervised learning* yaitu *clustering.*

2.6 *Extreme Gradient Boosting (XGBoost)*

Penelitian yang dilakukan (Chen & Guestrin, 2016) yang berjudul *XGBoost*: *A Scalable Tree Boosting System* merupakan sebuah *open source* untuk menerapkan sistem pembelajaran mesin yang efisien, cepat dan terukur. Algoritma ini merupakan kombinasi antara algoritma *Gradient Descent* dan *Boosting* yang disebut dengan istilah *Gradient Boosting Machine*. *XGBoost* digunakan untuk permasalahan *supervised learning* yang menggunakan data latih dengan beberapa fitur untuk memprediksi variabel. *Extreme Gradient Boosting* dikembangkan pada sebuah *Higgs Boson Competition,* dimana pada kompetisi ini, algoritma *XGBoost* menjadi algoritma yang paling banyak digunakan dikarenakan dapat melakukan

optimasi 10 kali lebih cepat dibandingkan dengan implementasi dari *Gradient Boosting Machine* lainnya.

Tugas melatih model adalah menemukan parameter terbaik yang paling sesuai dengan data latih dan label, dan untuk mengukur seberapa cocok model tersebut dengan data latih. Langkah-langkah dalam menerapkan algoritma *XGBoost* sebagai berikut:

1. Menentukan nilai *probability* awal dari data target, *class value* yaitu jumlah data yang akan diolah dapat dilihat pada Persamaan (2).

$$Probability(p) = \frac{\Sigma(Class\ Value)}{\Sigma(Class)} \quad (2)$$

2. Menentukan nilai *residual* dengan mengurangkan nilai *class* setiap data dengan nilai *probability* awal tertera seperti dalam Persamaan (3).

$$Residual\ (Y) = Class\ Value - Probability \quad (3)$$

3. Membuat *root* awal dari *classification tree* dengan *residual* yang telah ditentukan dengan menjumlahkan semua *residual* tersebut. Selanjutnya, membuat *leaf* atau dengan mengklasifikasikan berdasarkan *feature* yang ada.

4. Menghitung *similarity* atau kesamaan antara data dirumuskan seperti dalam Persamaan (4).

$$Similarity\ score = \frac{((\Sigma Residual)^2}{\Sigma(p\ x\ (1-p) + \lambda} \quad (4)$$

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

5. Setelah menghitung semua *similarity* dan setiap *leaf*, selanjutnya menghitung nilai *gain* dari *left similarity* dan *right similarity* dirumuskan seperti dalam Persamaan (5):

$$Gain = (Left_{similarity} + Right_{similarity}) - Root_{similarity} \quad (5)$$

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

6. Kemudian melakukan *iterasi* atau percabangan menggunakan *pruning* untuk menghitung selisih antara *gain* dari cabang paling bawah dari *tree* dengan nilai *gamma* yang sudah ditetapkan ditampilkan dengan Persamaan (6).

$$Gain - Gamma \tag{6}$$

Apabila dalam operasi kondisi tersebut mendapatkan hasil diatas <0, maka leaf tersebut dipangkas dan data tersebut tidak digunakan lagi. Namun, jika bernilai >0 artinya *leaf* tidak bisa dipangkas.

7. Kemudian menghitung *output value* dari setiap *leaf* dapat dilihat dengan Persamaan (7).

$$Output\ Value = \frac{((\Sigma Residual)}{\Sigma(p\ x\ (1-p) + \lambda} \tag{7}$$

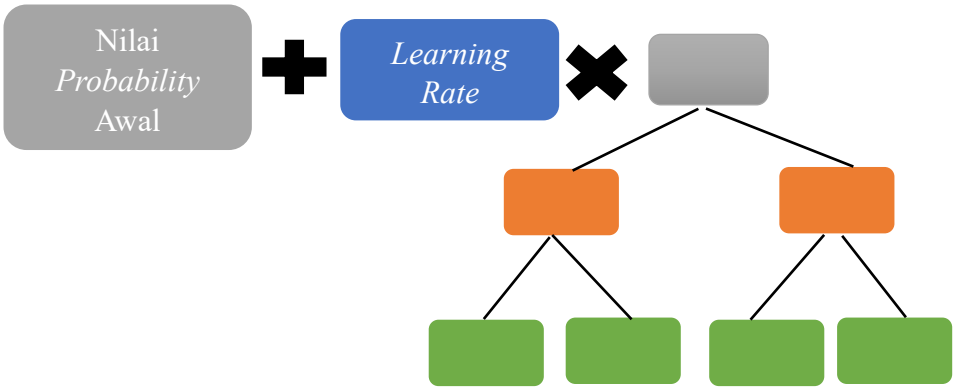
8. Setelah diketahui *output* dari setiap *leaf* perlu dilakukan *scale* dengan mengalikan dengan *learning rate* tertera pada Persamaan (8).

$$Scale\ data(P) = \log\left(\frac{p}{1} - p\right) + (learning\ rate\ x\ output\ value) \tag{8}$$

Nilai *learning rate* biasanya dengan *range* 0-1. Kemudian nilai tersebut dimasukkan kedalam Persamaan (9).

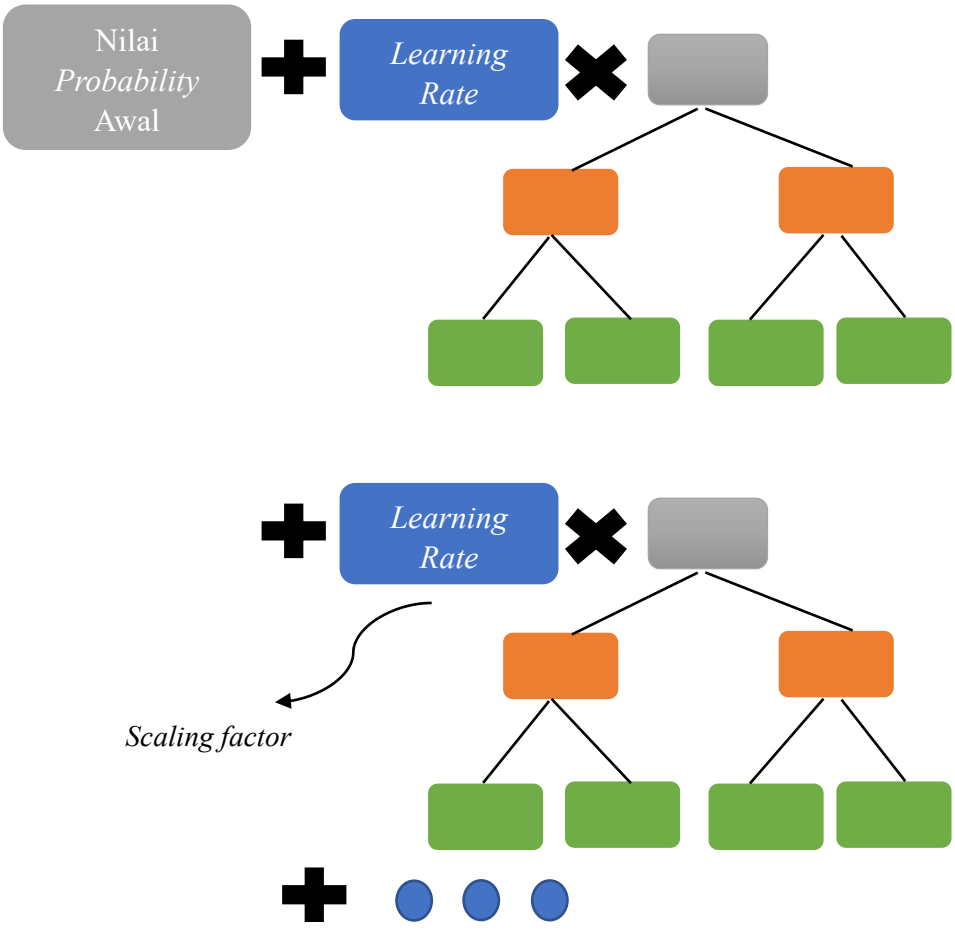
$$Probability\ baru = \frac{e^P}{1 + e^P} \tag{9}$$

Selanjutnya membuat *probability* baru dengan menggabungkan nilai-nilai seperti Gambar 7.



Gambar 7. *Probability* baru dalam *XGBoost*.

9. Mengulangi langkah-langkah dari nomor 4 dengan model *tree* dan kondisi *feature* yang berbeda sehingga nilai *residual* seminimal mungkin atau sampai mendapatkan jumlah *tree* yang maksimal.

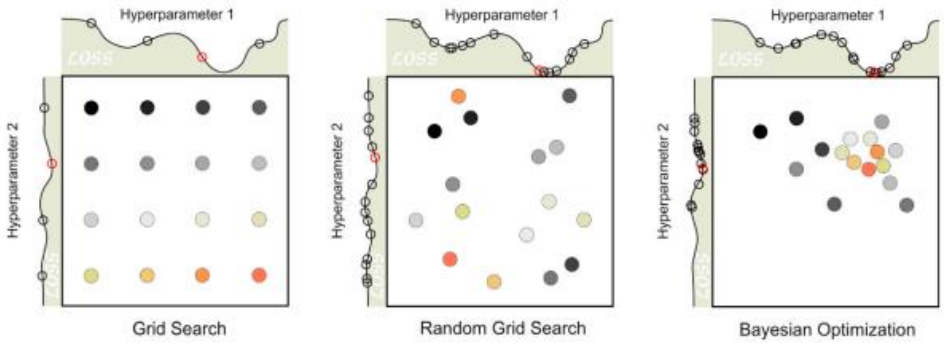


Gambar 8. Penggambaran Algoritma *XGBoost*.

2.7 *Tuning Hyperparameter*

Pada *machine learning*, terdapat beberapa nilai parameter yang diperkirakan dapat meningkatkan kinerja model dikenal dengan istilah *hyperparameter*. *Hyperparameter* digunakan untuk meningkatkan hasil kinerja pada algoritma, yang cukup mempengaruhi berbagai uji model. *Hyperparameter* dilakukan dengan cara menguji sekelompok *hyperparameter* pada batas yang telah ditentukan, pencarian hyperparameter dapat dilakukan dengan cara manual maupun dengan menguji kumpulan *hyperparameter* yang telah ditentukan sebelumnya (Muslim & Karo, 2020).

Nilai-nilai dari parameter model berasal dari data latih, serta mengacu pada bobot dan koefisien yang berasal dari data oleh algoritma. Parameter yang ditentukan misalnya *decision tree* untuk mengukur kedalaman *tree*. Terdiri berbagai *hyperparameter* yang membutuhkan penyetelan untuk kinerja algoritma yang optimal disebut dengan optimasi *hyperparameter*. Optimasi *hyperparameter* berdampak pada kinerja dari algoritma pembelajaran mesin yang telah terbukti secara teoritis dan empiris dari penelitian yang sudah ada. Optimasi *hyperparameter* membutuhkan waktu yang cukup lama jika dilakukan dengan pencarian manual, seperti *gridsearch*, dan *randomsearch*, dan alternatif lainnya yaitu melakukan optimasi *Bayesian* (Putatunda & Rama, 2018) yang ditampilkan pada Gambar 9.



Gambar 9. Ilustrasi *Tuning Hyperparameter* (Passos & Mishra, 2022).

2.7.1 *Grid Search*

Ilustrasi *hyperparameter* pada (Elgeldawi, et al., 2021) menjelaskan bahwa terdapat perbedaan dalam melakukan *tuning hyperparameter*. *Grid Search*

digunakan untuk melatih algoritma pembelajaran mesin untuk semua kombinasi dari *hyperparameter*, proses ini biasanya dipandu oleh performa metrik yang diukur menggunakan teknik *cross validation* pada data latih. Tujuan dari validasi ini untuk memastikan bahwa model terlatih memperoleh sebagian besar pola dari dataset. *Grid search* adalah metode *hyperparameter* yang paling mudah, tekniknya cukup membuat kisi dengan setiap kombinasi dari semua *hyperparameter* dengan nilai yang telah diberikan, menghitung skor setiap model, melakukan evaluasi, dan kemudian pemilihan model yang memberikan hasil terbaik.

2.7.2 *Random Search*

Teknik cara *random search* adalah dengan mengambil sampel ruang pencarian dan mengevaluasi set dari semua distribusi probabilitas yang telah ditentukan, dimana kombinasi acak dari *hyperparameter* digunakan untuk menemukan solusi terbaik untuk model yang akan dipertimbangkan. Jumlah evaluasi dalam *random search* harus diatur dari awal, sebelum proses pengoptimalan *hyperparameter* dimulai.

2.7.3 *Optimasi Bayes*

Optimasi *Bayes* merupkan infromasi pencarian algoritma, yang diartikan bahwa setiap iterasi algoritma belajar dari sebelumnya, dan hasil dari satu iterasi ke iterasi berikutnya. Pertama,beberapa kombinasi *hyperparameter* dipilih secara acak dan diuji, kemudian nilai *hyperparameter* dipilih secara acak untuk menghasilkan model fungsi yang pertama. Setelah draf pertama model diperoleh, ada dua metode untuk memilih kombinasi berikutnya dari ruang pencarian hyperparameter. Optimasi *bayes* menyerupai metode *random search* dengan mengambil sampel *subset* dari beberapa kombinasi *hyperparameter*, namun berbeda dalam kombinasi yang dipilih. Optimasi bayes tidak mengambil sampel setiap kombinasi, seperti *grid search*, dan pada saat yang bersamaan, ini mengahasilkan cara yang lebih sistematis daripada *random search*.

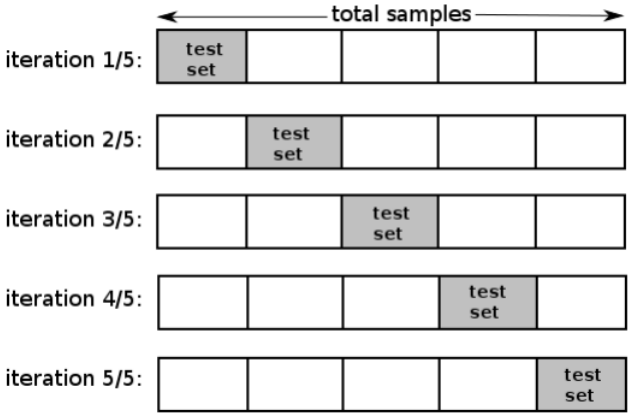
2.8 *Cross Validation*

Cross validation adalah metode statistik yang digunakan untuk memperkirakan akurasi dan mengevaluasi kinerja model pada pembelajaran mesin. *Cross validation* sangat membantu untuk menguji seberapa efektif model pembelajaran mesin saat data terbatas. Dalam melakukan *cross validation*, sebagian data harus dipisahkan untuk pengujian dan validasi, dan tidak akan digunakan untuk melatih model.

Cross validation juga dapat mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua bagian, yaitu untuk melakukan pembelajaran dan pelatihan pada model, dan untuk memvalidasi suatu model (Refaeilzadeh, et al., 2009). Adapun jenis jenis dari *cross validation* sebagai berikut:

2.8.1 *k-fold Cross Validation*

k-fold cross validation merupakan teknik umum untuk memperkirakan kinerja pengklasifikasian. Cara kerja *k-fold cross validation* yaitu pertama kali data akan dipartisi menjadi *k* yang sama dengan ukuran iterasi atau lipatan. Selanjutnya, *k* pada model pelatihan dan validasi dilakukan pengulangan atau iterasi sedemikian rupa sehingga dalam setiap iterasi. Iterasi data yang berbeda tidak berlaku untuk dilakukan validasi, dan *k-1 (k minus 1)*, iterasi digunakan untuk pembelajaran atau *training* (Refaeilzadeh, et al., 2009). Ilustrasi cara kerja *k-fold cross validation* dapat dilihat pada Gambar 10.

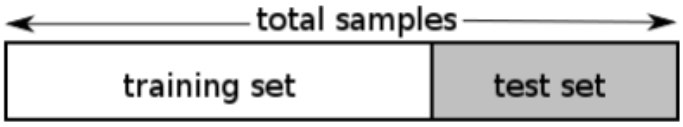


Gambar 10. Ilustrasi *k-fold Cross Validation* (Chlis, 2013).

Dalam permasalahan klasifikasi biner dimana setiap kelas terdiri dari 50% dari data, itu terbaik untuk mengatur sata sedemikian rupa pada setiap iterasi, serta masing-masing kelas terdiri dari sekitar sekitar setengah dari *instances*. Biasanya pada data *machine learning* maupun *data mining*, 10 *fold cross validation (k=10*) adalah yang paling umum digunakan.

2.8.2 *Holdout Cross Validation*

Menurut Chlis, 2013 menjelaskan bahwa *holdout cross validation* adalah metode *cross validation* yang paling sederhana, dengan membagi sampel yang ada menjadi ke dalam dua bagian. Pada set pelatihan terdiri dari sebagian besar sampel yang digunakan untuk melatih model sedangkan set uji dengan persentase yang lebih kecil dari sampel yang ada untuk mengevaluasi kemampuan generalisasi model. Ilustrasi *holdout cross validation* ditunjukkan seperti Gambar 11.

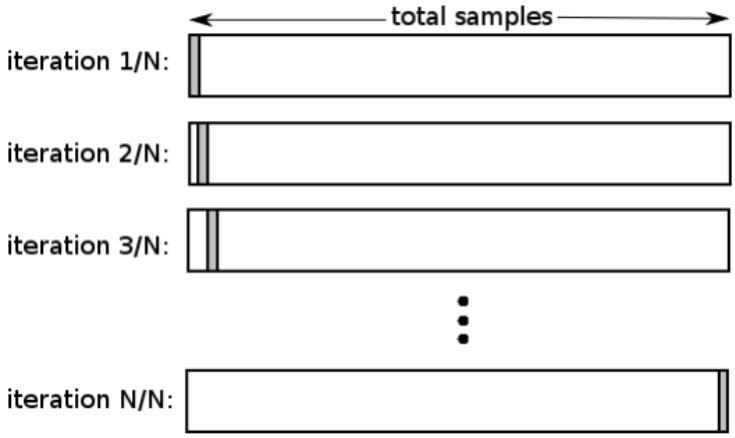


Gambar 11. Ilustrasi *Holdout Cross Validation* (Chlis, 2013).

Teknik evaluasi model pada *holdout cross validation* yang digunakan pada data pelatihan dalam menghitung kesalahan bergantung dari jenis masalah pada penelitian. Kelemahan dari *holdout* adalah prosedur tersebut tidak menggunakan semua data yang tersedia dan hasilnya bergantung pada pilihan untuk pelatihan atau tes terpisah (Refaeilzadeh, et al., 2009). Apabila data pelatihan atau pengujian tidak mempresentasikan data secara lengkap maka hasil data dapat menjadi tidak baik (Yadav & Shukla, 2016).

2.8.3 *Leave One Out Cross Validation*

Leave one out cross validation (LOOCV) adalah teknik validasi model khus dari *k-fold cross validation* dimana *k* sama dengan banyaknya jumlah data. Metode validasi ini banyak digunakan untuk data yang jarang maupun data yang sedikit di dalam bidang bioinformatika. Ilustrasi *leave one out cross validation* terlihat seperti Gambar 12.



Gambar 12. Ilustrasi *Leave One Out Cross Validation* (Chlis, 2013).

Penerapan metode ini dengan cara membagi *k=N*, dimana *N* adalah banyaknya data. Dimana *N-1* observasi digunakan untuk data latih dan 1 merupakan observasi sebagai data uji sehingga semua data berkesempatan untuk menjadi data latih dan daya uji. Perkiraan akurasi yang didapatkan dari metode validasi ini hamper tidak bias, akan tetapi memiliki varians yang sangat tinggi (Refaeilzadeh, et al., 2009).

2.9 *Confusion Matrix*

Confusion matrix memuat informasi mengenai aktual klasifikasi dan prediksi klasifikasi yang dilakukan oleh sistem, serta untuk mengevaluasi kinerja sistem dengan menyajikan data kedalam bentuk matriks. *Confusion matrix* diterapkan untuk klasifikasi biner maupun multi kelas (Kulkarni, et al., 2020). Berikut adalah contoh *confusion matrix* untuk klasifikasi biner pada Tabel 2.

Tabel 2. *Confusion Matrix* Klasifikasi Biner.

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

Adapun istilah yang digunakan pada *confusion matrix* yang tertera pada Tabel 2, diantaranya:

1. *True Negative* adalah jumlah prediksi negatif yang diklasifikasikan secara akurat.
2. *True Positive* adalah jumlah prediksi positif yang diklasifikasikan secara akurat.
3. *False Positive* adalah jumlah prediksi salah yang diklasifikasikan sebagai positif.
4. *False Negative* adalah jumlah prediksi salah dan diklasifikasikan sebagai negatif.

Menurut Kulkarni, et al., 2020, *confusion matrix* memiliki hasil statistik untuk penilaian dan evaluasi kinerja. Hasil tersebut diantaranya sebagai berikut:

2.9.1 *Accuracy*

Accuracy adalah persentase seberapa akuratnya suatu model dalam mengklasifikasikan (prediksi) secara benar oleh algoritma. Fungsi dari *accuracy* dapat dilihat pada Persamaan (10).

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{10}$$

2.9.2 *Precision*

Precision adalah persentase tingkat ketepatan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Fungsi dari *precision* dapat dilihat pada Persamaan (11):

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

2.9.3 *Recall*

Recall adalah persentase keberhasilan suatu model dalam menemukan kembali (observasi) sebuah informasi. Fungsi dari *recall* dapat dilihat pada Persamaan (12).

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

2.9.4 *F1 Score*

F1 score adalah perbandingan rata-rata nilai *recall* dan *precision* yang dibobotkan. *F1 score* digunakan apabila ingin membuat model klasifikasi dengan keseimbangan *recall* dan *precision* secara optimal. Fungsi dari *F1 score* dapat dilihat pada Persamaan (13).

$$F1\ Score = \ 2\text{x}\ \frac{(Precision\ x\ Recall)}{(Precision + Recall)} \tag{13}$$

III. METODOLOGI PENELITIAN

3.1 Tempat dan Waktu Penelitian

3.1.1 Tempat

Penelitian ini dilakukan di Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung yang beralamatkan di Jalan Soemantri Brojonegoro No. 1 Gedung Meneng, Bandar Lampung.

3.1.2 Waktu

Penelitian ini dilaksanakan dari mulai semester genap tahun ajaran 2020/2021 sampai bulan Desember 2022. Berikut adalah ringkasan dari alokasi penelitian ini sebagai berikut:

1. Pengambilan atau Pengumpulan Data
   Data *microarray* diambil dari jurnal (Kumar, et al., 2012) yang dapat diakses melalui website *National Center of Biotechnology Information* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283953/
2. *Cross Validation*
   Menggunakan *k-fold cross validation* dengan 10 kali iterasi.
3. Klasifikasi
   Penelitian ini dibuat dengan melatih model *extreme gradient boosting (boosting)* menggunakan data latih, serta melakukan proses prediksi.
4. *Hyperparameter*
   Menggunakan metode *gridsearch cross validation* untuk menentukan beberapa parameter yang digunakan dan optimal.
5. Evaluasi
   Melakukan evaluasi atau kinerja pada model.

3.2 Data dan Alat

3.2.1 Data

Penelitian ini menggunakan data *microarray* yang didapatkan pada jurnal yang berjudul *Application of microarray in breast cancer: An overview* pada tahun 2012. Data diambil dari *National Center for Biotechnology Information (NCBI)* yang dapat diakses melakui https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283953/. Tampilan dataset dapat dilihat pada Tabel 3.

Tabel 3. Data *Microarray*

Pasien	Kelas	X1	X2	X3	⋯	X22.213	X22.214	X22.215
1	0	2461.4	26.7	82.6	⋯	206.4	233.4	120.7
2	0	3435.7	159	243.4	⋯	296.5	390.7	168.2
3	0	1932.5	31.2	150.2	⋯	185.5	242.1	57.8
⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮
40	1	2390.3	115.4	73.6	⋯	224.9	267.2	56.8
41	1	2738.8	14.1	122.7	⋯	1751.8	83.6	247.8
42	1	3233.1	47.6	107.6	⋯	703.2	212.3	103.1

Data dipresentasikan ke dalam suatu matriks. Jumlah fitur yang digunakan untuk mengklasifikasikan seseorang yang terkena kanker payudara sebanyak 22.215 fitur. Fitur X1 sampai X22.215 tersebut diartikan sebagai gen kanker payudara yang telah diteliti. Setiap fitur memiliki nilai yang disebut nilai ekspresi gen. Jumlah kelas pada penelitian ini sebanyak 2 kelas yaitu 1 dan 0. Pada kelas 1 menunjukkan seseorang terkena kanker ganas, dan kelas 0 menunjukkan seseorang yang tidak terkena kanker jinak.

3.2.2 Alat

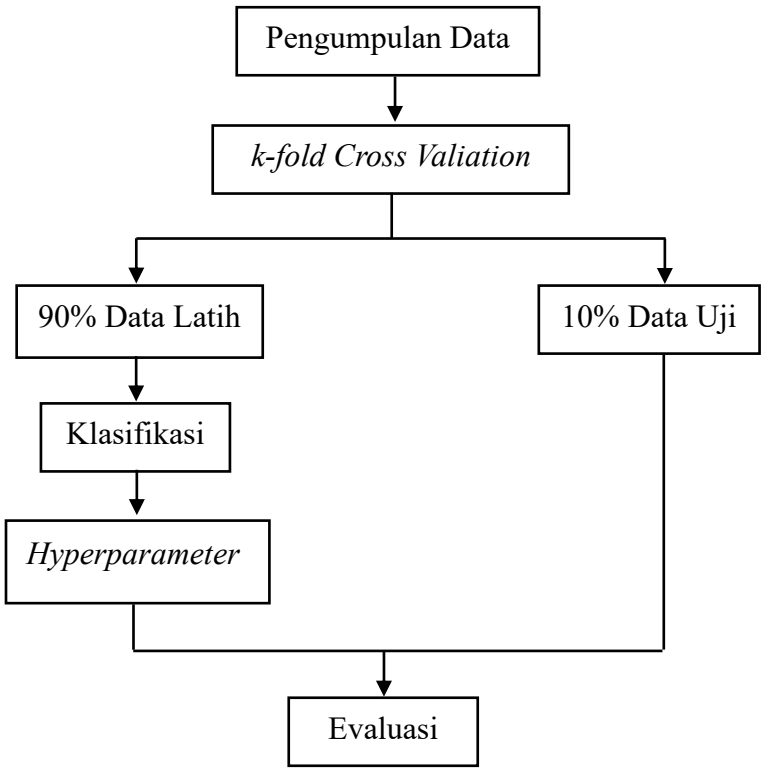
Penelitian ini dilakukan dengan menggunakan perangkat keras dan perangkat lunak sebagai berikut:

1. Perangkat keras: *Processor* Intel ® Celeron CPU N3350 @1.10GHz up to 2.4GHz, DDR3, RAM 4GB, SSD 500GB, HDD 500GB, System type 64-bit *Operating System, x-64-based processor.*

2. Perangkat lunak: *Operating System Windows 64-bit, Microsoft Excel* 2019*, Microsoft Word* 2019*, Microsoft PowerPoint* 2019*, Jupyter Notebook*, Python 3.

3.3 Metode

Tahapan metode yang dilakukan pada penelitian ini, digambarkan dalam diagram alur yang ditunjukkan oleh Gambar 13.



Gambar 13. Tahapan Penelitian Klasifikasi Metode *XGBoost*.

Adapun tahapan dari metode pada penelitian ini, diantaranya sebagai berikut:

3.3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini didapatkan dari penelitian yaitu (Kumar, et al., 2012), dimana data tersebut diperoleh dari *National Center for Biotechnology Information* yang terdiri dari ribuan ekspresi gen dari X1 sampai X22.215. Jumlah data sebanyak 42 pasien yang digolongkan ke dalam dua kelas yaitu kelas 0 sebanyak 24 pasien kanker jinak dan kelas 1 sebanyak 18 pasien kanker ganas.

3.3.2 Pembagian Data

Pada proses ini, data dibagi menjadi dua bagian untuk proses *testing* dan *training*. Pada tahapan ini akan dilakukan *cross validation* dengan *k-fold* . Pada *k-fold cross validation*, data diacak dengan pembagian data 10% untuk data *testing*, dan 90% untuk data *training*. Kemudian model dilatih menggunakan lipatan *k-1 (k minus 1)* dimana *k=10*, sebanyak 9 iterasi digunakan untuk melatih dan iterasi yang tersisa dipakai untuk pengujian. Proses diulang sebanyak 10 kali setiap iterasi k untuk set testing.

3.3.3 Klasifikasi

Dalam penelitian ini bertujuan untuk mengklasifikasikan sampel pasien kanker payudara menggunakan *XGBoost*, menjadi salah satu implementasi dari algoritma *boosting* dengan menerapkan teknik regulasi untuk mengurangi *overfitting*. Pada tahap ini, akan dilakukan perbandingan parameter yang akan digunakan sebelum dan sesudah *hyperparameter*.

3.3.4 *Hyperparameter*

Pada tahapan ini, melakukan proses *hyperparameter* menggunakan salah satu metode yaitu *gridsearch cross validation* dengan beberapa parameter seperti *base_score, max_depth, sub_sample, n_estimaotrrs, learning_rate, min_child_weight, gamma,* dan *colsample_bytree* untuk menghasilkan nilai optimal dari *range* pada setiap parameter. Serta untuk meningkatkan nilai akurasi dari sebelum melakukan *hyperparameter*.

3.3.5. Evaluasi

Tahapan terakhir dalam penelitian ini yaitu melakukan evaluasi atau menilai performa model *XGBoost* terhadap data *testing*, serta mendapatkan informasi mengenai seberapa baik klasifikasi dilakukan. Adapun penilaian kinerja model, seperti akurasi, *precision*, *recall*, dan *f1-score*.

V. SIMPULAN DAN SARAN

5.1 Simpulan

Dari penelitian yang telah dilakukan, didapatkan hasil dan dapat diambil kesimpulan sebagai berikut:

1. Penelitian ini menggunakan pembagian data dengan *k-fold cross validation,* yang dimana *n_splits* sebanyak 10 kali. Data diperoleh dari jurnal Kumar, et al., pada tahun 2012 yang berisikan data sebanyak 42 pasien dengan jumlah fitur sebanyak 22.215 ekspresi gen *microarray* kanker payudara. Didapatkan hasil nilai terbaik *hyperparameter* secara *grid search*, sehingga memberikan nilai dan peningkatan akurasi setelah dilakukan *hyperparameter*. Sebelum *hyperparameter* mendapatkan nilai akurasi sebesar 50%, presisi 50%, *recall* 50%, dan *f1-score* 50%. Kemudian setelah melakukan *tuning hyperparameter* mendapatkan nilai akurasi, presisi, *recall* %, dan *f1-score* masing-masing sebesar 76%, 86%, 60%, dan 71%.
2. Pada penelitian ini menggunakan beberapa parameter untuk meningkatkan performa model setelah proses *hyperparameter*. Parameter yang digunakan dengan nilai yang terbaik antara lain *base_score* 0.4, *max_depth* bernilai 1, *sub_sample* sebesar 0.38, *n_estimators* 70*, learning_rate* 0.2, *min_child_weight* bernilai 1.0, nilai *gamma* 0.1, dan parameter *colsample_bytree* sebesar 0.2.

5.2 Saran

Dari penelitian yang telah dilakukan terdapat saran untuk meningkatan penelitian sebagai berikut:

1. Disarankan untuk penelitian selanjutnya dapat menggunakan metode klasifikasi lainnya untuk dapat dibandingkan dengan hasil nilai akurasi pada penelitian ini.

2. Mencoba menggunakan data DNA *microarray* lainnya untuk melakukan klasifikasi *binary class* menggunakan *XGBoost* untuk mengetahui kinerja atau performa *XGBoost*.

DAFTAR PUSTAKA

Abdillah, A. A., Murfi, H., & Satria, Y. (2016). *Uji Kinerja Learning To Rank Dengan Metode Support Vector Regression*. *1*(January 2015), 14–25.

Andryan, M. R., Fajri, M., & Sulistyowati, N. (2022). Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosis Penyakit Kanker Payudara. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(1), 1. https://doi.org/10.26798/jiko.v6i1.500

Bayat, A. (2002). Science, medicine, and the future Bioinformatics. *Bmj*, *324*(April), 1018–1022. www.ebi.ac.uk

Bhatia, S., & Dahiya, R. (2015). Concepts and Techniques of Plant Tissue Culture Science. In *Modern Applications of Plant Biotechnology in Pharmaceutical Sciences*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-802221-4.00004-2

Bolstad, B. M. (2004). *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization* (p. 156). papers2://publication/uuid/8B996D4A-CD91-4F11-9F50-7B5E60EFC00C

Bustan, D. M. N. (2007). *Epidemiologi Penyakit Tidak Menular*. PT. Rineka Cipta.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Knowledge Discovery and Data Mining*. https://doi.org/http://dx.doi.org/10.1145/2939672.2939785

Chlis, N.-K. (2013). Comparison of Statistical Methods for Genomic Signature Extraction. *Tech. Univ. Crete, Chania, Greece, Tech. Rep*, *October*. https://doi.org/10.13140/2.1.2230.6563

Diani, R., Wisesty, U. N., & Aditsania, A. (2017). Analisis Pengaruh Kernel Support Vector Machine (SVM) pada Klasifikasi Data Microarray untuk Deteksi Kanker. *Indonesian Journal of Computing*, *2*(1), 109–118. https://doi.org/10.21108/indojc.2017.21.169

Dr. dr. Imam Rasjidi, S. (K)Onk. (2009). *Deteksi Dini Pencengahan Kanker Pada*

Wanita (S. (K)Onk Dr. dr. Imam Rasjidi & dr. L. Kusumo (eds.)).
Dufva, M. (2009). *DNA Microarrays for Biomedical Research* (Vol. 529, Issue 2). https://doi.org/10.1007/978-1-59745-538-1

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. *Informatics*, *8*(4), 1–21. https://doi.org/10.3390/informatics8040079

Firmansyah, R. R. D. L., Tjandrasa, H., & Arieshanti, I. (2012). Pengembangan Perangkat Lunak Prediktor Kanker Payudara Menggunakan Metode Elastic SCAD SVM dan Data DNA Microarray. *Jurnal Teknik ITS*, *1*(1), 216–221.

Globacan. (2019). *266 794 986*. *256*, 2018–2019.

Globocan. (2018). *7 632 819 272*. *876*, 2018–2019.

Globocan. (2020). The Global Cancer Observatory - All cancers. *International Agency for Research on Cancer - WHO*, *419*, 199–200. https://gco.iarc.fr/today/home

Gonzalo, R., & Sánchez, A. (2018). Introduction to Microarrays Technology and Data Analysis. In *Comprehensive Analytical Chemistry* (1st ed., Vol. 82). Elsevier B.V. https://doi.org/10.1016/bs.coac.2018.08.002

Handayani, A., Jamal, A., & Septiandri, A. A. (2017). Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *6*(4), 394–403. https://doi.org/10.22146/jnteti.v6i4.350

Iiris Hovatta, Kimppa, K., Laine, M. M., Lehmussola, A., Pasanen, T., Saarela, J., Saarikko, I., Saharinen, J., Tiikkainen, P., Toivanen, T., Tolvanen, M., Tuimala, J., Vihinen, M., & Wong, G. (2014). DNA Microarray Data Analysis. In *Scientific Computing Ltd* (Second, Vol. 4, Issue 1). CSC - the Finnish IT center for science.

Kalqutny, S. H., Pakki, S., & Muis, A. (2020). Potensi Pemanfaatan Teknik Molekuler Berbasis DNA dalam The Potential Use of DNA Based Molecular Techniques in The Study of Maize Downy Mildew. *Jurnal Ilmu Dan Teknologi Pertanian Agrosaintek*, *4*(1), 17–27.

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*.

Elsevier Inc. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

Kumar, A. (2010). Gene : Expression and Regulation (in Recent Advances in Life Sciences). In P. A. K. Rai & P. R. P. Sinha (Eds.), *Gene : Expression and Regulation* (1st ed., Issue February).

Kumar, R., Sharma, A., & Tiwari, R. K. (2012). Application of microarray in breast cancer: An overview. *Journal of Pharmacy and Bioallied Sciences*, *4*(1), 21–26. https://doi.org/10.4103/0975-7406.92726

Lamartine, J. (2006). The benefits of DNA microarrays in fundamental and applied bio-medicine. *Materials Science and Engineering C*, *26*(2–3), 354–359. https://doi.org/10.1016/j.msec.2005.10.068

Muslim, I., & Karo, K. (2020). Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. *Journal of Software Engineering, Information and Communication Technology*, *1*(1), 10–16.

Nugraha, W. (2021). Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi. *Jurnal Managemen Dan Informatika*, *9*(2), 3–8.

Nugraha, W., & Sasongko, A. (2022). Hyperparameter Tuning pada Algoritma Klasifikasi dengan Grid Search. *SISTEMASI : Jurnal Sistem Informasi*, *11*(2), 391–401.

Nuklianggraita, T. N., Adiwijaya, A., & Aditsania, A. (2020). On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier. *Jurnal Infotel*, *12*(3), 89–96. https://doi.org/10.20895/infotel.v12i3.485

Passos, D., & Mishra, P. (2022). A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems*, *223*(October 2021). https://doi.org/10.1016/j.chemolab.2022.104520

Putatunda, S., & Rama, K. (2018). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. *ACM International Conference Proceeding Series*, 6–10. https://doi.org/10.1145/3297067.3297080

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross Validation. *Dementia with Lewy Bodies: And Parkinson's Disease Dementia*, 1–8. https://doi.org/10.1007/978-0-387-39940-9_565

Sholihin, R. (2017). *Mengenal, Mencegah, & Mengatasi "Silent Killer" Kanker* (I. W. S.N (ed.)). Romawi Pustaka.

Simon, Annina, Deo, Mahima, Selvam, Venkatesan, Babu, & Ramesh. (2016). *An Overview of Machine Learning and its Applications. International Journal of Electrical Sciences & Engineering. Volume. 22-24. 1*(January), 22–24. https://www.researchgate.net/publication/289980169_An_Overview_of_Machine_Learning_and_its_Applications

Trevino, V., Falciani, F., & Barrera-Saldaña, H. A. (2007). DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, *13*(9), 30–39. https://doi.org/10.2119/2006

Yadav, S., & Shukla, S. (2016). Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, *Cv*, 78–83. https://doi.org/10.1109/IACC.2016.25

Yang, P., Yang, J. Y. H., Zhou, B. B., & Zomaya, A. Y. (2010). A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics*, *5*(4), 296–308. https://doi.org/10.2174/157489310794072508