MODEL HIBRID PEMILIHAN FITUR BERINTERAKSI MENGGUNAKAN PERATURAN KESATUAN KELAS DAN PENYEPUHLINDAPAN SIMULASI BAGI PENGELASAN KANSER PAYUDARA

SHAHIRATUL AMALINA BINTI ABD KARIM

UNIVERSITI KEBANGSAAN MALAYSIA

MODEL HIBRID PEMILIHAN FITUR BERINTERAKSI MENGGUNAKAN PERATURAN KESATUAN KELAS DAN PENYEPUHLINDAPAN SIMULASI BAGI PENGELASAN KANSER PAYUDARA

SHAHIRATUL AMALINA BINTI ABD KARIM

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH
IJAZAH DOKTOR FALSAFAH

INSTITUT VISUAL INFORMATIK
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2025

UNIVERSITI KEBANGSAAN MALAYSIA
The National University of Malaysia

PERAKUAN TESIS SARJANA / DOKTOR FALSAFAH
(DECLARATION OF MASTER / DOCTOR OF PHILOSOPHY THESIS)

A. MAKLUMAT PELAJAR/*STUDENT DETAILS*	
Nama *Name*	SHAHIRATUL AMALINA BINTI ABD KARIM
No. Pendaftaran *Registration No.*	P105766
Fakulti/Institut *Faculty/Institute*	INSTITUT INFORMATIK VISUAL (IVI)
Program Pengajian *Program of Study*	~~Sarjana~~ / Doktor Falsafah | *Masters / Doctor of Philosophy*
Tajuk Tesis *Thesis Title*	MODEL HIBRID PEMILIHAN FITUR BERINTERAKSI MENGGUNAKAN PERATURAN KESATUAN KELAS DAN PENYEPUHLINDAPAN SIMULASI BAGI PENGELASAN KANSER PAYUDARA
Mel-e/*Email*: shahiratulamalina@gmail.com	No. Telefon/*Tel.No.*: 011-14405255

B. PERAKUAN / *DECLARATION*

Merujuk kepada Dasar Harta Intelek UKM 2018, **tesis adalah hak milik UKM** seperti keterangan dibawah:

4.2.1 Harta Intelek yang direka cipta oleh pelajar termasuk tesis dan semua hasil penyelidikan ditulis oleh Pelajar UKM dalam tempoh pengajiannya di UKM direka cipta, dibangunkan atau dihasilkan menggunakan kemudahan, bahan, dana, atau sumber lain adalah hak milik UKM melainkan dinyatakan sebaliknya dalam Dasar ini atau dipersetujui sebaliknya oleh UKM dan Pelajar/Penaja.

*Referring to the UKM Intellectual Property Policy 2018, **the thesis is the property of UKM** as described below:*

4.2.1 Intellectual property created by the student, including the thesis and all research output produced by the UKM Student during their studies at UKM, including designed, developed or produced output using facilities, materials, funds, or other resources are the property of UKM unless otherwise stated in this Policy or otherwise agreed by UKM and the Student/Sponsor.

	RAHSIA ***CONFIDENTIAL***	Mengandungi maklumat rahsia di bawah AKTA RAHSIA RASMI 1972 *Consisting of classified information under the OFFICIAL SECRETS ACT 1972*
	TERHAD ***RESTRICTED***	Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi / badan di mana penyelidikan dijalankan *Consisting of RESTRICTED information which has been determined by the organisation/body where the research was conducted*
/	**AKSES TERBUKA/ TIDAK TERHAD** ***OPEN ACCESS/ NON-RESTRICTED***	Saya membenarkan tesis ini diterbitkan secara akses terbuka, teks penuh atau dibuat salinan untuk tujuan pengajian, pembelajaran & penyelidikan sahaja *I give permission for this thesis to be published through open access, full text or copied only for study, learning, and research purposes.*

Bagi kategori Akses Terbuka/Tidak Terhad, saya membenarkan tesis (Sarjana/Doktor Falsafah) ini di simpan di Perpustakaan Universiti Kebangsaan Malaysia (UKM)* dengan syarat-syarat kegunaan seperti berikut:

For the Open Access/Non-Restricted category, I permit this (Master's/Doctoral) Thesis to be kept in the Universiti Kebangsaan Malaysia (UKM) Library with the following conditions:

1. Perpustakaan UKM mempunyai hak untuk membuat salinan untuk tujuan pengajian,pembelajaran, penyelidikan sahaja.
   UKM Library has the right to reproduce the thesis only for study, learning and research purposes.

2. Perpustakaan Universiti Kebangsaan Malaysia dibenarkan membuat satu (1) salinan tesis ini untuk tujuan pertukaran antara institusi pengajian tinggi dan mana-mana badan/ agensi kerajaan, tertakluk kepada terma dan syarat.
   UKM Library is allowed to make one (1) copy of this thesis for exchange purposes among higher education institutions and any government body/agency, subject to the terms and conditions.

Saya mengakui maklumat & tesis yang diberikan adalah benar & terkini. *I acknowledge the information & the thesis provided are correct & current.* Tandatangan Pelajar: *Student's Signature* shakiratulamalina *Tarikh / Date: 27/11/2025*	Saya memperakukan maklumat & tesis yang diberikan adalah benar & terkini. *I verify that the information & the thesis provided are correct & current.* DR. UMMUL HANAN MOHAMAD RESEARCH FELLOW Institute of Visual Informatics Universiti Kebangsaan Malaysia Tandatangan Penyelia / *Supervisor's Signature*: Nama / *Name*: Cop Rasmi / *Official Stamp*: *Tarikh / Date: 4/12/2025*

C. PENGESAHAN FAKULTI/INSTITUT | *VERIFICATION OF FACULTY/INSTITUTE*

Tandatangan/*Signature*:
Nama/*Name*:
Tarikh/*Date*:
Cop Rasmi/*Official Stamp*:

MOHAMAD SALLEH SULIEMAN
Penolong Pendaftar Kanan
Institut Informatik Visual(IVI)
Universiti Kebangsaan Malaysia

**Kuatkuasa:15 Julai 2021

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

04 Disember 2025

SHAHIRATUL AMALINA
BINTI ABD KARIM
P105766

PENGHARGAAN

Dengan nama Allah yang Maha Pengasih lagi Maha Penyayang.

Alhamdulillah. Setinggi-tinggi kesyukuran kepada Allah S.W.T atas limpah rahmat dan hidayah-Nya telah memberikan saya kekuatan, ketabahan dan ilham untuk menyiapkan tesis ini dengan jayanya. Selawat dan salam ke atas junjungan besar Nabi Muhammad S.A.W, ahli keluarga baginda dan para sahabat baginda.

Pertama sekali, ucapan jutaan terima kasih yang tulus ikhlas saya tujukan kepada penyelia saya, Dr. Ummul Hanan Mohamad dan penyelia bersama, Dr. Puteri Nor Ellyza Nohuddin, PM Dr. Mohd Hanafi Ahmad Hijazi serta PM Dr. Akmal Sabarudin atas segala tunjuk ajar, bimbingan yang berterusan serta sokongan moral dan ilmiah sepanjang penyelidikan ini dijalankan. Segala ilmu dan panduan yang dicurahkan amat saya hargai.

Saya juga ingin merakamkan penghargaan kepada Institut Visual Informatik (IVI), Universiti Kebangsaan Malaysia (UKM) atas segala kemudahan penyelidikan, infrastruktur dan suasana akademik yang kondusif sepanjang tempoh pengajian saya. Tidak lupa juga penghargaan diberikan kepada felo-felo penyelidik, rakan-rakan seperjuangan serta kakitangan IVI yang telah membantu melalui perkongsian ilmu, teguran membina dan dorongan sepanjang proses penyelidikan ini.

Saya juga ingin mengucapkan terima kasih kepada Majlis Amanah Rakyat (MARA) atas sokongan kewangan dan pinjaman yang telah membolehkan saya melaksanakan penyelidikan ini dengan lancar dan teratur. Tidak dilupakan juga kepada UKM atas pemberian geran penerbitan yang telah membantu memudahkan proses penerbitan artikel hasil penyelidikan ini.

Tidak dilupakan, setinggi-tinggi penghargaan dan kasih sayang saya titipkan buat suami tercinta Muhammad Izzat Haziq Idris atas sokongan, doa, pengorbanan dan kesabaran yang tidak ternilai sepanjang perjalanan ini. Buat anak saya yang tersayang, Aleena Aleesha, kehadirannya menjadi sumber kekuatan dan inspirasi saya untuk terus melangkah. Ucapan terima kasih yang tidak terhingga ditujukan kepada kedua ibu bapa saya, Encik Abd Karim Sharif dan Puan Hamidah Jaafar Sidek atas doa, kasih sayang dan sokongan tidak berbelah bahagi sejak dari awal pengajian saya. Tidak lupa juga kepada ibu dan bapa mertua saya, Haji Idris Hamid dan Hajah Jamiah Darus yang sentiasa memberikan galakan dan sokongan moral. Ucapan terima kasih turut saya tujukan kepada seluruh ahli keluarga saya iaitu Mohd Syafiq, Nurul Farah, Mohd Faiz, Hidayatul Atiqah, Izzatul Syazwani dan Muhammad Naqib Firdaus serta Nur Balqis Husnina, Izz Thaqif dan Nur Diena Yusrina yang sentiasa berada di sisi dalam memastikan kelancaran usaha ini.

Akhir kata, sekalung terima kasih saya ucapkan kepada semua yang terlibat secara langsung atau tidak langsung dalam membantu saya menyiapkan tesis ini. Semoga segala kebaikan kalian diberkati oleh Allah S.W.T.

ABSTRAK

Kanser payudara merupakan antara penyebab utama kematian wanita di seluruh dunia. Ketepatan dalam proses pengelasan amat kritikal bagi menyokong diagnosis dan rawatan yang berkesan. Namun demikian, model sedia ada menghadapi beberapa kekangan penting dalam mengenal pasti subset fitur yang relevan serta berinteraksi, manakala proses penalaan hiperparameter secara manual dalam model Perlombongan Peraturan Kesatuan Kelas (CARM) lazimnya kurang cekap. Kajian ini mencadangkan satu kerangka hibrid untuk menangani kekangan tersebut melalui tiga objektif utama: (1) merumuskan model pemilihan fitur berasaskan pendiskretan kabur dan CARM bagi mengenal pasti fitur yang relevan, berinteraksi dan tidak bertindih (FD-CARI); (2) mereka bentuk satu kerangka pengesahan fitur berinteraksi menggunakan gabungan kaedah pengundian majoriti berwajaran dan teknik Kecerdasan Buatan yang Boleh Dijelaskan (XAI) berasaskan nilai SHAP; dan (3) membangunkan model hibrid FD-CARI+SA yang mengintegrasikan CARM dengan algoritma Penyepuhlindapan Simulasi (SA) untuk penalaan hiperparameter secara automatik. Kajian ini menggunakan pendekatan kuantitatif terhadap tiga set data kanser payudara iaitu WDBC, WPBC dan BCC yang merangkumi fitur klinikal, diagnostik dan biokimia. Data dipra-proses menggunakan pendiskretan kabur dan algoritma FD-CARI digunakan untuk memilih subset fitur. Fitur berinteraksi yang dipilih kemudiannya disahkan menggunakan nilai SHAP serta kaedahpengundian majoriti berwajaran, manakala penalaan hiperparameter pula dilaksanakan secara automatik menggunakan algoritma SA. Model dinilai menggunakan lima algoritma pengelasan iaitu Pokok Keputusan (DT), Hutan Rawak (RF), Bayes Naif (NB), Regresi Logistik (LR) dan Mesin Sokongan Vektor (SVM). Dapatan menunjukkan bahawa FD-CARI+SA berjaya mengenal pasti subset fitur yang lebih kecil namun lebih berkesan, dengan peningkatan kecil tetapi signifikan dalam skor AUC berbanding kaedah konvensional seperti CFS, FCBF, Consistency, Relief-F dan mRMR. Sebagai contoh, dalam set data WDBC, jumlah fitur berkurang daripada 4 kepada 3 dengan peningkatan skor AUC purata daripada 0.9857 kepada 0.9860. Bagi WPBC, fitur berkurang daripada 5 kepada 4 dan AUC meningkat daripada 0.7015 kepada 0.7035. Set data BCC pula menunjukkan prestasi stabil dengan empat fitur terpilih serta mencapai AUC konsisten pada 0.8044. Selain itu, penggunaan SHAP dalam proses pengesahan turut meningkatkan ketelusan dan kefahaman terhadap sumbangan fitur berdasarkan hubungan interaksi. Justeru, kajian ini menyumbang sebuah kerangka pemilihan fitur berasaskan interaksi yang mempunyai potensi kukuh untuk diaplikasikan dalam tugasan pengelasan data perubatan.

A HYBRID MODEL FOR INTERACTION-BASED FEATURE SELECTION USING CLASS ASSOCIATION RULES AND SIMULATED ANNEALING FOR BREAST CANCER CLASSIFICATION

ABSTRACT

Breast cancer remains one of the leading causes of death among women worldwide. Accurate classification is critical to support early diagnosis and effective treatment. However, existing models face challenges in identifying relevant and interacting feature subsets, and manual hyperparameter tuning in Class Association Rule Mining (CARM) model is often inefficient. This study proposes a hybrid framework to address these limitations through three main objectives: (1) to formulate a feature selection model based on fuzzy discretization and CARM to identify relevant, interacting and non-redundant features (FD-CARI); (2) to design a validation framework for the selected interactive features using a combination of weighted majority voting and Explainable Artificial Intelligence (XAI) technique based on SHAP values; and (3) to develop a hybrid FD-CARI+SA model that integrates CARM with Simulated Annealing (SA) for automated hyperparameter tuning. This study adopts a quantitative approach using three breast cancer datasets such as WDBC, WPBC and BCC which consists of clinical, diagnostic and biochemical features. Data were pre-processed using fuzzy discretization and the FD-CARI algorithm was used to choose feature subsets. The selected interactive features were validated using SHAP values and a weighted majority voting scheme. Hyperparameter optimization was performed automatically using SA algorithm. The model was evaluated using five classifiers such as Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR) and Support Vector Machine (SVM). Results showed that FD-CARI+SA successfully identified smaller yet more effective feature subsets, achieving slight but meaningful improvement in AUC scores compared to conventional methods such as CFS, FCBF, Consistency, Relief-F and mRMR. For example, in the WDBC dataset, the number of features was reduced from 4 to 3 with an AUC increase from 0.9857 to 0.9860. In WPBC dataset, features were reduced from 5 to 4 and AUC increased from 0.7015 to 0.7035. The BCC dataset demonstrated stable performance with four selected features that achieving a consistent AUC of 0.8044. Additionally, the use of SHAP-based validation enhanced transparency and interpretability of feature contributions based on interactions. Hence, this study contributes an interaction-aware selection framework with strong potential for application in medical classification tasks.

KANDUNGAN

Halaman

PENGAKUAN		ii
PENGHARGAAN		iii
ABSTRAK		iv
ABSTRACT		v
KANDUNGAN		vi
SENARAI JADUAL		xii
SENARAI ILUSTRASI		xviii
SENARAI SINGKATAN		xxii
SENARAI ISTILAH		xxiii
BAB I	**PENDAHULUAN**	
1.1	Pengenalan	1
1.2	Latar Belakang Kajian	1
1.3	Persoalan Kajian	3
1.4	Kenyataan Masalah	4
1.5	Objektif Kajian	6
1.6	Skop Kajian	7
	1.6.1 Jenis Data	7
	1.6.2 Fokus Teknikal	7
	1.6.3 Pengesahan Interaksi Fitur	8
	1.6.4 Pengoptimuman Hiperparameter	8
	1.6.5 Penilaian Prestasi Model	8
1.7	Batasan Kajian	9
1.8	Ringkasan Sumbangan Kajian	10
	1.8.1 Sumbangan Konseptual	10
	1.8.2 Sumbangan Metodologi	10
	1.8.3 Sumbangan Aplikasi Praktikal	11
1.9	Struktur Kandungan Tesis	11
BAB II	**KAJIAN KESUSASTERAAN**	
2.1	Pengenalan	13
2.2	Kanser Payudara dan Cabaran Diagnosis Awal	13

	2.2.1	Modaliti Pengimejan Diagnosis Kanser Payudara	14
	2.2.2	Sistem Diagnosis Berbantu Komputer	14
	2.2.3	Batasan CAD	16
2.3	Pemilihan Fitur		17
	2.3.1	Prosedur Pemilihan Fitur	18
	2.3.2	Kaedah Pemilihan Fitur	19
	2.3.3	Kriteria Pemilihan Fitur	21
	2.3.4	Pemilihan Fitur Pada Diagnosis Kanser Payudara	23
	2.3.5	Kaedah Pemilihan Fitur Berdasarkan Relevan, Pertindihan dan Interaksi	32
	2.3.6	Rumusan Isu Pemilihan Fitur	38
2.4	Analisis Kesatuan		39
	2.4.1	Algoritma Apriori	40
	2.4.2	Peraturan Kesatuan Kuat, Kelas dan Atomik	40
	2.4.3	Konsep Fitur Relevan, Bertindih dan Interaksi Berdasarkan Peraturan Kesatuan	41
	2.4.4	Perlombongan Peraturan Kesatuan Pada Penyelidikan Kanser Payudara	43
	2.4.5	Pemilihan Fitur Menggunakan Perlombongan Peraturan Kesatuan	49
2.5	Model Kecerdasan Buatan yang Boleh Dijelaskan		54
	2.5.1	Pendekatan Model XAI	55
	2.5.2	Penggunaan Model XAI dalam Penyelidikan Kanser Payudara	56
	2.5.3	Rumusan Isu Model XAI	57
2.6	Pengoptimuman Hiperparameter		58
	2.6.1	Pendekatan Pengoptimuman Hiperparameter	59
	2.6.2	Pengoptimuman Hiperparameter dalam Perlombongan Peraturan Kesatuan	61
	2.6.3	Rumusan Isu Pengoptimuman Hiperparameter	64
2.7	Kesimpulan		65
BAB III	**METODOLOGI KAJIAN**		
3.1	Pengenalan		66
3.2	Reka Bentuk Kaedah Kajian		66
3.3	Perisian dan Platform		72
3.4	Set Data Kajian		73
	3.4.1	Kanser Payudara Diagnostik Wisconsin (WDBC)	74
	3.4.2	Kanser Payudara Prognostik Wisconsin (WPBC)	75
	3.4.3	Kanser Payudara Coimbra (BCC)	76
3.5	Fasa 1: Pra-pemprosesan Data		77
	3.5.1	Pembersihan Data	78

	3.5.2 Penyeimbangan Data	78
	3.5.3 Penormalan Data	80
3.6	Fasa 2: Pemilihan Fitur Berinteraksi	82
	3.6.1 Pengenalpastian Calon Subset Fitur Melalui Penarafan Fitur Awal	82
	3.6.2 Pendiskretan Fitur Berterusan	86
	3.6.3 Set Kabur dan Fungsi Keahlian	87
	3.6.4 Prosedur Pendiskretan Kabur	88
	3.6.5 Pengekstrakan Subset Fitur Berasaskan Peraturan	89
3.7	Fasa 3: Pengesahan Fitur Berinteraksi	95
	3.7.1 Penilaian Kepentingan Fitur	96
	3.7.2 Penilaian Interaksi Fitur yang Kurang Relevan	98
3.8	Fasa 4: Pengoptimuman Hiperparameter Pemilihan Fitur	98
	3.8.1 Penalaan Hiperparameter Kaedah FD-CARI+SA	99
	3.8.2 Prosedur Algoritma Penyepuhlindapan Simulasi	99
3.9	Fasa 5: Pembinaan dan Penalaan Model Pengelasan	103
	3.9.1 Penyediaan Data bagi Pengujian	103
	3.9.2 Pengelasan Menggunakan Pembelajaran Mesin Terselia	103
	3.9.3 Penalaan Hiperparameter	105
3.10	Fasa 6: Penilaian Prestasi Model	106
	3.10.1 Penilaian Keberkesanan Model Pemilihan Fitur	106
	3.10.2 Pengesahan Fitur Berinteraksi	110
	3.10.3 Penilaian Keberkesanan Pengoptimuman Hiperparameter	110
3.11	Kesimpulan	111
BAB IV	**PEMILIHAN FITUR YANG RELEVAN, BERINTERAKSI DAN TIDAK BERTINDIH MENGGUNAKAN KAEDAH PERATURAN KESATUAN KELAS**	
4.1	Pengenalan	112
4.2	Hasil Pra-Pemprosesan Data dan Pengekstrakan Subset Fitur	112
	4.2.1 Pembersihan Data	112
	4.2.2 Penyeimbangan Data	113
	4.2.3 Penormalan Data	114
	4.2.4 Pengenalpastian Calon Subset Fitur Menggunakan Kaedah Penaraf	117
	4.2.4 Pendiskretan Kabur untuk Penyediaan Fitur dalam Perlombongan Peraturan	121

	4.2.5 Pengekstrakan Subset Fitur Relevan dan Berinteraksi	128
4.3	Hasil Model Pengelasan	155
4.4	Perbandingan Bilangan Subset Fitur Dipilih	164
4.5	Perbandingan Prestasi Model Pemilihan Fitur	167
	4.5.1 Set Data WDBC	167
	4.5.2 Set Data WPBC	173
	4.5.3 Set Data BCC	180
4.6	Ujian Signifikan	187
	4.6.1 Set Data WDBC	188
	4.6.2 Set Data WPBC	190
	4.6.3 Set Data BCC	191
4.7	Perbincangan	192
4.8	Kesimpulan	192
BAB V	**PENGESAHAN FITUR BERINTERAKSI BERDASARKAN NILAI SHAP DAN PENGUNDIAN MAJORITI BERWAJARAN**	
5.1	Pengenalan	193
5.2	Hasil Analisis Interaksi Fitur Berdasarkan Nilai SHAP	193
	5.2.1 Penilaian Awal Prestasi Model Pengelasan Sebelum Penjanaan Nilai SHAP	193
	5.2.2 Penjanaan Nilai Kepentingan Fitur Menggunakan SHAP	196
	5.2.3 Penjanaan Nilai Interaksi Fitur Menggunakan SHAP	208
	5.2.4 Penarafan Fitur Menggunakan Pengundian Majoriti Berwajaran	213
	5.2.5 Pengenalpastian Fitur Kurang Relevan Berdasarkan Nilai SHAP Berwajaran	218
	5.2.6 Penjanaan Nilai Interaksi SHAP bagi Fitur Kurang Relevan	220
5.3	Pengesahan Interaksi Fitur Kurang Relevan Berdasarkan Model SHAP	225
	5.2.1 Set Data WDBC	226
	5.2.2 Set Data WPBC	230
	5.2.3 Set Data BCC	235
5.4	Perbincangan	239
5.5	Kesimpulan	240

BAB VI	**PENAMBAHBAIKAN PRESTASI PEMILIHAN FITUR FD-CARI MELALUI PENGOPTIMUMAN HIPERPARAMETER BERASASKAN PENYEPUHLINDAPAN SIMULASI**	
6.1	Pengenalan	241
6.2	Penalaan Hiperparameter Menggunakan Algoritma Penyepuhlindapan Simulasi	242
6.3	Prestasi Model Pengelasan Berdasarkan Lelaran	248
6.4	Bilangan Subset Fitur Terpilih Berdasarkan Lelaran Terbaik	249
6.5	Perbandingan Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman Hiperparameter Menggunakan SA	250
	6.5.1 Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman bagi Set Data WDBC	251
	6.5.2 Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman bagi Set Data WPBC	252
	6.5.3 Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman bagi Set Data BCC	253
6.6	Perbandingan Model Pemilihan Fitur	254
	6.6.1 Perbandingan Prestasi FD-CARI+SA dengan Kaedah Pemilihan Fitur Piawai bagi Set Data WDBC	254
	6.6.2 Perbandingan Prestasi FD-CARI+SA dengan Kaedah Pemilihan Fitur Piawai bagi Set Data WPBC	255
	6.6.3 Perbandingan Prestasi FD-CARI+SA dengan Kaedah Pemilihan Fitur Piawai bagi Set Data BCC	257
6.7	Ujian Signifikan	258
	6.7.1 Keputusan Ujian Signifikan bagi Set Data WDBC	258
	6.7.2 Keputusan Ujian Signifikan bagi Set Data WPBC	260
	6.7.3 Keputusan Ujian Signifikan bagi Set Data BCC	261
6.8	Perbincangan	261
6.9	Kesimpulan	262
BAB VII	**KESIMPULAN DAN CADANGAN**	
7.1	Rumusan Kajian	263
7.2	Perbincangan dan Perbandingan Kritis	264
	7.2.1 Keunikan FD-CARI Berbanding Teknik Sedia Ada	264
	7.2.2 Perbandingan Pengesahan Interaksi Fitur	265

	7.2.3	Pengoptimuman FD-CARI Menggunakan Algoritma SA	265
	7.2.4	Perbandingan dengan Kajian Kesusasteraan	266
7.3	Implikasi Kajian		266
	7.3.1	Implikasi terhadap Teori	266
	7.3.2	Implikasi terhadap Metodologi	267
	7.3.3	Implikasi terhadap Amalan Praktikal	267
7.4	Sumbangan Kajian		268
	7.4.1	Sumbangan Teori	268
	7.4.2	Sumbangan Metodologi	269
	7.4.3	Sumbangan Praktikal	269
7.5	Keterbatasan dan Cadangan Kajian Lanjutan		269
7.6	Kesimpulan		270
RUJUKAN			272
Lampiran A	Hasil Prestasi Kaedah FD-CARI bagi Setiap Model Pengelasan		293
Lampiran B	Hasil Prestasi Kaedah FD-CARI+SA bagi Setiap Model Pengelasan dan Subset Fitur Akhir Bagi 20 Lelaran		298
Lampiran C	Senarai Penerbitan		304

SENARAI JADUAL

No. Jadual		Halaman
Jadual 2.1	Ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur pada pengelasan kanser payudara	29
Jadual 2.2	Ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur berdasarkan relevansi, pertindihan dan interaksi	36
Jadual 2.3	Ringkasan kajian kesusasteraan bagi kaedah ARM pada penyelidikan kanser payudara	48
Jadual 2.4	Ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur berdasarkan ARM	53
Jadual 2.5	Ringkasan kajian kesusasteraan bagi kaedah pengoptimuman hiperparameter ARM	63
Jadual 3.1	Senarai fitur bagi set data WDBC	74
Jadual 3.2	Kelas sasaran bagi set data WDBC	74
Jadual 3.3	Senarai fitur bagi set data WPBC	75
Jadual 3.4	Kelas sasaran bagi set data WPBC	76
Jadual 3.5	Senarai fitur bagi set data BCC	76
Jadual 3.6	Kelas sasaran bagi set data BCC	76
Jadual 3.7	Contoh sampel fitur *radius1* bagi set data WPBC	79
Jadual 3.8	Contoh sampel fitur *radius1* bagi set data WPBC	80
Jadual 3.9	Nilai fitur *radius1* selepas penormalan	81
Jadual 3.10	Kriteria pemangkasan bagi peraturan kesatuan kelas	93
Jadual 3.11	Pseudokod kaedah FD-CARI	94
Jadual 3.12	Pseudokod kaedah FD-CARI+SA	101
Jadual 3.13	Senarai hiperparameter bagi setiap model pengelasan	106
Jadual 3.14	Matriks kekeliruan	108
Jadual 4.1	Bilangan sampel sebelum dan selepas proses pembersihan data	113
Jadual 4.2	Bilangan sampel latihan sebelum dan selepas proses penyeimbangan data	113

Jadual 4.3	Nilai beberapa fitur terpilih sebelum dan selepas proses penormalan data bagi set data WDBC, WPBC dan BCC	114
Jadual 4.4	Calon fitur terpilih berdasarkan ambang SD bagi set data WDBC	118
Jadual 4.5	Calon fitur terpilih berdasarkan ambang SD bagi set data WPBC	119
Jadual 4.6	Calon fitur terpilih berdasarkan ambang SD bagi set data BCC	120
Jadual 4.7	Bilangan fitur sebelum dan selepas pengenalpastian calon subset fitur	121
Jadual 4.8	Julat nilai fitur terpilih sebelum dan selepas penyingkiran pencilan	123
Jadual 4.9	Set kabur bagi setiap contoh fitur	125
Jadual 4.10	Nilai kabur bagi setiap contoh fitur bagi sampel pertama	127
Jadual 4.11	Bilangan pencilan dan sampel yang diselaraskan bagi setiap contoh fitur	128
Jadual 4.12	Kategori kabur selepas pelarasan nilai pencilan	128
Jadual 4.13	Peraturan AR, CAR dan AAR bagi set data WDBC	129
Jadual 4.14	Peraturan AR, CAR dan AAR bagi set data WPBC	131
Jadual 4.15	Peraturan AR, CAR dan AAR bagi set data BCC	133
Jadual 4.16	Bilangan peraturan yang dipangkas dan senarai subset fitur berdasarkan α dan β bagi set data WDBC	136
Jadual 4.17	Bilangan peraturan yang dipangkas dan senarai subset fitur berdasarkan α dan β bagi set data WPBC	140
Jadual 4.18	Bilangan peraturan yang dipangkas dan senarai subset fitur berdasarkan α dan β bagi set data BCC	142
Jadual 4.19	Bilangan RFset sebelum dan selepas penyingkiran fitur bertindih bagi set data WDBC	144
Jadual 4.20	Bilangan RFset sebelum dan selepas penyingkiran fitur bertindih bagi set data WPBC	147
Jadual 4.21	Bilangan RFset sebelum dan selepas penyingkiran fitur bertindih bagi set data BCC	149
Jadual 4.22	Subset fitur unik berdasarkan α dan β bagi set data WDBC	150

Jadual 4.23	Subset fitur unik berdasarkan α dan β bagi set data WPBC	151
Jadual 4.24	Subset fitur unik berdasarkan α dan β bagi set data BCC	152
Jadual 4.25	Skor AUC bagi setiap model pengelasan bagi set data WDBC	153
Jadual 4.26	Skor AUC bagi setiap model pengelasan bagi set data WPBC	154
Jadual 4.27	Skor AUC bagi setiap model pengelasan bagi set data BCC	154
Jadual 4.28	Nilai skor metrik penilaian setiap model pengelasan bagi set data WDBC	155
Jadual 4.29	Nilai skor metrik penilaian setiap model pengelasan bagi set data WPBC	159
Jadual 4.30	Nilai skor metrik penilaian bagi setiap model pengelasan bagi set data BCC	162
Jadual 4.31	Subset fitur terpilih oleh kaedah pemilihan fitur bagi set data WDBC	165
Jadual 4.32	Subset fitur terpilih oleh kaedah pemilihan fitur bagi set data WPBC	165
Jadual 4.33	Subset fitur terpilih oleh kaedah pemilihan fitur bagi set data BCC	166
Jadual 4.34	Nilai skor AUC bagi setiap model pengelasan	167
Jadual 4.35	Nilai skor ACC bagi setiap model pengelasan	169
Jadual 4.36	Nilai skor SEN bagi setiap model pengelasan	170
Jadual 4.37	Nilai skor SPE bagi setiap model pengelasan	171
Jadual 4.38	Nilai skor F1 bagi setiap model pengelasan	172
Jadual 4.39	Nilai skor AUC bagi setiap model pengelasan	174
Jadual 4.40	Nilai skor ACC bagi setiap model pengelasan	176
Jadual 4.41	Nilai skor SEN bagi setiap model pengelasan	177
Jadual 4.42	Nilai skor SPE bagi setiap model pengelasan	178
Jadual 4.43	Nilai skor F1 bagi setiap model pengelasan	179
Jadual 4.44	Nilai skor AUC bagi setiap model pengelasan	181
Jadual 4.45	Nilai skor ACC bagi setiap model pengelasan	183

Jadual 4.46	Nilai skor SEN bagi setiap model pengelasan	184
Jadual 4.47	Nilai skor SPE bagi setiap model pengelasan	185
Jadual 4.48	Nilai skor F1 bagi setiap model pengelasan	186
Jadual 4.49	Skor AUC bagi setiap model pengelasan	188
Jadual 4.50	Hasil ujian signifikan pada kaedah pemilihan fitur terbaik	189
Jadual 4.51	Hasil ujian signifikan pada selain kaedah pemilihan fitur terbaik	189
Jadual 4.52	Skor AUC bagi setiap model pengelasan	190
Jadual 4.53	Hasil ujian signifikan pada kaedah pemilihan fitur terbaik	190
Jadual 4.54	Skor AUC bagi setiap model pengelasan	191
Jadual 4.55	Hasil ujian signifikan pada kaedah pemilihan fitur terbaik	191
Jadual 5.1	Skor AUC bagi setiap model pengelasan bagi set data WDBC	194
Jadual 5.2	Purata nilai SHAP bagi setiap model pengelasan bagi set data WDBC	197
Jadual 5.3	Purata nilai SHAP bagi setiap model pengelasan bagi set data WPBC	201
Jadual 5.4	Purata nilai mutlak SHAP bagi setiap model pengelasan bagi set data BCC	205
Jadual 5.5	Pasangan fitur dengan purata nilai interaksi SHAP bagi set data WDBC	209
Jadual 5.6	Pasangan fitur dengan purata nilai interaksi SHAP bagi set data WPBC	210
Jadual 5.7	Pasangan fitur dan purata nilai interaksi SHAP bagi set data BCC	212
Jadual 5.8	Kedudukan fitur berdasarkan purata nilai SHAP bagi set data WDBC	213
Jadual 5.9	Kedudukan fitur berdasarkan nilai SHAP berwajaran bagi set data WDBC	214
Jadual 5.10	Kedudukan fitur berdasarkan purata nilai mutlak SHAP bagi set data WPBC	215
Jadual 5.11	Kedudukan fitur berdasarkan nilai SHAP berwajaran bagi set data WPBC	216

Jadual 5.12	Kedudukan fitur berdasarkan purata nilai mutlak SHAP bagi set data BCC	217
Jadual 5.13	Kedudukan fitur berdasarkan nilai SHAP berwajaran bagi set data BCC	218
Jadual 5.14	Senarai fitur kurang relevan pada set data WDBC	218
Jadual 5.15	Senarai fitur kurang relevan pada set data WPBC	219
Jadual 5.16	Senarai fitur kurang relevan pada set data BCC	220
Jadual 5.17	Pasangan fitur kurang relevan dan purata nilai interaksi SHAP bagi set data WDBC	220
Jadual 5.18	Pasangan fitur kurang relevan dan purata nilai interaksi SHAP bagi set data WPBC	222
Jadual 5.19	Pasangan fitur kurang relevan dan purata nilai interaksi SHAP bagi set data BCC	224
Jadual 5.20	Tahap relevansi fitur terpilih bagi setiap set data	225
Jadual 5.21	Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-75 dalam set data WDBC	228
Jadual 5.22	Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-75 dalam set data WPBC	232
Jadual 5.23	Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-75 dalam set data BCC	237
Jadual 5.24	Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-50 dalam set data BCC	238
Jadual 6.1	Hasil lelaran proses pengoptimuman hiperparameter bagi set data WDBC	243
Jadual 6.2	Hasil lelaran proses pengoptimuman hiperparameter bagi set data WPBC	245
Jadual 6.3	Hasil lelaran proses pengoptimuman hiperparameter bagi set data BCC	247
Jadual 6.4	5 lelaran dengan skor AUC tertinggi bagi set data WDBC	248
Jadual 6.5	5 lelaran dengan skor AUC tertinggi bagi set data WPBC	249
Jadual 6.6	5 lelaran dengan skor AUC tertinggi bagi set data BCC	249
Jadual 6.7	Bilangan fitur selepas pemurnian subset fitur bagi set data WDBC	250

Jadual 6.8	Bilangan fitur selepas pemurnian subset fitur bagi set data WPBC	250
Jadual 6.9	Bilangan fitur selepas pemurnian subset fitur bagi set data BCC	250
Jadual 6.10	Perbandingan prestasi bagi set data WDBC	251
Jadual 6.11	Purata skor AUC bagi setiap model pengelasan bagi set data WDBC	251
Jadual 6.12	Perbandingan prestasi bagi set data WPBC	252
Jadual 6.13	Purata skor AUC bagi setiap model pengelasan bagi set data WPBC	252
Jadual 6.14	Perbandingan prestasi bagi set data BCC	253
Jadual 6.15	Purata skor AUC bagi setiap model pengelasan bagi set data BCC	253
Jadual 6.16	Skor AUC setiap kaedah pemilihan fitur bagi set data WDBC	255
Jadual 6.17	Skor AUC setiap kaedah pemilihan fitur bagi set data WPBC	256
Jadual 6.18	Skor AUC setiap kaedah pemilihan fitur bagi set data BCC	257
Jadual 6.19	Skor AUC bagi setiap model pengelasan bagi set data WDBC	259
Jadual 6.20	Hasil ujian signifikan pada kaedah pemilihan fitur terbaik	259
Jadual 6.21	Hasil ujian signifikan pada selain kaedah pemilihan fitur terbaik	260
Jadual 6.22	Skor AUC bagi setiap model pengelasan bagi set data WPBC	260
Jadual 6.23	Hasil ujian signifikan pada kaedah pemilihan fitur terbaik	260
Jadual 7.1	Perbandingan model FD-CARI dengan kajian kesusasteraan	266
Jadual 7.2	Pemetaan objektif, dapatan utama dan sumbangan kajian	268

SENARAI ILUSTRASI

No. Rajah		Halaman
Rajah 2.1	Carta alir sistem CADe dan CADx (Diadaptasi daripada Hassan et al. (2022))	16
Rajah 2.2	Prosedur pemilihan fitur (Sumber: Dash & Liu (1997))	18
Rajah 2.3	Hubungan relevansi dan pertindihan antara fitur (Al Nuaimi et al. 2019)	22
Rajah 3.1	Aliran metodologi kajian	67
Rajah 3.2	Reka bentuk kaedah kajian	71
Rajah 3.3	Contoh imej FNA (Mangasarian et al., 2008)	73
Rajah 3.4	Proses pra-pemprosesan data	78
Rajah 3.5	Proses pemilihan fitur	82
Rajah 3.6	Carta alir bagi proses pengenalpastian calon subset fitur melalui penarafan fitur awal	86
Rajah 3.7	Carta alir proses pendiskretan kabur	89
Rajah 3.8	Carta alir proses pengekstrakan subset fitur	90
Rajah 3.9	Proses pengesahan fitur berinteraksi menggunakan nilai SHAP dan pengundian majoriti berwajaran	96
Rajah 3.10	Carta alir algoritma penyepuhlindapan simulasi	101
Rajah 4.1	Set data (a) WDBC (b) WPBC (c) BCC sebelum dan selepas penyeimbangan data menggunakan SMOTE	114
Rajah 4.2	Perbandingan plot kotak bagi nilai fitur sebelum dan selepas penormalan bagi set data (a) WDBC (b) WPBC (c) BCC	117
Rajah 4.3	Plot kotak sebelum dan selepas penyingkiran pencilan bagi fitur (a) *radius_mean* (b) *area_mean* (c) *symmetry1* (d) *fractal_dimension1* (e) *Glucose* (f) *Insulin*	124
Rajah 4.4	Plot fungsi keahlian segi tiga bagi fitur (a) *radius_mean* (b) *area_mean* (c) *symmetry1* (d) *fractal_dimension1* (e) *Glucose* (f) *Insulin*	126
Rajah 4.5	Bilangan (a) AR (b) CAR (c) AAR berdasarkan α dan β pada WDBC	130

Rajah 4.6	Bilangan (a) AR (b) CAR (c) AAR berdasarkan α dan β pada WPBC	132
Rajah 4.7	Bilangan (a) AR (b) CAR (c) AAR berdasarkan α dan β pada BCC	134
Rajah 4.8	Purata matriks kekeliruan bagi model pengelasan (a) DT (b) RF (c) NB (d) LR (e) SVM bagi set data WDBC	157
Rajah 4.9	Purata lengkung ROC bagi setiap model pengelasan bagi set data WDBC	158
Rajah 4.10	Purata matriks kekeliruan bagi model pengelasan (a) DT (b) RF (c) NB (d) LR (e) SVM bagi set data WPBC	160
Rajah 4.11	Purata lengkung ROC bagi keseluruhan model pengelasan bagi set data WPBC	161
Rajah 4.12	Purata matriks kekeliruan bagi model pengelasan (a) DT (b) RF (c) NB (d) LR (e) SVM bagi set data BCC	163
Rajah 4.13	Purata lengkung ROC bagi keseluruhan model pengelasan bagi set data BCC	164
Rajah 4.14	Purata lengkungan ROC bagi model (a) DT (b) RF (c) NB (d) LR (e) SVM	168
Rajah 4.15	Perbandingan purata skor metrik penilaian bagi keseluruhan model pengelasan	173
Rajah 4.16	Purata lengkungan ROC bagi model (a) DT (b) RF (c) NB (d) LR (e) SVM	175
Rajah 4.17	Perbandingan purata skor metrik penilaian bagi keseluruhan model pengelasan	180
Rajah 4.18	Purata lengkungan ROC bagi model (a) DT (b) RF (c) NB (d) LR (e) SVM	182
Rajah 4.19	Perbandingan purata skor metrik penilaian bagi keseluruhan model pengelasan	187
Rajah 5.1	Purata lengkung ROC bagi set data WDBC	195
Rajah 5.2	Purata lengkung ROC bagi set data WPBC	195
Rajah 5.3	Purata lengkung ROC bagi set data BCC	196
Rajah 5.4	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model RF	198

Rajah 5.5	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model GB	199
Rajah 5.6	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model XGB	200
Rajah 5.7	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model RF	202
Rajah 5.8	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model GB	203
Rajah 5.9	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model XGB	204
Rajah 5.10	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model RF	206
Rajah 5.11	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model GB	207
Rajah 5.12	Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model XGB	208
Rajah 5.13	Graf rangkaian bagi 30 pasangan fitur berinteraksi tertinggi dalam set data WDBC	210
Rajah 5.14	Graf rangkaian bagi 30 pasangan fitur berinteraksi tertinggi bagi set data WPBC	211
Rajah 5.15	Graf rangkaian bagi 30 pasangan fitur berinteraksi tertinggi bagi set data BCC	212
Rajah 5.16	Graf rangkaian bagi 30 pasangan fitur kurang relevan bagi set data WDBC	222
Rajah 5.17	Graf rangkaian bagi 30 pasangan fitur kurang relevan bagi set data WPBC	223
Rajah 5.18	Graf rangkaian bagi 30 pasangan fitur kurang relevan bagi set data BCC	224
Rajah 5.19	Peta haba interaksi bagi (a) keseluruhan (b) kuantil ke-75 (c) fitur kurang relevan bagi set data WDBC	227
Rajah 5.20	Bilangan pasangan fitur kurang relevan pada kuantil ke-75 bagi set data WDBC	230
Rajah 5.21	Peta haba interaksi bagi (a) keseluruhan (b) kuantil ke-75 (c) fitur kurang relevan bagi set data WPBC	232

Rajah 5.22	Bilangan pasangan fitur kurang relevan pada kuantil ke-75 bagi set data WPBC	235
Rajah 5.23	Peta haba interaksi bagi (a) keseluruhan (b) kuantil ke-75 (c) fitur kurang relevan bagi set data BCC	236
Rajah 5.24	Peta haba interaksi bagi (a) kuantil ke-50 (b) fitur kurang relevan bagi set data BCC	238
Rajah 5.25	Bilangan pasangan fitur kurang relevan pada kuantil ke-50 bagi set data BCC	239
Rajah 6.1	Perbandingan skor AUC bagi set data WDBC	255
Rajah 6.2	Perbandingan skor AUC bagi set data WPBC	256
Rajah 6.3	Perbandingan skor AUC bagi set data BCC	258

SENARAI SINGKATAN

AAR	Peraturan kesatuan atomik
ACC	Ketepatan
ARM	Perlombongan Peraturan Kesatuan
AUC	Luas di bawah lengkung
CAD	Diagnosis berbantu komputer
CAR	Peraturan kesatuan kelas
CARM	Perlombongan Peraturan Kesatuan Kelas
CFS	*Correlation-based Feature Selection*
DT	Pokok Keputusan
FCBF	*Fast Correlation-based Filter*
FD-CARI	Pendiskretan Kabur–Peraturan Kesatuan Kelas Berinteraksi
GB	*Gradient Boost*
GS	*Grid Search*
LR	Regresi Logistik
mRMR	*Minimum Redundancy Maximum Relevance*
NB	Bayes Naif
RF	Hutan Rawak
SA	Penyepuhlindapan Simulasi
SD	Sisihan Piawai
SEN	Kepekaan
SHAP	Penjelasan Tambahan Shapley
SPE	Kekhususan
SVM	Mesin Vektor Sokongan
XAI	Kecerdasan Buatan yang Boleh Dijelaskan
XGB	*Extreme Gradient Boost*

SENARAI ISTILAH

Ambang	Threshold
Anteseden	Antecedent
Entropi	Entropy
Hingar	Noise
Kebolehfahaman	Understandable
Kebolehtafsiran	Interpretability
Kebolehjelasan	Explainability
Konsekuen	Consequent
Pemangkasan	Pruning
Pembelajaran mesin	Machine learning
Pencilan	Outlier
Plot kotak	Box plot
Terlebih padan	Over fitting

BAB I

PENDAHULUAN

1.1 PENGENALAN

Bab ini memberikan pengenalan kepada kajian yang telah dijalankan dengan memfokuskan kepada cabaran dalam pengelasan kanser payudara, keperluan pemilihan fitur yang berkesan serta potensi pendekatan gabungan seperti Perlombongan Peraturan Kesatuan Kelas (CARM), Kecerdasan Buatan yang Boleh Dijelaskan (*eXplainable Artificial Intelligent*, XAI) dan pengoptimuman hiperparameter. Perbincangan merangkumi latar belakang kanser payudara, sistem Diagnosis Berbantu Komputer (CAD), isu pemilihan fitur dan keperluan pemodelan yang cekap, telus dan boleh dipercayai dalam konteks klinikal. Selain itu, bab ini turut membincangkan kenyataan masalah, objektif kajian, skop dan batasan kajian, ringkasan sumbangan kajian serta struktur keseluruhan tesis.

1.2 LATAR BELAKANG KAJIAN

Kanser payudara merupakan antara isu kesihatan awam yang paling kritikal dengan jutaan wanita terjejas setiap tahun (Kim et al. 2025). Ia bukan sahaja merupakan kanser yang paling kerap didiagnosis di kalangan wanita, malah menjadi penyebab utama kematian berkaitan kanser (Michaels et al. 2023). Kadar kelangsungan hidup yang lebih rendah sering dikaitkan dengan kelewatan dalam proses diagnosis dan pengesanan awal (Barrios 2022). Justeru, pengesanan awal dan pengelasan yang tepat adalah amat penting bagi meningkatkan keberkesanan rawatan dan mengurangkan risiko kematian (Liew et al. 2021; Lim et al. 2022).

Kemajuan dalam teknologi perubatan telah menghasilkan pelbagai kaedah pengimejan dan sistem pengumpulan data yang canggih yang telah menghasilkan set data berdimensi tinggi dan kompleks. Ini secara langsung mewujudkan keperluan untuk teknik analisis data yang lebih efisien bagi menyaring maklumat penting dan membina model pengelasan yang boleh dipercayai. Sistem CAD telah diperkenalkan sebagai teknologi sokongan kepada pakar perubatan dalam mentafsir imej dan maklumat pesakit secara automatik. Dalam pembinaan sistem CAD, pemilihan fitur memainkan peranan utama dalam menentukan prestasi model pembelajaran mesin. Fitur yang tidak relevan, bertindih atau berlebihan bukan sahaja menjejaskan ketepatan model, tetapi juga meningkatkan kerumitan pengiraan dan menyukarkan interpretasi (Chhillar & Singh 2025; Li et al. 2017). Walaupun terdapat pelbagai kaedah pemilihan fitur moden yang diperkenalkan bagi set data perubatan, kebanyakannya masih tertumpu kepada pemilihan berasaskan kedudukan atau kepentingan individu tanpa menilai kestabilan subset fitur atau makna klinikal fitur yang dipilih (Wang et al. 2025).

Tambahan pula, kebanyakan kaedah pemilihan fitur tradisional hanya menilai hubungan antara satu fitur dengan kelas sasaran secara individu, tanpa mengambil kira interaksi antara fitur yang berpotensi memberikan nilai diagnostik yang lebih tinggi (Jakulin & Bratko 2004). Interaksi antara fitur merujuk kepada situasi di mana gabungan dua atau lebih fitur menghasilkan kesan yang lebih bermakna terhadap kelas sasaran berbanding analisis secara bersendirian. Kaedah pengoptimuman seperti *Red Fox Optimization* dan *Elephant Herding Optimization* juga dilaporkan berkesan dalam meningkatkan prestasi model, namun masih gagal menangkap hubungan interaksi antara fitur serta tidak menyediakan mekanisme interpretasi yang jelas (Khanna et al. 2024; Thatha et al. 2025). Oleh itu, kaedah pemilihan fitur yang mampu mengenal pasti fitur yang relevan, tidak bertindih dan berinteraksi amat diperlukan untuk menghasilkan model pengelasan yang lebih tepat.

Kaedah Perlombongan Peraturan Kesatuan Kelas (CARM) telah diperkenalkan sebagai salah satu pendekatan yang sesuai untuk mengenal pasti hubungan antara fitur dan kelas secara eksplisit (Subasi 2020). Kaedah CARM berupaya untuk mengekstrak peraturan berasaskan gabungan fitur dan mengenal pasti pola tersembunyi yang tidak dapat dikesan (Bartschat et al. 2019) oleh kaedah pemilihan fitur tradisional seperti

teknik penapis. Namun, perlaksanaan kaedah CARM secara optimum memerlukan penalaan pelbagai hiperparameter seperti ambang sokongan minimum dan keyakinan minimum yang secara konvensional dilakukan secara manual (Manikandan & Selvakumar 2022). Bagi menangani isu ini, algoritma Penyepuhlindapan Simulasi (SA) telah diperkenalkan sebagai pendekatan metaheuristik yang efisien untuk meneroka ruang carian hiperparameter secara automatik dan menyesuaikan diri dengan fungsi objektif (Nikolaev & Jacobson 2010). Algoritma SA meniru proses penyejukan logam dalam metalurgi dan sesuai untuk tugas pengoptimuman tidak linear seperti penalaan kaedah CARM.

Selain itu, penyelidikan kanser payudara sering menekankan ketepatan pengelasan semata-mata tanpa mengesahkan kesesuaian dan kepentingan fitur yang dipilih. Bagi memastikan bahawa fitur yang dipilih benar-benar berinteraksi dan menyumbang kepada pengelasan yang tepat, proses pengesahan fitur adalah penting (Shi et al. 2023). Dalam kajian ini, model Kecerdasan Buatan yang Boleh Dijelaskan (XAI) melalui teknik *SHapley Additive exPlanations* (SHAP) (Lundberg & Lee 2017) digunakan bagi menilai kepentingan dan interaksi antara fitur secara kuantitatif dan visual (Kok et al. 2023). Kekangan-kekangan ini mengukuhkan keperluan kepada kaedah pemilihan fitur yang bersifat interaksi, stabil dan boleh dijelaskan secara klinikal. Kaedah ini turut disokong oleh teknik pengundian majoriti berwajaran untuk mengukuhkan kebolehpercayaan dalam proses pengesahan akhir.

Secara keseluruhannya, latar belakang kajian ini menunjukkan bahawa terdapat kekangan dan keperluan mendesak terhadap satu kerangka pemilihan fitur berasaskan interaksi yang menyeluruh yang bukan sahaja memilih fitur yang bermakna secara statistik dan klinikal, tetapi juga mudah dijelaskan serta boleh dioptimumkan secara automatik.

1.3 PERSOALAN KAJIAN

a. Bagaimanakah model pemilihan fitur berasaskan peraturan boleh mengenalpasti subset fitur yang relevan, berinteraksi dan tidak bertindih bagi pengelasan kanser payudara?

b. Bagaimanakah model XAI dapat mengesahkan kepentingan dan interaksi antara fitur yang dipilih?

c. Bagaimanakah algoritma Penyepuhlindapan Simulasi (SA) dapat mengoptimumkan hiperparameter dalam pemilihan fitur dan meningkatkan prestasi pengelasan?

1.4 KENYATAAN MASALAH

Walaupun banyak kajian telah dijalankan dalam bidang pengelasan kanser payudara, masih terdapat beberapa cabaran utama yang belum ditangani sepenuhnya. Pertama, kebanyakan teknik pemilihan fitur sedia ada hanya menumpukan kepada hubungan satu-ke-satu antara fitur dan kelas sasaran tanpa mengambil kira interaksi antara fitur (Wan et al. 2021). Walaupun sesuatu fitur kelihatan tidak berkait secara langsung dengan kelas secara individu, apabila ia berinteraksi dengan fitur lain, gabungan tersebut berkemungkinan mempunyai perkaitan yang signifikan dengan kelas sasaran (Song et al., 2019; Wan et al., 2021). Ini menjadi satu kekurangan kepada kebanyakan teknik pemilihan fitur sedia ada yang menyebabkan maklumat penting yang tersirat dalam gabungan fitur berpotensi diabaikan, seterusnya menjejaskan ketepatan model (Guo et al. 2024; Vijayasarveswari et al. 2020; Wenjing Wang et al. 2023).

Sebagai contoh, dalam set data *Breast Cancer Coimbra* (BCC), fitur-fitur seperti *Body Mass Index* (BMI) dan *Glucose* mungkin secara individu memberikan beberapa maklumat tentang kebarangkalian penyakit kanser. Walau bagaimanapun, apabila fitur-fitur ini dipertimbangkan bersama, ia berpotensi memberikan petunjuk yang lebih kukuh tentang tahap kanser kerana peningkatan kadar glukosa dalam pesakit yang mempunyai BMI yang lebih tinggi mungkin lebih signifikan dalam menunjukkan risiko kanser payudara.

Oleh itu, kebergantungan kepada teknik pemilihan fitur yang mengabaikan interaksi antara fitur boleh menyebabkan model pengelasan kehilangan maklumat yang penting terutamanya dalam bidang yang kritikal seperti diagnosis kanser payudara. Tambahan pula, kebanyakan teknik sedia ada tidak menawarkan mekanisme penjelasan yang mencukupi bagi menyokong keputusan model, sekali gus menyukarkan

pemahaman dan kepercayaan pengguna bukan teknikal seperti pakar perubatan. Justeru, kajian ini penting untuk membangunkan satu kaedah pemilihan fitur berasaskan interaksi yang bukan sahaja tepat tetapi juga telus dan boleh dijelaskan agar dapat menyumbang ke arah pembangunan model pengelasan yang lebih boleh diguna pakai dalam konteks klinikal sebenar.

Kedua, proses pemilihan fitur yang rumit, bersifat kotak hitam (*black box*) dan sukar dijelaskan menyukarkan pemahaman serta penerimaan dalam kalangan pakar klinikal. Dalam bidang kesihatan, terdapat keperluan yang mendesak terhadap pembangunan model pengelasan yang boleh dijelaskan (*explainable*) bagi membolehkan keputusan disokong dengan bukti dan logik yang boleh difahami oleh pengguna bukan teknikal (Abrantes & Rouzrokh 2024; Shakhovska et al. 2024). Walaupun model pembelajaran mesin mampu mencapai ketepatan tinggi, kekurangan dari segi kebolehjelasan menimbulkan cabaran dalam penggunaannya untuk sokongan keputusan klinikal. Justeru, kajian ini mencadangkan mekanisme pengesahan fitur dua lapis yang menggabungkan pendekatan pengundian majoriti berwajaran dan model XAI seperti SHAP untuk mengenal pasti serta mengesahkan interaksi antara fitur yang dipilih melalui FD-CARI. Pengundian majoriti bertujuan memastikan kestabilan pemilihan fitur merentas pelbagai model pengelasan, manakala SHAP memberikan penjelasan kuantitatif dan visual terhadap sumbangan dan interaksi setiap fitur terhadap keputusan model. Pendekatan ini bukan sahaja dapat meningkatkan ketepatan dan kestabilan pengelasan model, malah dapat meningkatkan kebolehpercayaan dan penerimaan pengguna khususnya dalam CAD dan sistem sokongan keputusan klinikal.

Ketiga, walaupun kaedah pemilihan fitur berasaskan peraturan seperti kaedah CARM mempunyai kelebihan dalam mengenal pasti interaksi fitur, ia masih berhadapan kekangan dari segi penalaan hiperparameter yang dilakukan secara manual dan tidak sistematik**.** Nilai ambang seperti sokongan minimum, keyakinan, faktor kepastian dan lain-lain memainkan peranan penting dalam menentukan kualiti peraturan yang dihasilkan (Aljehani & Alotaibi 2024; Geng et al. 2025). Namun, pemilihan nilai ambang yang tidak sesuai boleh menyebabkan peraturan yang terlalu banyak atau terlalu sedikit, peraturan yang tidak relevan atau terlalu umum, serta kegagalan dalam mencerminkan hubungan sebenar antara fitur dan kelas sasaran.

Masalah pemilihan gabungan hiperparameter terbaik ini merupakan satu cabaran besar dalam menjamin prestasi optimum model terutamanya apabila ruang carian adalah luas dan bergantung kepada sifat set data. Pencarian secara manual bukan sahaja memakan masa, malah berisiko menghasilkan model yang tidak stabil dan kurang cekap (Aljehani & Alotaibi 2024) . Oleh itu, kajian ini mencadangkan penggunaan algoritma SA sebagai teknik pengoptimuman hiperparameter untuk mengatasi cabaran ini. Algoritma SA berupaya mengeksplorasi ruang carian secara meluas pada peringkat awal dan kemudiannya menyempitkan carian ke arah penyelesaian yang lebih baik. Algoritma ini dapat meningkatkan kecekapan proses penalan hiperparameter dan membantu kaedah FD-CARI mencapai prestasi pengelasan yang lebih stabil dan tepat.

Akhir sekali, kurangnya pendekatan menyeluruh yang menggabungkan teknik pemilihan fitur berasaskan interaksi, pengesahan menggunakan model XAI dan pengoptimuman automatik hiperparameter menjadikan pembinaan model pengelasan yang kukuh dan boleh dipercayai masih menjadi cabaran utama. Oleh itu, kajian ini berhasrat untuk membangunkan satu kerangka kerja hibrid yang dikenali sebagai FD-CARI+SA yang menggabungkan kelebihan kaedah CARM, algoritma SA dan model XAI iaitu SHAP bagi menangani kekangan ini secara menyeluruh.

1.5 OBJEKTIF KAJIAN

Kajian ini dijalankan dengan matlamat utama untuk membangunkan satu kerangka pemilihan fitur berasaskan interaksi yang boleh dijelaskan dan dioptimumkan secara automatik bagi pengelasan kanser payudara. Secara khusus, objektif kajian ini adalah seperti berikut:

a. Merumuskan model pemilihan fitur berasaskan kaedah CARM iaitu FD-CARI yang dapat mengenal pasti subset fitur yang relevan, berinteraksi dan tidak bertindih.

b. Mereka bentuk kerangka pengesahan fitur berinteraksi terpilih menggunakan model XAI berasaskan teknik SHAP dan kaedah pengundian majoriti berwajaran untuk meningkatkan ketelusan dan kebolehpercayaan kaedah FD-CARI.

c. Membangunkan kaedah hibrid FD-CARI+SA yang menggabungkan kaedah CARM dengan algoritma SA bagi penalaan hiperparameter secara automatik dan peningkatan prestasi model pengelasan kanser payudara.

1.6 SKOP KAJIAN

Skop kajian ini tertumpu kepada pembangunan dan penilaian kaedah pemilihan fitur berasaskan interaksi yang boleh dijelaskan dan dioptimumkan secara automatik untuk tujuan pengelasan kanser payudara. Ruang lingkup kajian ini merangkumi proses bermula daripada perolehan data, pembangunan algoritma, pengesahan fitur, pengoptimuman hiperparameter dan penilaian prestasi model seperti berikut:

1.6.1 Jenis Data

Kajian ini menggunakan tiga set data terbuka berkaitan kanser payudara iaitu:

- Kanser Payudara Diagnostik Wisconsin (WDBC) (Wolberg, Mangasarian, et al. 1995).
- Kanser Payudara Prognostik Wisconsin (WPBC) (Wolberg, Street, et al. 1995).
- Kanser Payudara Coimbra (BCC) (Patrcio et al. 2018).

1.6.2 Fokus Teknikal

Kajian ini menumpukan kepada pembangunan kaedah pemilihan fitur terbaru yang menggabungkan dua komponen utama iaitu:

- Pendiskretan Kabur: Model Logik Kabur digunakan untuk mendiskretkan fitur berterusan dalam set data kanser. Kaedah ini lebih baik kerana nilai fitur diwakili dalam julat [0,1] berbanding hanya 0 atau 1 (Zadeh 1965). Ini membolehkan hubungan antara fitur dan label kelas dikenal pasti dengan lebih tepat.

- CARM: Kaedah ini digunakan untuk mengh asilkan peraturan yang menghubungkan fitur dengan kelas sasaran. CARM membantu mengenal pasti fitur penting dan berinteraksi yang benar-benar menyumbang kepada proses pengelasan.

1.6.3 Pengesahan Interaksi Fitur

Fitur yang dipilih akan disahkan melalui dua pendekatan iaitu:

- Model SHAP untuk penilaian sumbangan dan interaksi fitur.
- Pengundian majoriti berwajaran melibatkan tiga model pembelajaran mesin iaitu Hutan Rawak (RF), *Gradient Boosting* (GB) dan *Extreme Gradient Boosting* (XGB). Kaedah ini meningkatkan kebolehpercayaan hasil dengan menggabungkan kekuatan ketiga-tiga model.

1.6.4 Pengoptimuman Hiperparameter

Kaedah FD-CARI yang dibangunkan dioptimumkan menggunakan algoritma SA iaitu teknik metaheuristik yang berkesan untuk meneroka ruang hiperparameter secara menyeluruh serta meningkatkan prestasi model dengan lebih cekap.

1.6.5 Penilaian Prestasi Model

Penilaian dilakukan menggunakan lima model pembelajaran mesin piawai bagi tugas pengelasan iaitu:

- Pokok Keputusan (DT)
- Hutan Rawak (RF)
- Bayes Naif (NB)
- Regresi Logistik (LR)
- Mesin Sokongan Vektor (SVM)

Bagi menguji gabungan kaedah pendiskretan kabur dan kaedah CARM yang dibangunkan, pengujian dilakukan dengan teknik pemilihan fitur piawai iaitu CFS, FCBF, Consistency, mRMR dan Relief-F. Model pengesahan bagi fitur berinteraksi yang menggunakan model SHAP serta pengundian majoriti berwajaran pula dibandingkan dengan fitur berinteraksi yang dipilih oleh kaedah FD-CARI. Penilaian

adalah berdasarkan lima metrik penilaian iaitu luas di bawah lengkung (AUC), ketepatan (ACC), kepekaan (SEN), kekhususan (SPE) dan skor F1. Kaedah pengoptimuman menggunakan algoritma SA yang dibangunkan juga dibandingkan dan dinilai dengan teknik pemilihan fitur piawai tersebut.

1.7 BATASAN KAJIAN

Kajian ini bersifat kuantitatif dan eksperimen serta tidak melibatkan data klinikal sebenar mahupun pengesahan lapangan secara klinikal bersama pakar perubatan. Fokus adalah kepada keberkesanan kaedah dari sudut teknikal dan kebolehjelasan menggunakan data sekunder. Kajian ini tidak menggunakan data perubatan tempatan daripada hospital atau institusi kesihatan dalam negara disebabkan oleh beberapa kekangan praktikal dan teknikal. Antaranya termasuk keperluan mendapatkan kelulusan etika penyelidikan perubatan yang ketat dan isu sensitiviti serta kerahsiaan pesakit. Di samping itu, data perubatan tempatan masih belum mencapai tahap kematangan dari segi keterbukaan, piawaian dan kebolehcapaian untuk penyelidikan berbanding set data terbuka antarabangsa seperti WDBC, WPBC dan BCC. Ketiadaan dokumentasi metadata yang lengkap, variasi dalam struktur data serta kekurangan repositori data berskala besar menjadikan pemprosesan dan pengesahan kajian menggunakan data tempatan satu cabaran besar. Oleh itu, kajian ini memilih untuk menggunakan set data terbuka yang telah diiktiraf dan banyak digunakan dalam komuniti penyelidikan sebagai langkah praktikal untuk membangunkan, menguji dan mengesahkan algoritma secara sistematik.

Selain itu, batasan kajian lain ialah perbandingan prestasi model pengelasan hanya pada lima kaedah pemilihan fitur piawai termasuk CFS, FCBF, Consistency, Relief-F dan mRMR. Kaedah CFS, FCBF dan mRMR menilai relevansi dan pertindihan fitur, manakala Consistency hanya menilai relevansi tanpa mengambil kira pertindihan atau interaksi. Relief-F pula menilai relevansi dan interaksi antara fitur. Namun, pemilihan kaedah-kaedah ini membolehkan perbandingan yang menyeluruh bagi setiap sudut kriteria pemilihan fitur.

1.8 RINGKASAN SUMBANGAN KAJIAN

Kajian ini menyumbang secara signifikan kepada bidang sains data dan pengelasan perubatan terutamanya dalam pembangunan algoritma pemilihan fitur yang bersifat berinteraksi, boleh dijelaskan dan dioptimumkan secara automatik. Sumbangan utama kajian ini terletak pada pembangunan dan penilaian kaedah FD-CARI+SA, manakala kerangka kerja yang dibentuk menyokong pelaksanaan menyeluruh pada sistem algoritma tersebut dalam konteks pengelasan kanser payudara. Secara keseluruhan, sumbangan kajian ini boleh diringkaskan kepada tiga kategori utama:

1.8.1 Sumbangan Konseptual

Kajian ini memperkenalkan kaedah pemilihan fitur berasaskan interaksi dengan gabungan pendiskretan kabur dan kaedah CARM bagi mengenal subset fitur yang lebih informatik dan tidak bertindih. Ia menyokong keperluan terhadap model pengelasan yang boleh dijelaskan, selaras dengan tuntutan sistem kesihatan yang memerlukan justifikasi terhadap keputusan model pembelajaran mesin. Selain itu, kerangka FD-CARI+SA yang dibangunkan memberikan struktur konseptual bagi pelaksanaan bersepadu pemilihan fitur, pengesahan interpretasi dan penalaan hiperparameter.

1.8.2 Sumbangan Metodologi

Dari aspek metodologi, kajian ini memberikan sumbangan melalui (i) pembangunan kerangka metodologi pemilihan fitur berasaskan interaksi yang inovatif, iaitu FD-CARI (*Fuzzy Discretization and Class Association Rule-based Interaction*), yang menggabungkan pendiskretan kabur dan CARM dalam satu aliran kerja tersusun untuk mengekstrak interaksi antara fitur serta menilai subset secara sistematik menggunakan metrik CARM, (ii) penggabungan FD-CARI dengan algoritma SA bagi membentuk satu kaedah hibrid FD-CARI+SA yang membolehkan penalaan automatik hiperparameter dijalankan secara automatik dan lebih cekap dan (iii) pengenalan mekanisme pengesahan dua lapis bagi subset fitur terpilih melalui gabungan nilai SHAP untuk menilai sumbangan dan interaksi fitur, dan pengundian majoriti berwajaran untuk memastikan kestabilan pemilihan merentas pelbagai model pengelasan berasaskan pokok.

1.8.3 Sumbangan Aplikasi Praktikal

Pelaksanaan dan penilaian menyeluruh ke atas tiga set data kanser payudara terbuka (WDBC, WPBC, BCC) menunjukkan bahawa rangka kerja yang dicadangkan mampu (i) mengurangkan bilangan fitur secara konsisten tanpa menjejaskan prestasi model, (ii) menyediakan penjelasan visual dan kuantitatif terhadap sumbangan serta hubungan antara fitur melalui SHAP, dan (iii) mengekalkan prestasi pengelasan yang stabil merentas pelbagai model. Walaupun peningkatan ketepatan yang diperoleh adalah kecil dan tidak bertujuan untuk intepretasi klinikal pada peringkat ini, kajian ini berfungsi sebagai bukti konsep yang menunjukkan kebolehlaksanaan rangka kerja FD-CARI+SA untuk data tabular klinikal dan biokimia dalam persekitaran eksperimen. Rangka kerja ini tidak bertujuan sebagai alat sedia aplikasi klinikal, tetapi menyediakan asas metodologi yang kukuh bagi pembangunan lanjut ke arah sistem sokongan keputusan yang lebih telus dan boleh dijelaskan pada masa hadapan.

1.9 STRUKTUR KANDUNGAN TESIS

Tesis ini disusun kepada tujuh bab yang merangkumi keseluruhan proses penyelidikan di mana ia bermula dengan pengenalpastian masalah sehinggalah kepada perbincangan dapatan dan cadangan untuk kajian masa hadapan. Ia dimulakan dengan Bab I yang menerangkan latar belakang kajian, persoalan kajian, kenyataan masalah, objektif kajian, skop kajian, batasan kajian dan ringkasan sumbangan utama kajian.

Bab II mengupas kajian-kajian terdahulu berkaitan pemilihan fitur, teknik-tenik pemilihan fitur, model XAI dan teknik pengoptimuman hiperparameter. Bab ini membantu mengenal pasti jurang penyelidikan yang ingin diatasi melalui kajian ini.

Bab III menghuraikan reka bentuk penyelidikan termasuk proses pra-pemprosesan data, perlaksanaan kaedah FD-CARI, mekanisme pengesahan fitur berinteraksi menggunakan model SHAP dan pengundian majoriti berwajaran, serta kaedah pengoptimuman menggunakan algoritma SA. Aliran kerja kajian ditunjukkan secara visual melalui carta alir dan pseudokod.

Bab IV membentangkan hasil pemilihan fitur menggunakan FD-CARI ke atas tiga set data kanser payudara. Analisis prestasi berdasarkan pelbagai model pengelasan turut dijelaskan bagi menilai keberkesanan subset fitur yang dipilih.

Bab V membincangkan hasil pengesahan terhadap fitur berinteraksi yang dipilih dengan menggabungkan model SHAP dan pengundian majoriti berwajaran.

Bab VI membincangkan hasil penalaan automatik hiperparameter utama dalam FD-CARI menggunakan algoritma SA serta perbandingan model sebelum dan selepas pengoptimuman.

Bab VII merumuskan keseluruhan dapatan kajian, menyenaraikan sumbangan kepada bidang ilmu serta mencadangkan hala tuju dan penambahbaikan untuk kajian lanjutan.

BAB II

KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Bab ini mengupas kajian kesusasteraan yang berkaitan dengan pemilihan fitur, interaksi fitur dan teknik-teknik sedia ada dalam diagnosis kanser payudara. Kajian terdahulu banyak menekankan kepentingan fitur secara individu, tetapi kurang meneliti interaksi kompleks antara fitur yang boleh meningkatkan prestasi model pengelasan. Keperluan kepada kaedah pemilihan fitur yang mempertimbangkan relevansi, interaksi dan menangani pertindihan bagi menghasilkan subset fitur yang padat dan bermakna. Bab ini juga menekankan potensi integrasi kaedah seperti *Association Rule Mining* (ARM), *eXplainable AI* (XAI) dan pengoptimuman hiperparameter sebagai asas kepada pembangunan model pengelasan kanser payudara yang lebih tepat dan boleh ditafsir.

2.2 KANSER PAYUDARA DAN CABARAN DIAGNOSIS AWAL

Kanser payudara merupakan salah satu jenis kanser yang paling kerap berlaku di kalangan wanita di seluruh dunia termasuk Malaysia. Menurut statistik daripada *World Health Organization* (World Health Organization 2024) dan Kementerian Kesihatan Malaysia (Azizah et al. 2019), kanser ini bukan sahaja menyumbang kepada peratusan tertinggi dalam jumlah kes kanser wanita, malah menjadi penyumbang utama kepada kadar kematian berkaitan kanser. Diagnosis awal dan tepat adalah kritikal kerana ia secara langsung meningkatkan kadar kelangsungan hidup pesakit dan membolehkan intervensi perubatan dijalankan pada peringkat awal (Liew et al. 2021; Lim et al. 2022).

Cabaran utama dalam diagnosis kanser payudara berkait rapat dengan kepelbagaian data klinikal yang perlu dianalisis. Data ini merangkumi ciri morfologi tumor seperti bentuk, tekstur dan struktur sel, faktor biokimia termasuk hormon dan

protein, serta maklumat klinikal seperti status nodus limfa, saiz tumor dan sejarah perubatan pesakit. Hubungan antara fitur-fitur ini sering bersifat tidak linear dan kompleks menyebabkan interaksi antara fitur sukar dikesan secara manual.

Oleh itu, kajian ini memberi tumpuan kepada pemilihan subset fitur yang relevan, berinteraksi dan tidak bertindih bagi menyokong pembangunan model ramalan kanser payudara yang berkesan dan boleh ditafsir dalam konteks klinikal. Terdapat tiga set data telah dipilih iaitu WDBC dan WPBC yang berasaskan ciri morfologi serta BCC yang mengandungi penanda bio (*biomarker*) pada darah untuk menunjukkan kebolehlaksanaan kaedah ini dalam pelbagai jenis data berkaitan kanser payudara.

2.2.1 Modaliti Pengimejan Diagnosis Kanser Payudara

Pengesanan kanser payudara merupakan bidang yang penting dalam pengimejan perubatan di mana pelbagai kaedah digunakan untuk meningkatkan ketepatan diagnosis dan menambahbaik hasil rawatan pesakit. Antara modaliti pengimejan yang paling kerap digunakan termasuk mamografi (MG), ultrabunyi (US), imbasan tomografi berkomputer (CT), pengimejan resonans magnetik (MRI), sitopatologi (CY) dan histopatologi (HP).

2.2.2 Sistem Diagnosis Berbantu Komputer

Sistem Diagnosis Berbantu Komputer (CAD) merupakan teknologi sokongan yang digunakan untuk membantu ahli radiologi dan pakar klinikal mentafsir imej perubatan khususnya dalam diagnosis kanser payudara. CAD mula diperkenalkan oleh Winsberg et al. (1967) dan kini digunakan secara meluas sebagai rutin klinikal untuk meningkatkan ketepatan diagnosis dan mengurangkan keterlepasan kes kanser (Lamb et al. 2020). Dalam konteks pengimejan kanser payudara, CAD berfungsi sebagai “mata kedua" (Keen et al. 2018) dengan menggunakan model pembelajaran mesin untuk mengenal pasti kawasan mencurigakan dalam imej dan membantu doktor dalam membuat keputusan yang lebih tepat. Fungsi ini juga dapat menjimatkan masa pakar dan mengekalkan konsistensi penilaian (Z. Guo et al. 2022).

a. Prosedur Sistem CAD

Terdapat dua komponen utama CAD iaitu Sistem Pengesanan Berbantu Komputer (CADe) dan Sistem Diagnostik Berbantu Komputer (CADx) (Wang & Liandong, 2019). CADe memfokuskan kepada pengesanan awal keabnormalan dan kawasan minat (*Region of Interest*, ROI) tumor manakala CADx memfokuskan pada tugas pengelasan tumor pada benigna atau malignan. Prosedur umum CAD diilustrasi seperti dalam Rajah 2.1.

i. Pemerolehan imej: Pemerolehan imej berkualiti tinggi daripada prosedur biopsi.

ii Pra pemprosesan imej: Peningkatan kualiti imej seperti penyingkiran hingar dan pembetulan artifak untuk memudahkan pengekstrakan fitur yang tepat. Teknik yang digunakan seperti penyamaan histogram, penapisan Gaussian dan peningkatan kontras (Avcı & Karakaya 2023; Mohd Sagheer & George 2020; Wu et al. 2013)

iii. Segmentasi: Pembahagian imej kepada ROI dengan struktur abnormaliti seperti massa dan kalsifikasi (Michael et al. 2021).

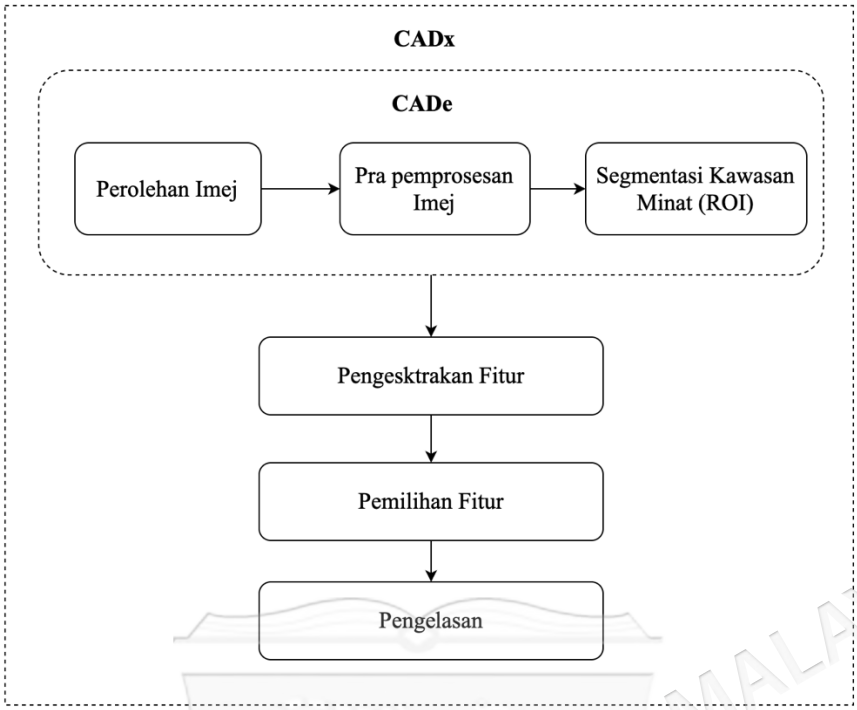
iv. Pengekstrakan fitur: Pengekstrakan fitur melibatkan pengenalpastian fitur yang berkaitan daripada ROI yang telah disegmentasi. Fitur-fitur ini boleh dikategorikan kepada beberapa jenis (Al-Thelaya et al. 2023):

- Fitur morfologi: Bentuk, saiz dan ciri-ciri sempadan lesi.
- Fitur tekstur: Taburan spatial intensiti piksel dan corak
- Fitur intensiti: Kecerahan atau kegelapan kawasan tertentu.
- Fitur statistik: Nilai yang diperoleh histogram intensiti atau analisis statistik peringkat lebih tinggi.

v. Pemilihan fitur: Penyingkiran fitur yang berlebihan atau tidak relevan yang boleh menurunkan prestasi sistem CAD (Li et al., 2017).

vi. Pengelasan: Pengkategorian fitur yang diekstrak ke dalam kelas yang telah ditetapkan oleh model pembelajaran mesin.

Hasil pengelasan dengan data klinikal lain digabungkan untuk menghasilkan laporan diagnostik yang lengkap. Hasil CAD kemudiannya dibentangkan kepada pakar radiologi dan doktor untuk penilaian lanjut.



Rajah 2.1 Carta alir sistem CADe dan CADx (Diadaptasi daripada Hassan et al. (2022))

2.2.3 Batasan CAD

Walaupun penggunaan CAD terbukti meningkatkan kepekaan pengesanan serta menyokong proses penghasilan keputusan klinikal, terdapat beberapa keterbatasan yang perlu diberi perhatian. Antaranya ialah kadar positif palsu (*false positive*) yang tinggi, yang boleh menyebabkan pesakit menjalani ujian lanjut atau rawatan tidak perlu (Lazard et al. 2023; Miglioretti et al. 2024). Di samping itu, keberkesanan CAD sangat bergantung kepada kualiti imej perubatan dan menjadikannya prestasi tidak konsisten antara pusat kesihatan (Herath et al. 2025).

Perkembangan kecerdasan buatan (AI) khususnya pembelajaran mesin dan pembelajaran mendalam telah memperluas potensi CAD (Thakur et al. 2024). Model AI digunakan secara meluas untuk pengelasan imej kanser payudara dan menghasilkan ketepatan yang tertinggi berbanding kaedah tradisional (Ahn et al. 2023). Namun demikian, cabaran besar tetap wujud. Kebanyakan model AI berfungsi sebagai "*black box*" di mana model menghasilkan keputusan tanpa memberikan penjelasan yang jelas mengenai bagaimana keputusan tersebut dicapai. Keadaan ini menimbulkan keraguan

dalam kalangan pengamal klinikal kerana interpretasi keputusan amat penting dalam konteks perubatan (Sadeghi et al. 2024).

Secara keseluruhan, meskipun CAD dan AI telah membawa kemajuan besar dalam diagnosis kanser payudara, teknologi ini masih berdepan kekangan dari aspek kebolehtafsiran, kebolehpercayaan dan konsistensi prestasi. Jurang ini menekankan keperluan kepada kaedah alternatif yang bukan sahaja mampu memberikan ketepatan pengelasan yang baik, tetapi juga menghasilkan pengetahuan yang boleh dijelaskan kepada pakar klinikal. Pendekatan seperti peraturan kesatuan dengan sokongan teknik pendiskretan kabur membentuk asas kepada kajian ini dan sekali gus berpotensi untuk mengisi jurang tersebut.

2.3 PEMILIHAN FITUR

Era masa kini adalah era berinformasi. Teknologi seperti *Internet of Things* (IoT) dan Perkomputeran Awan telah menyebabkan jumlah data yang sangat besar dijana setiap hari (Oussous et al. 2018). Walau bagaimanapun, data hanya berguna jika ia diproses dengan betul untuk mengekstrak maklumat yang berguna. Pertumbuhan data ini menimbulkan cabaran besar dalam pengurusannya di mana kaedah manual tidak lagi praktikal (Tang et al. 2014). Data dengan dimensi yang sangat tinggi ini juga menyebabkan "sumpahan kedimensian" (*curse of dimensionality*) yang boleh menjejaskan prestasi model (Hastie et al. 2009).

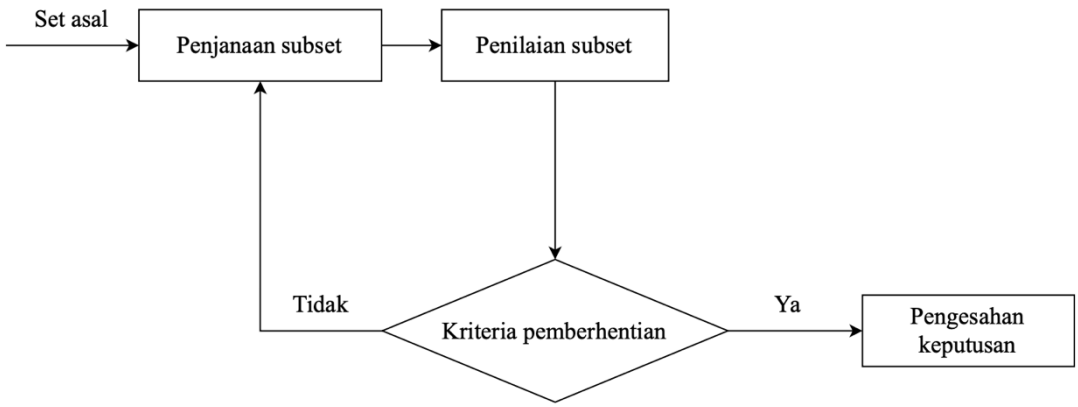
Sumpahan kedimensian berlaku dengan kehadiran bilangan fitur yang besar yang membawa kepada masalah "terlalu padan" (Vong et al., 2019). Masalah ini berlaku apabila model mempelajari data latihan dengan terlalu baik termasuk hingar yang membawa kepada prestasi buruk pada data baharu yang tidak kelihatan. Model juga telah menjadi terlalu kompleks dan gagal untuk menyamaratakan corak asas dalam data dan membawa kepada ramalan yang tidak tepat pada apa-apa data selain set latihan. Ini akan menyebabkan prestasi model pengelasan merosot. Untuk menangani masalah ini, pelbagai teknik pengurangan dimensi telah dikaji oleh para penyelidik. Pengurangan dimensi terbahagi kepada dua pendekatan utama iaitu (i) pengekstrakan fitur dan (ii) pemilihan fitur (Tang et al. 2014).

Pengekstrakan fitur merujuk kepada pembentukan fitur baru dengan dimensi yang lebih rendah. Namun, pendekatan ini boleh menyulitkan analisis lanjut kerana fitur baharu tersebut kehilangan makna fizikal yang menjadikannya sukar dari segi kebolehbacaan, kebolehtafsiran dan ketelusan (Remeseiro & Bolon-Canedo 2019). Contoh-contoh teknik ini termasuk PCA (Hotelling 1933; Pearson 1901), Analisis Diskriminan Linear (LDA) (Fisher 1936) dan Analisis Korelasi Kanonik (CCA) (Hotelling 1936).

Sebaliknya, pemilihan fitur memilih subset fitur asal yang relevan dan mengekalkan makna fizikal fitur yang menjadikannya lebih mudah untuk dianalisis dan ditafsirkan. Ia juga membantu mengurangkan hingar, memperbaiki prestasi model dan memastikan data yang lebih bersih dan bermakna (Li et al., 2017). Contoh teknik pemilihan termasuk *Correlation-based Feature Selection* (CFS) (Hall 2000), *Fast Correlation-based Filter* (FCBF) (Yu & Liu 2003) dan Relief (Šikonja & Kononenko 2003). Walaupun pemilihan fitur boleh mengurangkan sedikit ketepatan, ia membantu dalam kebolehtafsiran dan pengekstrakan pengetahuan terutamanya dalam bidang penjagaan kesihatan (Remeseiro & Bolon-Canedo, 2019).

2.3.1 Prosedur Pemilihan Fitur

Prosedur pemilihan fitur terdiri daripada beberapa langkah utama seperti penjanaan subset, penilaian subset, kriteria pemberhentian dan pengesahan keputusan seperti yang ditunjukkan dalam Rajah 2.2 (Dash & Liu 1997).



Rajah 2.2 Prosedur pemilihan fitur (Sumber: Dash & Liu (1997))

a. Penjanaan Subset

Proses ini melibatkan penjanaan subset fitur yang boleh bermula dengan set kosong, set lengkap semua fitur atau subset fitur secara rawak. Kaedah carian yang digunakan untuk menghasilkan subset termasuk:

- Carian lengkap: Meneroka semua kemungkinan subset fitur.
- Carian berturutan**:** Memilih atau membuang fitur secara progresif berdasarkan kriteria tertentu.
- Carian rawak**:** Menggunakan elemen rawak dalam proses pemilihan untuk mengelakkan terperangkap dalam optimum tempatan.

b. Penilaian Subset

Penilaian subset bertujuan menilai kualiti subset fitur yang dihasilkan berdasarkan kriteria tertentu. Terdapat dua jenis kriteria penilaian (i) kriteria bebas iaitu menggunakan fitur dalaman data untuk menilai subset tanpa melibatkan model pembelajaran (ii) kriteria bergantung iaitu menggunakan model pembelajaran untuk menilai prestasi subset fitur.

c. Kriteria Pemberhentian

Kriteria pemberhentian menentukan bila proses pemilihan fitur harus dihentikan. Beberapa kriteria pemberhentian seperti had yang ditentukan, carian selesai dan tiada perubahan signifikan.

d. Pengesahan Keputusan

Prosedur pengesahan bertujuan memastikan bahawa subset fitur yang dipilih relevan dan berkesan. Ini dilakukan dengan membandingkan hasil subset terhadap penanda aras atau hasil daripada kaedah pemilihan fitur lain.

2.3.2 Kaedah Pemilihan Fitur

Kaedah pemilihan fitur boleh dibahagikan kepada tiga kategori utama iaitu (i) penapis, (ii) pembalut dan (iii) terbenam.

a. Teknik Penapis

Teknik penapis menilai fitur secara bebas tanpa melibatkan algoritma pembelajaran (Tang et al. 2014). Kaedah penapis bermula dengan mengambil semua fitur dalam set data dan kemudian menilai setiap fitur berdasarkan ukuran statistik seperti maklumat bersama, korelasi dan jarak antara fitur untuk memilih subset fitur terbaik dan seterusnya diberikan nilai atau skor (Venkatesh & Anuradha 2019). Kaedah ini pantas dan cekap dari segi pengiraan dan memiliki kapasiti generalisasi yang baik dan sesuai untuk digunakan pada peringkat awal pemilihan fitur (Remeseiro & Bolon-Canedo 2019).

Teknik penapis boleh dikategorikan kepada kaedah univariat dan multivariat (Jundong Li et al. 2017). Kaedah univariat menilai setiap fitur secara individu tanpa mengambil kira hubungan atau kebergantungan antara fitur. Fitur-fitur ditaraf berdasarkan kepentingannya berdasarkan kelas sasaran. Sebaliknya, kaedah multivariat menilai fitur-fitur secara bersama membolehkan analisis hubungan antara fitur. Kaedah ini lebih berkesan dalam menangani fitur yang tidak relevan dan bertindih. Teknik seperti CFS, FCBF, Consistency (Manoranjan Dash & Liu 2003) dan Relief-F (Šikonja & Kononenko 2003) adalah contoh kaedah multivariat (Remeseiro & Bolon-Canedo 2019).

b. Teknik Pembalut

Teknik pembalut menggunakan algoritma pembelajaran untuk menilai subset fitur (Kumar 2014). Model akan dilatih pada setiap subset fitur dan prestasi dinilai berdasarkan metrik seperti ketepatan pengelasan atau ralat model. Proses ini dilakukan secara iteratif di mana fitur-fitur ditambah atau dikeluarkan untuk menghasilkan subset terbaik. Walaupun teknik ini mampu menghasilkan generalisasi yang baik, kos pengiraan menjadi sangat tinggi kerana model pengelas perlu dilatih berkali-kali untuk setiap kombinasi subset fitur terutamanya pada set data berdimensi tinggi (Li et al., 2017).

Untuk menangani cabaran ini, pelbagai strategi seperti carian berturutan (Guyon & Elisseeff 2003) seperti *Sequential Forward Selection* (SFS) dan *Sequential Backward*

Selection (SBS), *hill-climbing* (Skalak 1994), carian *Branch and Bound* (Kohavi & John 1997) dan algoritma genetik (GA) (Golberg 1989) sering digunakan. Strategi-strategi ini membantu dalam mencari penyelesaian yang optimum yang dapat meningkatkan prestasi pembelajaran.

c. Teknik Terbenam

Teknik terbenam menggabungkan pemilihan fitur ke dalam proses latihan model pembelajaran (Kumar 2014). Fitur-fitur yang relevan dikenal pasti semasa proses pengoptimuman model. Teknik ini mengelakkan keperluan untuk melatih semula model bagi setiap subset fitur menjadikannya lebih cekap berbanding teknik pembalut (Venkatesh & Anuradha 2019).

Contoh kaedah terbenam termasuk pemangkasan seperti *Recursive Feature Elimination* (RFE) pada SVM, mekanisme terbina dalam model seperti ID3 dan C4.5 yang memilih fitur secara automatik semasa proses pembelajaran dan model regularisasi seperti *Least Absolute Shrinkage and Selection Operator* (LASSO).

2.3.3 Kriteria Pemilihan Fitur

Terdapat tiga kriteria penting bagi memilih fitur iaitu (i) relevan (ii) bertindih dan (iii) interaksi antara fitur bagi memastikan subset fitur yang dipilih meningkatkan prestasi model pengelasan.

a. Relevansi

Kajian John et al. (1994) telah memperkenalkan tiga kategori utama jenis fitur berdasarkan kepentingannya terhadap pemboleh ubah sasaran. Kategori-kategori tersebut adalah (i) fitur sangat relevan, (ii) fitur kurang relevan dan (iii) fitur tidak relevan.

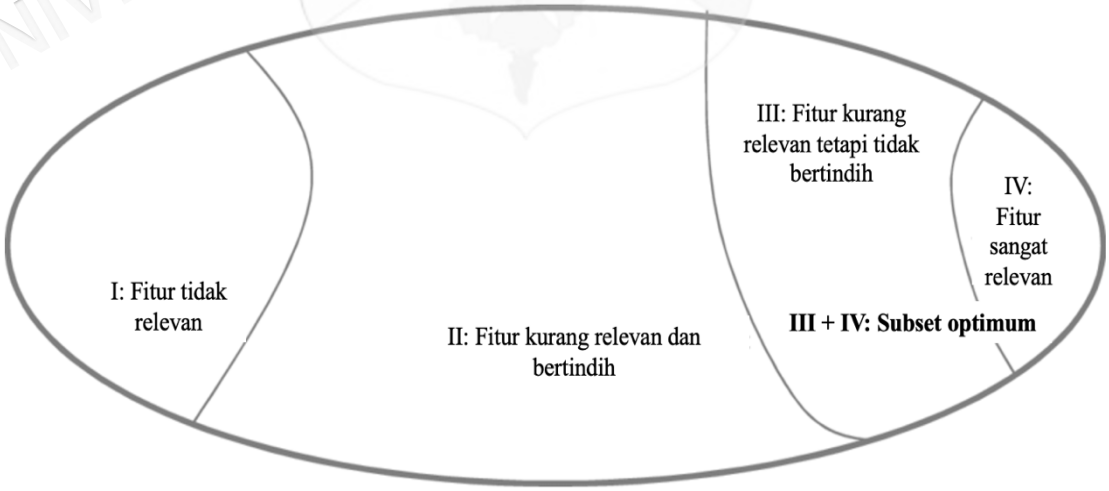
Fitur sangat relevan mempunyai tahap relevansi yang tinggi di mana ia mempunyai hubungan langsung dan signifikan dengan kelas sasaran. Sebaliknya, fitur kurang relevan mempunyai tahap relevansi yang rendah di mana fitur ini memberi maklumat yang kurang signifikan secara individu dengan kelas sasaran. Fitur ini

biasanya tidak dimasukkan dalam subset fitur akhir melainkan terdapat interaksi dengan fitur lain. Justeru itu, penghapusan fitur kurang relevan perlu dilakukan dengan berhati-hati kerana ia mungkin memberikan nilai maklumat tambahan dalam meramalkan kelas sasaran. Fitur tidak relevan pula tidak mempunyai hubungan yang bermakna dengan kelas sasaran dan biasanya disingkirkan bagi memastikan model lebih cekap.

b. Pertindihan

Pertindihan dalam pemilihan fitur berlaku apabila dua atau lebih fitur memberikan maklumat yang serupa atau hampir sama. Fitur yang bertindih ini tidak menambah nilai baru kepada model, malah ia meningkatkan kompleksiti data dan boleh membawa kepada masalah "terlalu padan" jika dimasukkan dalam model (Koller & Sahami 1996). Oleh itu, pengenalpastian dan penghapusan fitur adalah langkah penting dalam pemilihan fitur untuk meningkatkan kecekapan model.

Hubungan relevansi dan pertindihan antara fitur dapat dilihat pada Rajah 2.3. Keseluruhan set fitur dibahagikan kepada empat subset iaitu: (i) fitur tidak relevan (I), (ii) fitur kurang relevan dan bertindih (II), (iii) fitur kurang relevan tetapi tidak bertindih (III) dan (iv) fitur sangat relevan (IV). Subset optimum terdiri daripada fitur dalam kategori III dan IV.



Rajah 2.3 Hubungan relevansi dan pertindihan antara fitur (Al Nuaimi et al. 2019)

c. Interaksi

Konsep fitur berinteraksi diperkenalkan oleh Jakulin & Bratko (2004). Interaksi antara fitur dapat memberikan maklumat tambahan apabila digabungkan berbanding jumlah maklumat yang diberikan secara individu. Justeru itu, penyingkiran fitur secara tidak sengaja boleh menyebabkan prestasi pengelasan yang rendah kerana nilai gabungan antara fitur ini tidak dimanfaatkan sepenuhnya (Liang et al. 2019). Kajian terkini telah menunjukkan prestasi yang meningkat terhadap model pengelasan (Liang et al., 2019; Wan et al., 2021; Wang et al., 2021; Zeng et al., 2015).

2.3.4 Pemilihan Fitur Pada Diagnosis Kanser Payudara

Bahagian ini membincangkan pelbagai teknik pemilihan fitur yang telah digunakan pada pelbagai modaliti pengimejan seperti MG, US, CT, MRI, CY dan HP dalam kajian kanser payudara.

a. Mamografi

Terdapat pelbagai teknik penapis univariat digunakan pada imej MG. Sebagai contoh, kajian oleh Amin et al. (2023) menunjukkan keberkesanan teknik skor Fisher yang digunakan pada set data LE, CESM dan LE-CESM mencapai ketepatan 96.87% menggunakan SVM. Selain itu, kaedah Relief-F telah terbukti berkesan dalam kajian seperti Singh et al. (2022) dan Li et al. (2020) di mana ia mengatasi prestasi kaedah univariat dalam set data yang lebih kompleks. Dalam set data INbreast, Relief-F dan model pengelas kNN menghasilkan ketepatan 90.4% dalam menilai fitur yang relevan (Singh et al., 2022) . Begitu juga, pada set data BCDR, Relief-F membantu Interpolasi Peraturan Kabur (FRI) untuk mencapai ketepatan sebanyak 91.65% (Li et al., 2020).

Selain itu, teknik mRMR digunakan oleh Duarte et al. (2020) untuk memilih tekstur fitur daripada mikrokalsifikasi pada set data DDSM. Teknik ini berjaya memilih 13 fitur daripada 2432 fitur dan menghasilkan AUC sebanyak 0.945 menggunakan model pengelas Analisis Diskriminasi Fisher Tempatan (LFDA). Kajian Samee et al. (2022) telah menggabungkan teknik *Pearson Correlation Coefficient* (PCC), jarak Euclidean dan pekali Kosinus dan *Mutual Infomation* (MI) untuk memilih fitur dalaman

yang diekstrak daripada model CNN pada set data INbreast dan mencapai 98.50% ACC, 98.06% SEN dan 98.99% SPE.

Teknik pembalut seperti RFE juga digunakan oleh Al Rahman et al. (2023) untuk pemilihan fitur pada model pengelas RF, DT, XGB dan CatBoost menggunakan set data SEER. Hasil kajian menunjukkan RF mencapai skor AUC paling tinggi iaitu 0.91. Selain itu, Al-Fahaidy et al. (2022) menggunakan teknik SFS pada set data MIAS dan model pengelas SVM mengelaskan imej normal dan abnormal serta pengelasan imej abnormal sebagai benigna atau malignan. Hasil eksperimen menunjukkan 100% ACC untuk pengelasan normal atau abnormal, manakala pengelasan benigna atau malignan mencapai 87.1% ACC.

Kajian-kajian terkini menunjukkan keberkesanan teknik metaheuristik dalam pemilihan fitur dan pengelasan tumor menggunakan imej MG. Sebagai contoh, Ketabi et al. (2021) menggunakan GA untuk memilih fitur dan model pengelas SVM menghasilkan 90% ACC pada set data DDSM. Selain itu, Boumaraf et al. (2020) menerapkan GA yang diubah suai bersama *Backpropagation Neural Network* (BPNN) dan mencapai ACC sebanyak 84.5%. Kajian oleh Thawkar & Ingolikar (2020) pula menggunakan GA dengan beberapa model pengelas seperti AdaBoost, RF dan DT. Hasil kajian mendapati AdaBoost mencapai 96.15% ACC. Selain itu, algoritma Pengoptimuman Cacing Bercahaya (GWO) juga digunakan oleh Singh & Sivakkumar (2021) menggunakan GWO bersama *Pulsed Coupled Neural Network* (PCNN) dan mencapai 96.43% ACC pada set data MIAS. Di samping itu, algoritma Pengoptimuman Belalang (GOA) digunakan oleh Sha et al. (2020) bersama pengelas CNN dan mencapai 92% ACC pada set data MIAS dan DDSM. Kalita et al. (2022) pula menggunakan algoritma Titisan Air Pintar (IWD) dengan pelbagai pengelas seperti SVM, NB dan RF dan telah mencapai ACC sebanyak 99% pada pengelas SVM.

Teknik hibrid metaheuristik digunakan oleh Stephan et al. (2021) menggunakan gabungan Koloni Lebah Buatan (ABC) dan Pengoptimuman Paus (WO) untuk pemilihan fitur bersama pengelas ANN. Kajian ini menggunakan pelbagai set data MG DDSM, MIAS dan INbreast dan mencatatkan keputusan cemerlang dengan ACC sebanyak 98.8% pada DDSM, 98.7% pada MIAS dan 99.1% pada INbreast. Thawkar

et al. (2021) pula menggunakan gabungan Pengoptimuman Kerumunan Zarah (PSO) dan GA dikenali sebagai MOPSO dan model ANN pada set data DDSM dan mencapai 97.54% ACC. Kajian Pramanik et al. (2023) menggunakan gabungan algoritma Sosial-Pemandu-Ski dan Adaptif Beta Pendakian Bukit dan model kNN mencapai 96.07% ACC pada set data DDSM.

b. Ultrabunyi

Teknik penapis telah digunakan secara meluas dalam kajian pengimejan US. Rehman et al. (2023) menggunakan teknik entropi bagi pemilihan fitur dan model CNN dalam pengelasan tumor benign dan malignan pada set data BUSI dan menghasilkan prestasi tinggi dengan ACC sebanyak 97.04%. Selain itu, kajian oleh Daoud et al. (2020) dan Daoud et al. (2019) menggunakan mRMR bersama model SVM pada set data BUSI dan mencatatkan 96.1% ACC. Teknik CFS juga telah digunakan dalam kajian Shia et al. (2021) bersama model *Local Weighted Learning* (LWL) bagi pengelasan tumor benigna dan malignan dan mencapai prestasi yang baik dengan skor AUC 0.847.

Teknik pembalut seperti SFS juga digunakan dalam kajian Muhtadi (2022) dan model RF pada set data OASBUD dan menghasilkan ACC sebanyak 93.01%. Selain itu, Mishra et al. (2021) menggunakan RFE pada set data BUSI menghasilkan prestasi tinggi iaitu mencatatkan 97.4% ACC apabila digabungkan dengan model AdaBoost. Teknik metaheuristik juga telah digunakan bagi menangani masalah set data ultrabunyi yang kompleks. Cruz-Ramos et al. (2023) menggabungkan algoritma GA dan MI pada model Adaboost pada set data BUSI dan menghasilkan 96.1% ACC. Sementara itu, Alhussan et al. (2023) menggabungkan algoritma Pengoptimuman Dinamik Burung Dipper Berleher (DTO) dengan algoritma PSO bersama model CNN telah menghasilkan ACC sebanyak 98.1%. Selain itu, Khanna et al. (2021) menggunakan algoritma BGWO pada set data yang sama bersama pengelas SVM dan mencatatkan 84.6% ACC dan skor AUC sebanyak 0.97.

c. Imbasan Tomografi Berkomputer

Kajian Kuo et al. (2023) menggunakan teknik SFS untuk memilih fitur dalam mengelaskan tumor benigna dan malignan berdasarkan set data imej imbasan CT. Hasil

kajian mencapai 99.43% ACC, 98.82% SEN dan 100% SPE menggunakan model pengelas SVM.

d. Pengimejan Resonans Magnetik

Teknik skor Fisher pada imej MRI digunakan oleh Mutlu et al. (2020) dalam set data MRI yang mengandungi 84 sampel. Model RF telah mencatatkan 90.36% ACC, 96.25% SEN dan 83.33% SPE. Kajian lain oleh Çetinel et al. (2020) turut menggunakan skor Fisher dan model SVM untuk membangunkan sistem sokongan keputusan bagi segmentasi lesi payudara dan menghasilkan 86% ACC. Selain itu, Whitney et al. (2021) menggunakan teknik PCC dan model LDA mencapai skor AUC sebanyak 0.797. Teknik pembalut *Stepwise* dengan model SVM digunakan oleh Li et al. (2022) dan mencapai skor AUC sebanyak 0.90. Kajian Liu et al. (2020) pula membandingkan teknik *Stepwise*, SFS dan SBS bersama model SVM. Kajian telah mendapati bahawa SBS menghasilkan prestasi tertinggi sebanyak 94.52% SEN dan 90% SPE.

e. Sitologi

Kajian Jain et al. (2021) menggunakan teknik korelasi digunakan pada set data WDBC pada model SVM, RF dan MLP. Hasil kajian mendapati model RF mencatatkan ACC tertinggi sebanyak 96.50%. Selain itu, Pacheco (2022) mengkaji teknik kotak-Chi dan MI dan mencapai 96.49% ACC pada gabungan kotak-Chi dengan AdaBoost. Kajian Rahman et al. (2020) pula mengkaji teknik kotak-Chi, MI dan PCC dengan model ANN pada set data WOBC dan mencapai 99% ACC. Selain itu, Minnoor & Baths (2023) menggunakan teknik korelasi dan RF mencatatkan ACC tertinggi sebanyak 99.30% pada model RF pada set data WDBC. Teknik CFS mencapai 97% ACC pada kajian Hussain et al. (2021) apabila digunakan dengan model pengelas gabungan iaitu NB dan DT manakala Chhillar & Singh (2025) mencapai 99.34% ACC dengan model gabungan MLP, SVM dan XGB menggunakan teknik yang sama.

Teknik pembalut juga kerap digunakan pada data sitologi. Sebagai contoh, RFE digunakan secara meluas dalam kajian oleh Ono & Mitani (2022), Aamir et al. (2022) dan Prowal et al. (2021). Kajian Ono & Mitani (2022) mencatatkan 95.34% ACC apabila RFE digabungkan dengan RF pada set data WDBC manakala kajian Aamir et

al. (2022) mencatatkan ACC sehingga 98.33% apabila gabungan teknik RFE dan korelasi digunakan pada model MLP. Kajian Prowal et al. (2021) juga menunjukkan prestasi tinggi dengan ANN-RFE dengan mencatatkan 98.36% ACC pada set data WDBC.

Teknik metaheuristik iaitu algoritma GA digunakan oleh Ayoola & Ogunfunmi (2022) dan Rezaeipanah & Ahmadi (2022) untuk pemilihan fitur. Kajian Ayoola & Ogunfunmi (2022) telah menghasilkan 99.12% ACC dengan model RF pada set data WDBC manakala Rezaeipanah & Ahmadi (2022) telah mencapai 99.35% ACC dengan MLP pada set data WOBC. Teknik metaheuristik yang unik digunakan pada CY seperti Sannasi Chakravarthy et al. (2022) menggunakan algoritma Pengoptimuman Ebola pada set data WDBC dan mencapai 97.19% ACC dengan SVM. Begitu juga, Jovanovic et al. (2022) menggunakan algoritma Pengoptimuman Penguin Maharaja bersama XGB pada set data WOBC dan mencapai 97.21% ACC. Kajian Dalwinder et al. (2020) pula mencapai 97.85%, 98.37% dan 82.79% ACC pada set data WOBC, WDBC dan BCC masing-masing menggunakan algoritma Singa Semut menggunakan model BPNN.

Selain itu, teknik hibrid metaheuristik juga dikaji oleh Elkorany et al. (2022) WO dan algoritma Pepatung menghasilkan 97.89% ACC pada set data WDBC dan 99.27% ACC pada WOBC apabila digabungkan dengan SVM. Kajian oleh Stephan et al. (2021) pula mengkaji hibrid algoritma ABC dan WO pada WOBC, WDBC dan WPBC dan telah mencatatkan ACC sebanyak 99.2%, dan 98.5%, 96.3% pada masing-masing menggunakan model pengelas ANN. Kajian Murugesan et al. (2021) pula telah memperkenalkan kaedah hibrid menggunakan kawanan kucing, Kumpulan Krill dan Pengambilan Makanan Bakteria menggunakan model BPNN dan mencapai 96.83% ACC pada set data WDBC. Selain itu, Khanna et al. (2024) mencapai 97.96% ACC menggunakan teknik hibrid Pengoptimuman Berasaskan Pengajaran Pembelajaran (TLBO) dan Penggembalaan Gajah (EHO) pada set data WDBC menggunakan model pengelas kNN.

f. Histologi

Kajian Umer et al. (2023) menggunakan gabungan Relief-F dan PCC sebagai teknik pemilihan fitur pada model SVM, Pokok Ensembel dan CNN. Hasil kajian mendapati

model CNN mencatatkan 92.7% ACC dalam pengelasan tumor. Selain itu, Carvalho et al. (2020) mengkaji pengelasan tumor pada set data BreakHis menggunakan CFS, Relief-F dan IG dengan model RF dan SVM. Teknik gabungan Relief-F dan RF telah mencatatkan prestasi terbaik dengan 97.6% ACC. Teknik pembalut SFS juga diuji pada kajian Lagree et al. (2021) dengan model kNN, LR, RF, SVM dan XGB. Kajian ini telah menunjukkan gabungan model pengelasan oleh LR dan RF mencatatkan skor AUC sebanyak 0.836. Selain itu, algoritma GA juga dikaji oleh Reshma et al. (2022) pada set data BreakHis dan CNN dan mencatatkan 92.44% ACC.

Jadual 2.1 menunjukkan ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur pada pengelasan kanser payudara.

Jadual 2.1 Ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur pada pengelasan kanser payudara

Rujukan	Imej Modaliti	Set Data	Kaedah Pemilihan Fitur	Model Pengelasan	Hasil Keputusan
Amin et al. (2023)	MG	LE, CESM, LE-CSM	Skor Fisher	SVM	ACC=96.87%
Singh et al. (2022)	MG	INbreast	Relief-F	kNN	ACC=90.4%
Li et al. (2020)	MG	BCDR	Relief-F	FRI	ACC=91.65%
Duarte et al. (2020)	MG	DDSM	MI, mRMR	LFDA	AUC=0.884
Samee et al. (2022)	MG	INbreast	PCC, Jarak Euclidean, Pekali Kosinus, MI	CNN	ACC=98.5%
Al Rahman et al. (2023)	MG	SEER	RFE	RF, DT. XGB, CatBoost	AUC=0.91
Al-Fahaidy et al. (2022)	MG	MIAS	SFS	SVM	ACC=87.1%
Samala et al. (2021)	MG	DDSM	SFS	LDA, SVM, RF	LDA AUC=0.91
Ketabi et al. (2021)	MG	DDSM	GA	SVM	ACC=90%
Boumaraf et al. (2020)	MG	DDSM	GA	BPNN	ACC=84.5%
Thawkar & Ingolikar (2020)	MG	DDSM	GA	AdaBoost, RF, DT	AdaBoost ACC=96.15%
Singh & Sivakkumar (2021)	MG	MIAS	GWO	PCNN	ACC=96.43%
Sha et al. (2020)	MG	MIAS, DDSM	GOA	CNN	ACC=92%
Rajendran et al. (2022)	MG	MIAS	GOA + Burung gagak	MLP	ACC=97.1%
Kalita et al. (2022)	MG	Mini-MIAS	IWD	SVM, NB, RF	SVM ACC=99%
Stephan et al. (2021)	MG	DDSM, MIAS, INbreast	ABC, WO	ANN	(ACC) DDSM=98.8%, MIAS= 98.7%, INbreast=99.1%
Thawkar et al. (2021)	MG	DDSM	PSO + GA	ANN	ACC=97.54%
Pramanik et al. (2023)	MG	DDSM	Pemandu Ski+ Pendakian Bukit	kNN	ACC=96.07%

bersambung...

...sambungan

Rehman et al. (2023)	US	BUSI	Entropi	CNN	ACC=97.04%
Daoud et al. (2020)	US	BUSI	mRMR	SVM	ACC=96.1%
Shia et al. (2021)	US	Persendirian	CFS	LWL	AUC=0.847
Muhtadi (2022)	US	OASBUD	SFS	RF	ACC=93.01%
Mishra et al. (2021)	US	BUSI	RFE	AdaBoost	ACC=97.4%
Cruz-Ramos et al. (2023)	US	BUSI	GA + MI	AdaBoost	ACC=96.1%
Alhussan et al. (2023)	US	BUSI	DTO + PSO	CNN	ACC=98.1%
Khanna et al. (2021)	US	BUSI	BGWO	SVM	ACC=84.6%
Kuo et al. (2023)	CT	Persendirian	SFS	SVM	ACC=99.43%
Mutlu et al. (2020)	MRI	Persendirian	Skor Fisher	RF	ACC=90.36%
Çetinel et al. (2020)	MRI	Persendirian	Skor Fisher	SVM	ACC=86%
Whitney et al. (2021)	MRI	Persendirian	PCC	LDA	AUC=0.797
Li et al. (2022)	MRI	Persendirian	Stepwise	SVM	AUC=0.90
Liu et al. (2020)	MRI	Persendirian	Stepwise, SFS, SBS	SVM	SEN=94.52%
Jain et al. (2021)	CY	WDBC	Korelasi	SVM, RF, MLP	RF ACC=96.50%
Pacheco (2022)	CY	WDBC	Kotak-Chi, MI	AdaBoost	Kotak-Chi ACC=96.49%
Rahman et al. (2020)	CY	WOBC	Kotak-Chi, MI, PCC	DT, SVM, kNN, ANN	Kotak-Chi+ANN ACC=99%
Minnoor & Baths (2023)	CY	WDBC	PCC + RFE + LR + ANOVA	RF	ACC=99.30%
Hussain et al. (2021)	CY	WDBC	CFS	NB+DT	ACC=97%
Chhillar & Singh (2025)	CY	WDBC	CFS	MLP+SVM+XGB	ACC=99.34%
Ono & Mitani (2022)	CY	WDBC	RFE	RF	ACC=95.34%

bersambung...

...sambungan

Aamir et al. (2022)	CY	WDBC	RFE + Korelasi	MLP	ACC=98.33%
Prowal et al. (2021)	CY	WDBC	RFE	ANN	ACC=98.36%
Alghera iry et al. (2022)	CY	WDBC	SFS	LR	ACC=98.2%
Ayoola & Ogunfunmi (2022)	CY	WDBC	GA	RF	ACC=99.12%
Rezaeipanah & Ahmadi (2022)	CY	WOBC	GA	MLP	ACC=99.35%
Sannasi Chakravarthy et al. (2022)	CY	WDBC	Ebola	SVM	ACC=97.19%
Jovanovic et al. (2022)	CY	WOBC	Penguin Maharaja	XGB	ACC=97.21%
Dalwinder et al. (2020)	CY	WOBC, WDBC, BCC	Singa Semut	BPNN	(ACC) WOBC=97.85%, WDBC=98.37%, BCC=82.79%
Elkorany et al. (2022)	CY	WDBC, WOBC	WO, Pepatung	SVM	(ACC) WDBC=97.89%, WOBC=99.27%
Stephan et al. (2021)	CY	WOBC, WDBC, WPBC	ABC, WO	ANN	(ACC) WOBC=99.2%, WDBC=98.5%, WPBC=96.3%
Murugesan et al. (2021)	CY	WDBC	Kawanan Kucing + Kumpulan Krill + Pengambilan Makanan Bakteria	BPNN	ACC=96.83%
Khanna et al. (2024)	CY	WDBC	Pengajaran Pembelajaran + Penggembalaan Gajah	kNN	ACC-97.96%
Umer et al. (2023)	HP	BHI	Relief-F + PCC	SVM, Pokok Ensembel, CNN	CNN ACC=92.7%
Carvalho et al. (2020)	HP	BreakHis	CFS, Relief-F, IG	RF, SVM	Relief-F+RF ACC=97.6%
Lagree et al. (2021)	HP	Persendirian	SFS	kNN, LR, RF, SVM, XGB	LR+RF AUC=0.836
Reshma et al. (2022)	HP	BreakHis	GA	kNN, NB, Transformasi Diskret, SVM, CNN	CNN ACC=92.44%

Berdasarkan Jadual 2.1, pelbagai kaedah pemilihan fitur telah digunakan dalam pengelasan kanser payudara merentasi set data seperti DDSM, INbreast, MIAS, BUSI, WDBC, WOBC, BCC dan BreakHis. Kaedah tradisional seperti Relief-F, MI, mRMR, PCC, RFE, SFS dan skor Fisher lazimnya menilai kerelevanan fitur secara individu dan mengurangkan pertindihan sebelum digunakan bersama model pembelajaran mesin seperti SVM, RF, kNN, NN dan AdaBoost, dan telah menunjukkan ketepatan sekitar 85–99%. Perkembangan selanjutnya menunjukkan penggunaan semakin meluas kaedah metaheuristik seperti GA, PSO, GWO, GOA, ABC, IWD serta pendekatan hibrid yang turut mencatatkan prestasi tinggi sehingga sekitar 97–99%. Walau bagaimanapun, analisis keseluruhan mendapati bahawa kedua-dua kaedah tradisional dan metaheuristik masih menumpukan kepada pemilihan fitur berdasarkan kerelevanan individu tanpa menilai hubungan interaksi antara fitur atau menyediakan mekanisme pengesahan fitur yang telus. Kekangan ini mewujudkan keperluan kepada kaedah pemilihan fitur berasaskan interaksi yang lebih sistematik serta disertakan proses pengesahan berasaskan XAI bagi meningkatkan kebolehjelasan dan kebolehpercayaan model.

2.3.5 Kaedah Pemilihan Fitur Berdasarkan Relevan, Pertindihan dan Interaksi

Pemilihan fitur berdasarkan interaksi telah menjadi kaedah yang semakin penting dalam pembelajaran mesin terutamanya apabila berhadapan dengan set data berdimensi tinggi. Model ini bertujuan untuk memilih fitur yang relevan dan berinteraksi, serta menyingkirkan fitur yang tidak relevan dan bertindih. Fitur yang relevan dapat meningkatkan prestasi model, manakala fitur yang tidak relevan dan bertindih boleh mengganggu prestasi model pengelasan dan menurunkan keberkesanannya.

Beberapa kaedah yang popular untuk menangani fitur bertindih termasuk *Mutual Information Feature Selection* (MIFS) (Battiti 1994), *Joint Mutual Information* (JMI) (Meyer et al. 2008; Yang & Moody 1999), *Conditional Mutual Information Maximization* (CMIM) (Fleuret 2004; Gang Wang et al. 2004) dan *Minimum Redundancy Maximum Relevance* (mRMR) (Peng et al. 2005). MIFS adalah kaedah awal yang menggunakan teori maklumat bersama untuk memilih fitur berdasarkan relevansi dengan kelas sasaran sambil mempertimbangkan pertindihan antara fitur. Walau bagaimanapun, ia sering kali terlalu sensitif terhadap parameter pertindihan.

Untuk memperbaiki kelemahan ini, JMI dan CMIM diperkenalkan untuk menangani hubungan gabungan dan bersyarat antara fitur dan kelas sasaran.

Selain itu, kriteria yang sangat penting tetapi sering diabaikan ialah interaksi antara fitur (Deisy et al., 2010; Gu et al., 2020; Jakulin & Bratko, 2003, 2004). Interaksi fitur merujuk kepada situasi di mana fitur mungkin mempunyai korelasi yang lemah dengan kelas sasaran, tetapi apabila digabungkan dengan fitur lain, ia boleh menghasilkan hubungan korelasi yang kuat (Liu et al., 2023). Interaksi antara fitur penting untuk meningkatkan ketepatan pengelasan walaupun fitur itu sendiri mungkin tidak memberikan maklumat yang relevan secara individu. Jakulin & Bratko (2003) menunjukkan dalam masalah *Exclusive OR* (XOR) bahawa dua fitur yang tidak relevan secara individu dapat meningkatkan prestasi apabila digabungkan.

Beberapa algoritma telah dibangunkan untuk menangani interaksi antara fitur. Antaranya ialah algoritma INTERACT Zhao & Liu (2007, 2009) yang menggabungkan teori *Symmetry Uncertainty* (SU) dan *Consistency Contribution* (CC) untuk menilai sejauh mana penyingkiran fitur akan mempengaruhi konsistensi data. Selain itu, algoritma *Interaction Weight based Feature Selection* (IWFS) (Zeng et al., 2015) menggunakan SU dan MI untuk mengenal pasti fitur yang berinteraksi dan berkesan untuk set data sintetik dan dunia nyata. Kedua-dua algoritma ini menunjukkan keberkesanan dalam meningkatkan ketepatan ramalan berbanding dengan kaedah lain seperti FCBF dan Relief-F. Konsep pemilihan berdasarkan interaksi diperluaskan dengan algoritma *Neighborhood Interaction Weight based Feature Selection* (NIWFS) yang telah diperkenalkan oleh Zeng, Zhang, Zhang & Zhang (2015) dengan menggabungkan set kasar kejiranan untuk menangani fitur-fitur berangka dan kategori dalam set data bercampur.

Untuk menangani hubungan yang lebih kompleks, algoritma *Maximum-Relevance and Maximum-Interaction* (MRMI) (Lianxi Wang et al. 2021) mengimbangi relevansi dan interaksi fitur. Algoritma ini memilih fitur yang memberikan sumbangan terbesar kepada kelas sasaran sambil mengoptimumkan interaksi antara fitur. Walau bagaimanapun, MRMI berhadapan dengan cabaran dalam kebolehskalaan terutamanya pada set data berdimensi tinggi. Selain itu, algoritma *Nonlinear Dynamic Conditional*

Relevance Feature Selection (NDCRFS) yang diperkenalkan oleh Zhang (2021) menggunakan tiga metrik utama daripada teori maklumat iaitu MI, CMI dan pemberat dinamik.

Interaksi tiga hala untuk menilai relevansi dan pertindihan fitur turut dikaji oleh algoritma *Feature Interaction Maximisation* (FIM) (Bennasar et al. 2013). FIM memastikan hanya fitur yang saling melengkapi dipilih. Walau bagaimanapun, kaedah ini bergantung pada penyaringan fitur dan tidak boleh menghentikan pemilihan fitur secara automatik yang memerlukan penyesuaian manual. Omana & Jesu (2017) pula mengukuhkan lagi kemajuan dalam kaedah pemilihan fitur berdasarkan skor MI melalui algoritma *Joint Feature Interaction Maximization* (JFIM) menilai interaksi berbilang fitur dan kelas sasaran serta mengurangkan pertindihan dan meningkatkan ketepatan pengelasan.

Kaedah MI mempunyai batasannya di mana ia tidak mampu menangkap hubungan bersyarat antara fitur apabila terdapat pengaruh fitur lain. Untuk mengatasi kelemahan ini, penyelidik telah mengembangkan CMI yang menilai maklumat tambahan yang disumbangkan oleh fitur dengan mengambil kira pengaruh fitur lain. CMI digunakan secara meluas dalam algoritma seperti *Conditional Mutual Information based Feature Selection considering Interaction* (CMIFSI) (Liang et al. 2019) yang direka untuk menangkap interaksi dengan lebih tepat. CMIFSI menggabungkan algoritma SFS dengan CMI untuk menilai relevansi fitur secara iteratif sambil mengelakkan fitur bertindih. Di samping itu, kekangan penilaian tahap interaksi yang lebih tinggi dikaji oleh (Shishkin et al. 2016) melalui algoritma *Conditional Mutual Information with Complementary and Opposing Teams* (CMICOT) menggunakan kaedah *greedy estimation* dan perwakilan binari.

Kompleksiti dan berdimensi tinggi pada set data telah mendorong penyelidik untuk mengkaji algoritma yang mempertimbangkan interaksi antara fitur seperti *Backward Dropping Algorithm* (BDA) (Wang et al., 2012), *Sequential-Interaction Group-Selection* (SIGS) (Luo & Chen 2021) dan *High Dimensional Selection with* Interactions (HDSI) (Jain & Xu, 2021). BDA menggunakan kaedah SBS untuk membuang fitur yang tidak relevan secara sistematik dan memastikan hanya fitur yang

relevan dikekalkan. Sebaliknya, HDSI menggunakan teknik LASSO yang digabungkan dengan istilah interaksi. Dengan menggunakan sampel secara rawak, HDSI mampu menilai hubungan kompleks antara fitur dan memastikan subset fitur yang dipilih signifikan secara statistik. SIGS pula membezakan antara fitur mudah dan fitur komposit yang menggabungkan fitur dan interaksi.

Penggunaan ARM untuk menilai interaksi antara fitur juga dikaji seperti *Feature Subset Selection Algorithm based on Association Rule Mining* (FEAST) (Wang & Song, 2012). Peraturan kesatuan kelas digunakan untuk mengenal pasti fitur relevan dan interaksi. Kajian ini dikembangkan dengan teknik *Partial Least Squares Regression* (PLSR) bagi menentukan ambang sokongan dan keyakinan (Wang & Song, 2013). Begitu juga, algoritma *Extended Association Rules As Features* (EARAF) (Lin & Gao, 2021) direka untuk data kategori, data berterusan dan data yang tidak seimbang. Selain itu, algoritma *Feature Selection Method based on Frequent and Associated Itemsets* (FS-FAI) ((Mamdouh Farghaly & Abd El-Hafeez 2023) menggunakan ARM untuk memilih fitur yang relevan namun ia memerlukan penyesuaian ambang secara manual.

Jadual 2.2 menunjukkan ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur berdasarkan relevansi, pertindihan dan interaksi.

Jadual 2.2 Ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur berdasarkan relevansi, pertindihan dan interaksi

Rujukan	Kaedah Pemilihan Fitur	Konsep Teori	Kriteria Pemilihan Fitur			Hasil Kajian	Limitasi Kajian
			Relevansi	Pertindihan	Interaksi		
(Battiti, 1994)	MIFS	MI	√	√		Fitur relevan dan bertindih diperoleh	Sensitif pada parameter pertindihan
(Meyer et al., 2008; Yang & Moody, 1999)	JMI	MI	√	√		Fitur relevan dan bertindih diperoleh	Tidak mempertimbangkan interaksi
(Fleuret, 2004; G. Wang et al., 2004)	CMIM	MI	√	√		Fitur relevan dan bertindih diperoleh	Tidak mempertimbangkan interaksi
(Peng et al., 2005)	mRMR	MI	√	√		Fitur relevan dan bertindih diperoleh	Tidak mempertimbangkan interaksi
Zhao & Liu (2007, 2009)	INTERACT	MI, SU	√	√	√	Fitur berinteraksi diperoleh	Terhad pada hubungan dua hala interaksi
Zeng et al. (2015)	IWFS	MI, SU	√	√	√	Fitur berinteraksi diperoleh	Terhad pada hubungan dua hala interaksi
Zeng, Zhang, Zhang, & Zhang (2015)	NIWFS	MI, SU, Set Jiran Kasar	√	√	√	Fitur berinteraksi diperoleh	Bergantung kepada definisi kejiranan
Wang et al. (2021)	MRMI	MI, SU	√	√	√	Fitur berinteraksi diperoleh	Kos pengiraan tinggi pada set data besar
Zhang (2021)	NDCRFS	MI, CMI	√	√	√	Ketepatan pengelasan lebih tinggi	Prestasi bagi beberapa set data tidak memuaskan
Bennasar et al. (2013)	FIM	MI	√	√	√	Fitur berinteraksi tiga hala	Ketidakupayaan berhenti memilih secara automatik
Omana & Jesu (2017)	JFIM	JMIM	√	√	√	Ketepatan pengelasan lebih tinggi	Terhad kepada penggunaan domain
(Liang et al., 2019)	CMIFSI	CMI	√	√	√	Fitur berinteraksi diperoleh	Kestabilan algoritma perlu dikaji
(Shishkin et al., 2016)	CMICOT	CMI	√	√	√	Tahap fitur berinteraksi lebih tinggi	Model pembelajaran digunakan terhad
(Wang et al., 2012)	BDA	SBS	√	√	√	Fitur berinteraksi diperoleh pada set data besar	Kos pengiraan tinggi

bersambung…

…sambungan

(Jain & Xu, 2021)	HDSI	LASSO	√	√	√	Fitur berinteraksi diperoleh	Tidak dikaji dengan model pembelajaran
Luo & Chen (2021)	SIGS	Korelasi	√	√	√	Fleksibel pada set data besar	Terhad pada satu domain
Guangtao Wang & Song (2012)	FEAST	ARM	√	√	√	Fitur berinteraksi diperoleh	Penyesuaian ambang secara manual
(Wang & Song, 2013)	FEAST	ARM, PLSR	√	√	√	Penentuan ambang ARM dan fitur berinteraksi diperoleh	Hanya menggunakan metrik asas
Lin & Gao (2021)	EARAF	ARM	√	√	√	Masa dan ruang dioptimumkan	Hanya menggunakan metrik asas
Mamdouh Farghaly & Abd El-Hafeez (2023)	FS-FAI	ARM	√	√	√	Bilangan fitur berkurang	Penyesuaian ambang secara manual

Berdasarkan Jadual 2.2, kaedah seperti MIFS, JMI, CMIM, dan mRMR menumpukan kepada relevansi dan pertindihan fitur. Walaupun kaedah-kaedah ini berkesan dalam memilih fitur yang relevan, namun ia tidak mempertimbangkan interaksi kompleks antara fitur yang boleh mengehadkan prestasi model apabila hubungan fitur saling bergantung. Kaedah seperti INTERACT, IWFS, MRMI, FIM, JFIM, CMIFSI, dan CMICOT pula dibangunkan untuk mengenal pasti interaksi antara dua atau tiga fitur. Kaedah ini memberikan kelebihan dengan menangkap pola kompleks, tetapi ada batasan seperti kos pengiraan tinggi atau keterbatasan pada interaksi dua hala sahaja. Di samping itu, kaedah FEAST, EARAF, dan FS-FAI menggunakan prinsip ARM untuk memilih fitur berinteraksi dan mengurangkan bilangan fitur secara efektif. Walau bagaimanapun, mereka memerlukan penyesuaian ambang secara manual dan penggunaan metrik asas yang boleh menjejaskan kestabilan algoritma. Keseluruhannya, kekangan-kekangan ini menunjukkan keperluan kepada pendekatan pemilihan fitur berinteraksi yang lebih sistematik, stabil dan boleh dijelaskan.

2.3.6 Rumusan Isu Pemilihan Fitur

Antara isu utama yang sering timbul ialah ketidakmampuan kaedah tradisional untuk mengesan interaksi antara fitur. Pendekatan univariat atau penilaian satu demi satu tidak mencukupi dalam mengenal pasti fitur yang hanya menjadi penting apabila digabungkan dengan fitur lain. Hal ini menyebabkan banyak informasi penting dalam bentuk interaksi tidak dimanfaatkan secara optimum (Cheng 2024). Selain itu, terdapat masalah pertindihan maklumat khususnya pada data kesihatan. Terdapat banyak fitur mempunyai hubungan korelasi tinggi yang boleh mengelirukan model dan memberi kesan negatif terhadap kestabilan ramalan. Jika tidak ditangani, perkara ini boleh menyebabkan “terlalu padan” atau prestasi yang tidak konsisten (Shanti 2024).

Masalah seterusnya adalah kebolehskalaan dan kecekapan pengiraan terutama dalam situasi dengan data yang berdimensi tinggi. Dalam hal ini, kaedah seperti kaedah CARM yang dicadangkan dalam objektif kajian menunjukkan potensi dalam mengenal pasti fitur berdasarkan hubungan dengan kelas sasaran (Jimenez et al. 2019; Tangherloni et al. 2024). Namun, kaedah ini juga menghadapi masalah ledakan

kombinatorial iaitu peningkatan secara eksponen pada bilangan kemungkinan kombinasi apabila bilangan fitur bertambah (Mamdouh Farghaly & Abd El-Hafeez 2023). Sebagai contoh, jika terdapat 20 fitur, jumlah subset kombinasi yang mungkin ialah 2^{20} iaitu 1,048,576. Hal ini boleh menyebabkan terlalu banyak peraturan dijana secara berlebihan termasuk yang bersifat berulang, tidak signifikan atau tidak relevan. Proses ini bukan sahaja meningkatkan kos pengiraan, tetapi juga boleh menjejaskan keberkesanan dan kebolehfahaman model. Oleh itu, strategi seperti penapisan dan penggunaan metrik tambahan selain metrik asas diperlukan untuk mengekang impak isu ini (Alam et al. 2021; Kaushik et al. 2023).

Dalam menangani isu kebolehfahaman, penggunaan kaedah logik kabur telah terbukti berkesan dalam konteks klinikal. Kaedah ini membolehkan pemodelan data berdasarkan keanggotaan beransur di mana ia menyokong penerangan hasil model dalam bentuk linguistik seperti "risiko tinggi" atau "simptom sederhana" yang seiring dengan cara pengamal perubatan memahami maklumat pesakit. Kajian menunjukkan bahawa teknik pemilihan fitur berasaskan kabur bukan sahaja membantu menjelaskan hubungan kompleks antara fitur malah menghasilkan model berasaskan peraturan yang mudah difahami oleh pengguna akhir (Jimenez et al. 2019; Tangherloni et al. 2024).

2.4 ANALISIS KESATUAN

Analisis kesatuan merupakan salah satu teknik perlombongan data yang digunakan untuk mengenal pasti hubungan atau corak menarik dalam set data yang besar. Teknik ini menjadi semakin relevan dalam pelbagai bidang terutamanya apabila jumlah data yang disimpan dalam pangkalan data bertambah dengan pesat. Konsep asas analisis kesatuan melibatkan set item dan set transaksi.

Peraturan Kesatuan (*Association Rule*, AR) ditulis dalam bentuk $A \Rightarrow B$ yang menunjukkan hubungan antara item A (anteseden) dan item B (konsekuen) dalam set transaksi. Bagi menilai sesuatu AR, dua ukuran utama digunakan iaitu sokongan (*Supp*) dan keyakinan (*Conf*). Sokongan merujuk kepada peratusan transaksi yang mengandungi set item $A \cup B$, manakala keyakinan merujuk kepada peratusan transaksi yang mengandungi B sekiranya A wujud. Sesuatu AR hanya dianggap menarik sekiranya ia memenuhi kriteria minimum yang telah ditetapkan iaitu ambang sokongan

minimum (*minSupp*) dan keyakinan minimum (*minConf*). Nilai ambang ini biasanya ditentukan oleh pengguna atau pakar bidang bergantung kepada keperluan kajian atau domain tertentu (Wang & Song, 2012).

2.4.1 Algoritma Apriori

Salah satu algoritma yang sering digunakan dalam Perlombongan Peraturan Kesatuan (*Association Rule Mining*, ARM) ialah algoritma *Apriori*. Algoritma ini diperkenalkan oleh Agrawal et al. (1993) yang melibatkan proses prinsip "penjanaan dari bawah ke atas". Ia bermaksud proses bermula dengan mengenal pasti set item kerap terkecil seperti *1-itemset* sebelum memperluaskannya ke *2-itemset*, *3-itemset* dan seterusnya sehingga tiada lagi set kerap yang boleh dijana.

Prinsip utama *Apriori* menyatakan bahawa jika sesuatu set item adalah kerap, maka semua subsetnya juga mesti kerap. Sebaliknya, jika suatu set item gagal memenuhi ambang *minSupp,* maka semua supersetnya boleh diabaikan. Proses algoritma ini terdiri daripada dua fasa utama. Fasa pertama melibatkan pengecaman semua set item kerap secara iteratif, manakala fasa kedua menumpukan kepada penjanaan peraturan kesatuan berdasarkan itemset kerap dengan menilai keyakinan peraturan.

2.4.2 Peraturan Kesatuan Kuat, Kelas dan Atomik

Peraturan kesatuan kuat (SAR) merujuk kepada peraturan bentuk A ⇒ B yang memenuhi ambang *minSupp* dan *minConf*.

Peraturan kesatuan kelas (CAR) pula ialah bentuk khusus SAR yang ditulis sebagai A ⇒ C yang digunakan untuk menghubungkan nilai fitur dengan nilai sasaran. Dalam konteks ini, anteseden (A) terdiri daripada satu atau lebih nilai fitur, manakala konsekuen (C) hanya mengandungi satu nilai sasaran. CAR membantu dalam pembinaan model ramalan kerana ia menunjukkan hubungan langsung antara fitur-fitur yang diperoleh daripada data dengan kelas sasaran yang ingin diramalkan.

Peraturan kesatuan atomik (AAR) pula ialah bentuk paling asas SAR di mana anteseden dan konsekuen masing-masing hanya mengandungi satu item. AAR sangat berguna untuk mengenal pasti pertindihan maklumat antara fitur.

2.4.3 Konsep Fitur Relevan, Bertindih dan Interaksi Berdasarkan Peraturan Kesatuan

Definisi fitur relevan, fitur bertindih dan fitur berinteraksi berdasarkan peraturan kesatuan dibincangkan secara terperinci seperti berikut:

a. Fitur Relevan

Fitur relevan merujuk kepada nilai fitur yang memberikan sumbangan bermakna terhadap penerangan konsep sasaran. Nilai fitur f_{ij} bagi fitur dianggap relevan jika dan hanya jika ia muncul dalam anteseden pada CAR. Bagi mentakrifkan fitur yang relevan (RF), definisi nilai fitur yang relevan (RFV) perlu ditakrifkan terdahulu seperti berikut:

Definisi 2.5. Nilai Fitur Relevan (RFV). Nilai tertentu f_{ij} bagi fitur F_i adalah relevan dengan konsep sasaran Y jika dan hanya jika:

$$\exists r \in \text{CARset} , f_{ij} \in r.\text{Ante} \quad \ldots(2.1)$$

di mana CARset ialah himpunan peraturan kelas yang diperoleh dan r.Ante merujuk kepada anteseden daripada peraturan r. Jika f_{ij} tidak memenuhi syarat ini, ia dianggap sebagai nilai fitur yang tidak relevan (iRFV).

Definisi 2.6. Fitur Relevan (RF). Fitur F_i adalah relevan dengan konsep sasaran Y jika dan hanya jika:

$$\exists f_{ij} \in F_i , \{f_{ij} | f_{ij} = RFV\} \neq \emptyset \quad \ldots(2.2)$$

Jika tidak, fitur F_i dianggap sebagai fitur yang tidak relevan (iRF). Ini bermakna sesuatu fitur adalah relevan apabila sekurang-kurangnya salah satu nilai fiturnya adalah nilai fitur yang relevan. Sebaliknya, bagi fitur yang tidak relevan, semua nilai fiturnya adalah tidak relevan.

b. Fitur Bertindih

Terdapat situasi di mana fitur relevan mungkin bertindih apabila dua atau lebih nilai fitur sangat berkorelasi dan sering muncul serentak dalam anteseden AR. Fenomena ini berlaku kerana perlombongan set item kerap tidak dibangunkan untuk mengesan pertindihan dalam data. Bagi menangani masalah ini, AAR digunakan untuk mengenal pasti korelasi antara nilai fitur dan menentukan nilai yang bertindih. Nilai fitur f dalam set nilai fitur (FVset) dianggap bertindih jika ia muncul sebagai konsekuen dalam AAR dan anteseden peraturan juga berada dalam FVset. Definisi ini boleh dirumuskan sebagai:

Definisi 2.7. Nilai Fitur Bertindih (RDFV). Nilai tertentu f dalam FVset dikatakan bertindih jika dan hanya jika:

$$\exists r \in \text{AARset}, (f = r.\text{Cons}) \wedge (r.\text{Ante} \subseteq \text{FVset}) \quad \ldots(2.3)$$

di mana r.Ante dan r.Cons masing-masing merujuk kepada anteseden dan konsekuen daripada peraturan r. Ini bermakna nilai fitur dianggap bertindih apabila ia muncul sebagai konsekuen dalam AARset dan anteseden peraturan juga berada dalam set nilai fitur yang sama.

Definisi 2.8. Fitur Bertindih (RDF). Fitur F_i bertindih jika dan hanya jika:

$$\forall f_{ij} \in F_i \, , \{f_{ij} | f_{ij} = \text{RDFV atau iRF}\} \neq \emptyset \quad \ldots(2.4)$$

di mana setiap nilai dalam fitur tersebut adalah RDFV atau sebahagian nilainya adalah bertindih manakala selebihnya iRF.

c. Fitur Berinteraksi

Biarkan $\text{FVset} = \{f_1, f_2, \ldots, f_k\}$ ialah set nilai fitur dengan k nilai fitur. Setiap ahli dalam FVset sepadan dengan satu nilai bagi satu fitur dalam Fset. Andaikan $A \subset \text{FVset}$, $\text{A} \neq \emptyset$ dan $B = \text{FVset} - A$. Jadikan y sebagai satu nilai bagi konsep sasaran Y dan biarkan $Conf(r)$ menjadi ukuran keyakinan bagi peraturan r manakala r_F, r_A dan r_B ialah CAR

di mana masing-masing FVset $\Rightarrow \{Y = y\}, A \Rightarrow \{Y = y\}$ dan $B \Rightarrow \{Y = y\}$. Definisi nilai fitur berinteraksi dapat dirumuskan seperti berikut:

Definisi 2.9. Nilai Fitur Berinteraksi yang ke-k. Nilai fitur ke-k dalam FVset dikatakan saling berinteraksi antara satu sama lain jika dan hanya jika:

$$Conf(r_F) > Conf(r_A) \wedge Conf(r_F) > Conf(r_B) \quad \ldots(2.5)$$

di mana interaksi berlaku apabila keyakinan peraturan bagi gabungan fitur, $Conf(r_F)$ lebih tinggi daripada keyakinan peraturan nilai individu, $Conf(r_A)$ dan $Conf(r_B)$. Set nilai fitur A dan B ini dikatakan saling berinteraksi antara satu sama lain. Fitur berinteraksi pula dapat ditakrifkan seperti berikut:

Definisi 2.9. Fitur Interaksi. Biarkan $Fset = \{F_1, F_2, \ldots, F_k\}$ menjadi subset fitur yang terdiri daripada k fitur dan VAset sebagai set tugasan nilainya. Fitur-fitur $F_1, F_2, \ldots, F_k$ dikatakan saling berinteraksi antara satu sama lain jika dan hanya jika:

$$\{fset = \begin{matrix} \exists fset \in \text{VAset}, \\ \text{FVset dengan interaksi nilai } F_k \end{matrix}\} \neq \emptyset \quad \ldots(2.6)$$

di mana ia bermaksud bahawa terdapat sekurang-kurangnya satu kombinasi tugasan nilai fitur $fset$ dalam VAset yang mengandungi nilai-nilai fitur yang berinteraksi khususnya pada fitur ke-k. Interaksi antara fitur dalam subset tertentu boleh dianalisis melalui interaksi tugasan nilainya.

2.4.4 Perlombongan Peraturan Kesatuan Pada Penyelidikan Kanser Payudara

ARM telah terbukti membantu dalam penyelidikan kanser payudara terutamanya dalam mengenal pasti corak dan hubungan penting dalam set data klinikal dan genomik. Sebagai contoh, Arslan et al. (2020) menggunakan *Classification based on Association Rules* (CBAR) dan *Regularized Class Association Rules* (RCAR) untuk set data WDBC dan mencapai 95.4% ACC. Selain itu, Ahmed et al. (2019) menggabungkan CARM dengan DNN untuk mengelaskan data klinikal kanser ini. Kajian ini mencapai 100% ACC pada set data WOBC menggunakan 872 peraturan yang telah dijana. Walau

bagaimanapun, risiko “terlalu padan” dan kekangan set data terhad menjadi cabaran utama.

Penggunaan ARM juga digunakan oleh Sowan et al. (2023) di mana mereka memperkenalkan pendekatan baru yang menggabungkan *Weighted Average of Relative Frequency* (WARF) dan PSO dengan *Predictive Classification Based on Associations* (PCBA) bagi pengelasan kanser. Walau bagaimanapun, kos pengiraan dan penalaan parameter yang berkaitan dengan PSO menghadirkan cabaran besar terutamanya untuk set data bersaiz besar. Sementara itu, Kharya et al. (2022) menangani isu kematian akibat kanser payudara yang sering berpunca daripada kelewatan pengesanan dan rawatan. Kajian tersebut mencadangkan rangka kerja hibrid *Weighted Bayesian Belief Network* (WBBN). Hasil kajian mencapai 97.18% ACC dan mengenal pasti corak signifikan melalui peraturan berwajaran. Walau bagaimanapun, kebergantungan pada pemberian berat yang tepat mengehadkan kebolehgunaannya.

Pengelasan massa kanser pada mamografi dikaji oleh Keyvanpour et al. (2021) di mana kajian tersebut mencadangkan *Weighted Association Rule Mining* (WARM) yang secara automatik mengenal pasti corak kerap secara berasingan dan menghasilkan peraturan dengan memberikan pemberat pada corak tersebut. Hasil kajian mendapati bahawa WARM memberikan penambahbaikan ketara dalam nilai ACC dan SEN. Namun, kaedah ini memerlukan penambahbaikan dari segi penghasilan peraturan positif dan peraturan negatif untuk meningkatkan ketepatan pengelasan.

Di samping itu, Deshpande et al. (2015) mencadangkan *Texture Based Associative Classifier* (TBAC) untuk meningkatkan pengelasan mamografi di mana ia melibatkan penjanaan AR berdasarkan hubungan antara fitur tekstur. AR ini dimodelkan untuk mengelaskan mamografi kepada tiga kategori iaitu normal, benigna atau malignan. Hasil kajian menunjukkan ACC sebanyak 94.66% mengatasi model pengelasan yang lain. Kelebihan teknik TBAC ialah ia mampu menangani data tidak seimbang namun ia memberi ACC lebih rendah untuk kelas malignan iaitu 84%. Teknik ini juga terhad kepada set data MIAS dan memerlukan validasi pada set data yang lebih besar dan pelbagai.

Beberapa kajian telah berusaha memperbaiki teknik ARM dengan kaedah yang inovatif. Liewlom (2021) meneliti masalah ketidakseimbangan nisbah kelas dalam set data, Kajian ini memberi tumpuan kepada teknik *Classification Based on Associations* (CBA) dengan mencadangkan ukuran prestasi baru yang dipanggil *Profitability-of-Interestingness* (PoI) untuk memangkas peraturan positif dengan *Supp* atau *Conf* yang rendah tetapi penting. Hasilnya menunjukkan bahawa peraturan yang dipangkas menggunakan PoI memberikan ketepatan yang setara dengan CARM tradisional tetapi lebih mudah difahami kerana jumlah peraturan yang lebih rendah. Pengelasan berdasarkan peraturan yang dipangkas juga lebih ringkas dan bebas daripada peraturan yang mengelirukan. Namun, pendekatan ini terhad kepada set data kecil dan memerlukan pengesahan lanjut pada set data pelbagai.

Di samping itu, Alwidian et al. (2018) memperkenalkan algoritma *Weighted Classification Based on Association Rules* (WCBA) untuk meningkatkan ketepatan pengelasan kanser. Hasil kajian menunjukkan bahawa WCBA menghasilkan peraturan yang lebih tepat dan relevan dengan prestasi yang lebih baik dalam kebanyakan kes. Namun begitu, kebergantungan pada pakar domain untuk pemberian berat bagi fitur dan kebolehgunaan terhad di luar set data kanser payudara membawa kepada keperluan kepada penyelidikan yang lebih baik pada masa hadapan.

Pengenalpastian pola menggunakan ARM telah dikaji oleh para penyelidik di mana teknik ini membantu mengenal pasti faktor risiko dan pola prognostik pada kanser payudara. Dhanaseelan & Sutha (2021) memperkenalkan algoritma *Improved Fuzzy Frequent Pattern Mining* (IFFP) yang menggunakan logik kabur untuk mengenal pasti pola penting pada set data WOBC. Hasil kajian menunjukkan bahawa kehadiran sel-sel abnormal dalam tumor seperti kehadiran nukleus yang tidak normal menjadi penentu utama untuk diagnosis. Sebaliknya, fitur yang berkaitan dengan pembelahan sel tumor tidak memberikan sumbangan besar kepada pengenalpastian kanser payudara. Walaupun algoritma ini menunjukkan prestasi yang sangat baik dalam pengesanan kanser payudara, batasan utamanya ialah penggunaan set data yang agak kecil yang mungkin tidak mencerminkan variasi dunia sebenar.

Sementara itu, Oladipupo et al. (2021) memperkenalkan *Interval Type-2 Fuzzy Association Rule Mining (IT2FARM)* yang menggabungkan teknik pengkaburan *Hao* dan *Mendel* serta algoritma *FP-Growth* untuk menganalisis data kanser payudara. Kajian ini telah mengenal pasti bahawa pola berkaitan keseragaman bentuk sel dan kehadiran nukleus abnormal merupakan faktor penting bagi diagnostik kanser ini. Pendekatan ini juga berjaya mengurangkan peraturan berlebihan dan penggunaan memori menjadikannya lebih cekap. Namun, kebergantungan kepada input pakar mungkin membawa perbezaan pada hasil keputusan. Di samping itu, Jajroudi et al. (2024) menggunakan algoritma *Apriori* untuk mengenal pasti pola diagnostik kanser payudara dengan menganalisis fitur radiomik yang diekstrak daripada mamografi. Menurut analisis, fitur utama untuk tumor benigna termasuk sempadan jisim, penekanan jangka panjang mendatar dan entropi pembezaan. Sebaliknya, tumor malignan dikaitkan dengan kontras tinggi, korelasi dan penilaian klinikal tertentu.

Terdapat juga beberapa kajian memberi tumpuan kepada pola risiko dan gaya hidup yang cenderung ke arah kanser ini. Sebagai contoh, Kabir et al. (2018) menggunakan algoritma ARM dan model logit binari untuk menganalisis risiko berdasarkan faktor demografi dan sejarah perubatan menggunakan set data *Breast Cancer Surveillance Consortium* (BCSC). Kajian ini mendapati bahawa individu bukan Hispanik tanpa sejarah keluarga kanser payudara, tiada biopsi sebelumnya dan BMI dalam julat normal cenderung tidak menghidap kanser. Sebaliknya, individu Hispanik berusia 85 tahun ke atas dengan kelahiran pertama sebelum umur 20 tahun, kepadatan payudara rendah dan pernah menjalani biopsi berisiko tinggi menghidap kanser. Walau bagaimanapun, kajian ini memerlukan pengesahan lanjut untuk memastikan keberkesanannya pada populasi lain.

Selain itu, Li et al. (2019) menggunakan pendekatan *Item Association Rules* (IAR) dan model LR, DT, RF, XGB, GB dan MLP untuk mengenal pasti faktor risiko kanser payudara yang boleh diubah suai dan tidak boleh diubah suai. Hasil kajian mendapati faktor gaya hidup seperti diet tidak seimbang, kurang tidur dan gaya hidup tidak sihat yang dapat diubah untuk mengurangkan risiko kanser payudara. Namun, faktor risiko tidak boleh diubah seperti sejarah keluarga kanser dan usia haid pertama yang lewat dikenal pasti sebagai penyumbang kepada peningkatan risiko.

Pola yang lebih spesifik berdasarkan lokasi dan demografi menggunakan ARM turut dikenal pasti. Adam & Wieder (2024) menggunakan *Temporal Association Rule* (TAR) untuk mengenal pasti perbezaan dalam kesan buruk terhadap rawatan susulan pesakit kanser payudara terutamanya dalam kalangan pesakit Afrika-Amerika. Penggunaan algoritma *FP-Growth* menunjukkan bahawa terdapat perbezaan ketara dalam jenis kesan buruk yang berkaitan dengan rawatan bergantung kepada bangsa, tahap penyakit iaitu lokal atau metastatik dan tempat rawatan iaitu hospital atau klinik persendirian. Kajian ini menghadapi batasan seperti ketiadaan pembezaan antara kesan buruk ringan dan teruk serta keperluan pengesahan tambahan pada populasi yang lebih muda. Selain itu, Alharith et al. (2024) menggunakan algoritma *FP-Growth* untuk mengenal pasti pola kanser di Sudan. Kajian ini menganalisis pola berdasarkan faktor seperti geografi, jantina dan kumpulan umur. Namun, kajian ini terhad kepada ketiadaan sistem rekod kesihatan elektronik yang seragam di Sudan dan menyukarkan analisis data yang lebih menyeluruh.

Jadual 2.3 menunjukkan ringkasan kajian kesusasteraan bagi kaedah ARM pada penyelidikan kanser payudara.

Jadual 2.3 Ringkasan kajian kesusasteraan bagi kaedah ARM pada penyelidikan kanser payudara

Rujukan	Set Data	Kaedah ARM	Kaedah Tambahan	Objektif	Hasil Kajian	Limitasi Kajian
Ahmed et al. (2019)	WOBC	CARM	DNN	Pengelasan	ACC=100%	Risiko "terlalu padan"
Arslan et al. (2020)	WDBC	RCAR, CBAR	-	Pengelasan	ACC=95.4%	Hanya metrik asas digunakan
Sowan et al. (2023)	WOBC, BCR	CBA	WARF, PSO	Pengelasan	(ACC) BCR=75.6%, WOBC=97.3%	Kos pengiraan tinggi
Kharya et al. (2022)	WOBC	ARM	WBBN	Pengelasan	ACC=97.18%	Kebergantungan pada berat
Keyvanpour et al. (2021)	MIAS	WARM	-	Pengelasan	ACC=95.15%	Set data terhad
Deshpande et al. (2015)	MIAS	TBAC	-	Pengelasan	ACC=94.66%	Set data terhad
Liewlom (2021)	BCR	CBA	PoI	Pengelasan	Peraturan ringkas dan mudah difahami	Hanya metrik asas digunakan
Alwidian et al. (2018)	BCR, WOBC	WCBA	HM	Pengelasan	(ACC) BCR=73.26%, WOBC=97.4%	Kebergantungan pada pakar domain
Dhanaseelan & Sutha (2021)	WOBC	IFFP	Logik kabur	Pengenalpastian corak	Fitur penting dikenalpasti	Set data terhad
Oladipupo et al. (2021)	WOBC	IT2FARM	Hao dan Mendel	Pengenalpastian corak	Pengurangan peraturan dan lebih cekap	Kebergantungan pada input pakar
Jajroudi et al. (2024)	DDSM	ARM	CART, ANN, SVM	Pengenalpastian corak	Pengenalpastian fitur bagi pesakit	Tiada penilaian statistik
Kabir et al. (2018)	BCSC	ARM	Model logit binari	Pengenalpastian corak	Pengenalpastian risiko bagi pesakit	Pengesahan lanjut pada populasi lain
Li et al. (2019)	Persendirian	IAR	LR, DT, RF, XGB, GB, MLP	Pengenalpastian corak	AUC GB=0.9655	Ketidakseimbangan data
Adam & Wieder (2024)	SEER	TAR	-	Pengenalpastian corak	Pengenalpastian kesan selepas pembedahan	Pengesahan lanjut pada populasi lain
Alharith et al. (2024)	Persendirian	FP-Growth	-	Pengenalpastian corak	Pengenalpastian faktor kanser	Set data terhad

Berdasarkan Jadual 2.3, pelbagai set data digunakan pada ARM termasuk WOBC, WDBC, BCR, MIAS, DDSM, BCSC, SEER dan beberapa set data persendirian. Kaedah ARM digunakan secara meluas dalam kajian-kajian ini sama ada dalam bentuk klasik seperti CBA, RCAR, CBAR, WCBA atau digabungkan dengan kaedah lain seperti DNN, WBBN, PoI, Logik Kabur, CART, ANN dan SVM bagi meningkatkan prestasi ramalan dan kefahaman klinikal. Objektif utama kajian-kajian ini adalah untuk melakukan pengelasan serta pengenalpastian corak dan fitur penting dalam data kanser payudara. Hasil kajian menunjukkan bahawa ketepatan (ACC) yang tinggi dapat dicapai dalam kebanyakan set data. Sebagai contoh, WOBC (97–100%), WDBC (95.4%), MIAS (94–95%) dan BCR (73–97%). Penggunaan ARM yang digabungkan dengan algoritma tambahan bukan sahaja meningkatkan kecekapan dan mengurangkan bilangan peraturan, tetapi juga membolehkan penemuan corak atau fitur penting. Walau bagaimanapun, meskipun ARM mempunyai potensi yang tinggi dalam mengenal pasti corak serta fitur penting dalam data kanser payudara, beberapa kekangan masih wujud, antaranya ketiadaan mekanisme pemilihan fitur yang benar-benar menilai interaksi antara fitur, kebergantungan kepada ambang peraturan yang ditetapkan secara manual, serta kekurangan proses pengesahan yang telus untuk menilai sumbangan setiap fitur dalam model. Kekangan-kekangan ini secara langsung mengukuhkan keperluan kajian ini untuk membangunkan model hibrid pemilihan fitur berasaskan interaksi yang mempunyai proses pengesahan dan penalaan automatik yang lebih sistematik.

2.4.5 Pemilihan Fitur Menggunakan Perlombongan Peraturan Kesatuan

ARM adalah salah satu kaedah yang semakin digunakan untuk pemilihan fitur dengan mengenal pasti hubungan antara fitur. Kajian awal seperti Xie et al. (2009) menerapkan algoritma *Apriori* sebagai teknik pemilihan fitur di mana ia menapis atribut dengan nilai *Supp* yang tinggi dengan kelas sasaran. Walaupun kaedah ini menunjukkan keupayaan untuk mengurangkan bilangan fitur secara signifikan sambil mengekalkan ketepatan pengelasan yang dapat diterima, namun ia menghadapi cabaran utama dalam hal skalabiliti. Pengiraan masa algoritma juga menjadi tinggi terutamanya apabila digunakan secara langsung untuk melombong peraturan kesatuan tanpa pengoptimuman.

Selain itu, Rajeswari (2015) pula mengoptimumkan algoritma *Apriori* untuk pemilihan fitur dengan menambah strategi pengurangan set data. Dengan saiz set data yang lebih kecil, algoritma *Apriori* menjadi lebih cekap kerana bilangan iterasi dan masa pengiraan dikurangkan. Walau bagaimanapun, kaedah ini mempunyai kekangan di mana pengurangan set data ini hanya menumpukan kepada satu kelas sasaran sahaja dan mengabaikan kelas lain. Ia boleh menjadi masalah jika semua kelas dalam set data adalah sama penting.

Pemilihan fitur berdasarkan ARM ini juga dikembangkan dengan pengenalan kaedah yang lebih cekap dan disesuaikan untuk pelbagai domain. Sebagai contoh, Wang & Song (2012) memperkenalkan kaedah ARM menggunakan algoritma *FP-Growth* untuk mengenal pasti fitur relevan dan berinteraksi serta menyingkirkan fitur bertindih. Keberkesanan FEAST diuji menggunakan set data sintetik dan dunia sebenar dan telah menunjukkan peningkatan prestasi pengelasan dan keupayaan mengurangkan bilangan fitur berbanding algoritma pemilihan fitur lain. Walau bagaimanapun, nilai ambang *Supp* dan *Conf* yang sesuai diperlukan berbeza mengikut jenis set data. Oleh itu, Wang & Song (2013) telah menambah baik algoritma FEAST dalam kajian lanjutan dengan memperkenalkan kaedah PLSR untuk meramalkan nilai ambang tersebut.

Kaedah ARM bagi set data berdimensi tinggi telah memberi cabaran kepada penyelidik disebabkan oleh kos pengiraan yang tinggi, jumlah set calon yang banyak dan keperluan untuk melakukan pelbagai imbasan pangkalan data. Oleh itu, Harikumar et al. (2017) telah mencadangkan pendekatan baharu untuk memperbaiki algoritma *Apriori* dengan menggabungkan IG dan penguraian QR iaitu suatu kaedah dalam algebra linear. Walau bagaimanapun, pendekatan ini mempunyai kekangan di mana pemilihan nilai ambang pada IG memerlukan penalaan yang teliti kerana nilai ambang yang tidak sesuai boleh menjejaskan prestasi algoritma.

Di samping itu, terdapat juga penyelidik yang menyasarkan untuk menangani masalah dimensi tinggi dalam data besar dan data imej menggunakan pemilihan fitur berasaskan ARM. Sebagai contoh, kajian pertama oleh Kaoungku, Suksut, et al. (2017) mencadangkan algoritma yang menggunakan CARM untuk mengenal pasti fitur yang relevan dan menyingkirkan fitur yang tidak diperlukan. Kaedah ini berjaya

mengurangkan bilangan fitur dan meningkatkan ACC pengelasan tetapi beberapa set data menunjukkan ACC lebih rendah berbanding data asal tanpa pemilihan fitur. Seterusnya, kajian kedua oleh Kaoungku, Kerdprasop, et al. (2017) telah memperkenalkan strategi gabungan beberapa kaedah pemilihan fitur termasuk ARM, CFS, Consistency, korelasi, IG serta teknik pengelompokan *k-Means* untuk mengelompokkan skor kedudukan fitur. Strategi ini berjaya mengurangkan dimensi set data sambil meningkatkan ketepatan pengelasan, tetapi keberkesanan pada set data tertentu tidak konsisten. Dalam kajian berikutnya, Kaoungku et al. (2018) memfokuskan kepada data imej berdimensi tinggi seperti imej satelit dan mencadangkan pendekatan tiga fasa yang menggabungkan ARM dan kriteria Lebar Siluet. Walaupun pendekatan ini menghasilkan model pengelasan yang tepat dan berjaya menyelesaikan masalah dimensi tinggi, namun kekangan utamanya termasuk kebergantungan kepada parameter pengelompokan yang memerlukan penalaan yang intensif.

Selain itu, Veroneze et al. (2024) turut memberi tumpuan kepada keselamatan siber. Kajian ini mencadangkan algoritma *Feature Selection based on Associative Classification* (FSbAC) yang menggunakan teknik ARM untuk mengenal pasti fitur penting bagi setiap jenis *packer* secara spesifik. Hasil kajian menunjukkan FSbAC berjaya mengurangkan bilangan fitur yang diperlukan dan meningkatkan kecekapan proses pengelasan. Walau bagaimanapun, keberkesanan FSbAC sedikit terjejas apabila digunakan pada set data dunia sebenar dan menunjukkan keperluan untuk menjadikannya lebih kuat terhadap ancaman baharu.

Dalam aplikasi penjagaan kesihatan, Veeramuthu & Meenakshi (2014) membincangkan keperluan untuk sistem pengelasan automatik yang dapat menangani pelbagai jenis imej CT. Kajian ini mencadangkan sistem yang menggunakan ARM untuk mengenal pasti fitur yang relevan berdasarkan set kerap dengan label kelas. Fitur ini kemudian digunakan untuk pengelasan imej menggunakan model pengelas kNN dan NB. Hasil eksperimen menunjukkan peningkatan ACC sebanyak 8.89% untuk kNN dan 2.96% untuk NB apabila pengurangan fitur dilakukan dengan ARM. Walaupun sistem ini menunjukkan keberkesanan dalam proses pengurangan fitur, namun kajian ini memerlukan penyelidikan lanjut untuk meningkatkan prestasi pengelasan imej

perubatan yang lebih kompleks. Selain itu, Alam et al. (2021) pula mencadangkan kaedah untuk mengenal pasti faktor prognosis penting bagi mesothelioma malignan (MM) iaitu sejenis kanser jarang berlaku yang berasal daripada sel mesothelial. Kajian ini menerapkan teknik ARM seperti algoritma *Apriori* dan *FP-Growth*. Hasil kajian menunjukkan lima faktor prognosis utama untuk MM yang dikenal pasti melalui analisis nilai *Supp*, *Conf* dan *Lift*. Walau bagaimanapun, kajian ini menghadapi kekangan di mana penambahan set data berskala besar berpotensi meningkatkan masa pelaksanaan secara signifikan.

Terdapat juga domain lain yang dikaji menggunakan ARM sebagai teknik pemilihan fitur. Sebagai contoh, Pacheco-Ortiz et al. (2020) memperluaskan ARM dalam domain pendidikan khususnya untuk membangunkan sistem Ujian Adaptif Berkomputer (CAT). Sistem ini bertujuan meningkatkan proses penilaian pelajar dengan memilih item bagi ujian secara adaptif berdasarkan pengetahuan peserta ujian. Kajian ini menganalisis empat algoritma ARM iaitu *Apriori*, *FP-Growth, PredictiveAprior* dan *Tertius* untuk mengenal pasti algoritma yang paling sesuai bagi pemilihan item. Hasil kajian menunjukkan algoritma *Apriori* sebagai kaedah terbaik kerana ia menghasilkan peraturan dengan *Supp* dan *Conf* yang tinggi dalam masa yang lebih singkat. Namun, kajian ini mempunyai kekangan di mana ia masih dalam peringkat eksperimen awal dan belum diuji secara menyeluruh pada populasi pelajar yang lebih luas.

Secara keseluruhannya, kajian kesusasteraan menunjukkan trend yang semakin ketara ke arah penggunaan pemilihan fitur berasaskan ARM untuk mengenal pasti fitur yang relevan, mengesan interaksi penting antara fitur dan mengurangkan pertindihan. Namun begitu, beberapa cabaran utama masih wujud, terutamanya keperluan penalaan ambang yang optimum dan ketiadaan proses pengesahan yang telus bagi fitur berinteraksi melalui teknik XAI. Jadual 2.4 merumuskan kajian-kajian terdahulu berkaitan kaedah pemilihan fitur berasaskan ARM.

Jadual 2.4 Ringkasan kajian kesusasteraan bagi kaedah pemilihan fitur berdasarkan ARM

Rujukan	Domain	Kaedah ARM	Kaedah Tambahan	Hasil Kajian	Limitasi Kajian
Xie et al. (2009)	Pelbagai	Apriori	-	Bilangan fitur berkurang	Kompleksiti masa
Chawla (2010)	Kejuruteraan perisian	ARN	-	Pemahaman interaksi antara fitur	Digunakan hanya pada data keratan rentas
S. J. Zhang & Zhou (2011)	Pelbagai	CARM	DT	Bilangan fitur berkurang	Kompleksiti pengiraan
Waseem et al. (2013)	Pelbagai	ARM	JRipper	Bilangan fitur berkurang	Prestasi pengelasan tidak konsisten
Rajeswari (2015)	Kesihatan	Apriori	DT	Masa pengiraan berkurang	Hanya menumpukan satu kelas sasaran
Guangtao Wang & Song (2012)	Pelbagai	FP-Growth	DT, NB, PART	Interaksi antara fitur dikenal pasti	Bergantung pada nilai ambang
Wang & Song (2013)	Pelbagai	FP-Growth + PLSR	NB, IB1, DT, PART, SVM	Peningkatan ketepatan	Hanya metrik asas digunakan
Harikumar et al. (2017)	Kewangan	Apriori	IG, QR, Entropi	Bilangan fitur berkurang	Bergantung pada nilai ambang
Kaoungku, Suksut, et al. (2017)	Pelbagai	CARM	-	Bilangan fitur berkurang	Prestasi pengelasan tidak konsisten
Kaoungku, Kerdprasop, et al. (2017)	Pelbagai	ARM	CFS, Korelasi, IG, k-Means	Bilangan fitur berkurang	Prestasi pengelasan tidak konsisten
Kaoungku et al. (2018)	Imej satelit	ARM	k-Means, Lebar Siluet	Bilangan fitur berkurang	Hanya metrik asas digunakan
Sheikhan et al. (2011)	Keselamatan siber	FAR	ARTMAP	Bilangan fitur berkurang	Set data tunggal
Veroneze et al. (2024)	Keselamatan siber	Apriori	CBA, DT, Ripper, XGB	Bilangan fitur dan masa pengiraaan berkurang	Prestasi pengelasan rendah pada set data dunia sebenar
Mamdouh Farghaly & Abd El-Hafeez (2023)	Pengelasan teks	Apriori	DT, NB, LR	Bilangan fitur berkurang	Hanya metrik asas digunakan
Veeramuthu & Meenakshi (2014)	Kesihatan	Apriori	kNN, NB	Bilangan fitur berkurang dan prestasi meningkat	Hanya metrik asas digunakan
Alam et al. (2021)	Kesihatan	Apriori, FP-Growth	-	Pengesahan faktor klinikal dengan kaedah ARM dan FS	Peningkatan masa pelaksanaan pada set data bersaiz besar
Pacheco-Ortiz et al. (2020)	Pendidikan	Apriori, FP-Growth, PredictiveApriori, Tertius	-	Masa pengiraan Apriori singkat	Belum diuji pada populasi pelajar yang lebih luas
Li et al. (2019)	Kejuruteraan	Apriori	Kebarangkalian	Prestasi pengelsan tinggi	Hubungan fitur kurang jelas

Berdasarkan Jadual 2.4, terdapat kajian terdahulu menunjukkan penggunaan pelbagai kaedah ARM seperti *Apriori*, *FP-Growth*, CARM, ARN dan FAR untuk mengurangkan bilangan fitur dan mengenal pasti interaksi antara fitur dalam pelbagai domain termasuk kesihatan, kejuruteraan, kewangan, keselamatan siber, pendidikan, imej satelit dan pengelasan teks. Beberapa kajian menggabungkan ARM dengan kaedah tambahan seperti DT, NB, SVM, k-Means dan *Information Gain* untuk meningkatkan ketepatan pengelasan. Secara umumnya, hasil kajian menunjukkan pengurangan fitur yang konsisten dan peningkatan prestasi model terutamanya apabila ARM digabungkan dengan algoritma tambahan atau pelbagai metrik. Namun, beberapa isu utama timbul termasuk kos pengiraan dan masa pelaksanaan yang tinggi bagi set data bersaiz besar atau berdimensi tinggi, prestasi pengelasan yang tidak konsisten, kebergantungan kepada nilai ambang tertentu, serta penggunaan hanya metrik asas seperti *Support*, *Confidence* dan *Lift* tanpa penilaian tambahan. Terdapat kajian juga hanya menumpukan pada satu kelas sasaran dengan menggunakan set data tunggal atau tidak menguji kebolehgunaan kaedah pada populasi yang lebih luas. Ini menunjukkan bahawa walaupun ARM berkesan untuk penapisan fitur dan pemahaman interaksi, namun masih wujud jurang dalam kebolehpercayaan dan pengoptimuman yang memerlukan pendekatan tambahan bagi domain dunia sebenar. Justeru, ringkasan dalam jadual ini secara langsung menyokong objektif kajian dan mengukuhkan keperluan kepada pendekatan yang dicadangkan dalam kajian ini.

2.5 MODEL KECERDASAN BUATAN YANG BOLEH DIJELASKAN

Kecerdasan buatan yang boleh dijelaskan (*eXplainable Artificial Intelligent*, XAI) adalah satu cabang dalam AI yang bertujuan untuk mengatasi kekurangan model "kotak hitam" tradisional dengan memberikan ketelusan dan kefahaman terhadap bagaimana sesuatu keputusan dihasilkan (Gunning & Aha 2019).

Keperluan untuk model XAI semakin penting pada model pembelajaran yang kompleks seperti pembelajaran dalaman kerana kerumitan reka bentuknya menjadikan sesuatu keputusan sukar diterangkan walaupun prestasi ketepatannya tinggi. Sebaliknya, keputusan bagi model mudah seperti DT lebih mudah diterangkan kerana reka bentuknya yang lebih ringkas namun prestasi ketepatannya tidak setinggi model kompleks (Linardatos et al. 2021). Isu justifikasi yang kurang jelas ini menjadi lebih

ketara dalam sektor kritikal seperti penjagaan kesihatan (Challen et al. 2019) kerana keputusan perlu dipercayai sepenuhnya. XAI memainkan peranan penting dalam meningkatkan ketelusan dalam penggunaan AI. Ketelusan ini penting untuk memupuk kepercayaan di kalangan pengguna (Chamola et al. 2023). Tidak seperti model tradisional yang sukar difahami, XAI memberikan penjelasan yang jelas tentang bagaimana keputusan dihasilkan dan menjadikannya lebih boleh dipercayai (Albahri et al. 2023). Sebagai contoh, dalam bidang perubatan, XAI dapat membantu doktor memahami rasional di sebalik cadangan rawatan (Amann et al. 2020; Žlahtič et al. 2024).

2.5.1 Pendekatan Model XAI

Pelbagai model XAI telah dikembangkan dan boleh dikategorikan berdasarkan tahap, skop dan jenis kaedah (Singh et al., 2020). Salah satu kategori utama model XAI ialah berdasarkan tahap pelaksanaannya iaitu *ante-hoc* dan *pasca-hoc*. Kaedah *ante-hoc* direka yang bersifat intrinsik di mana struktur model jelas dan mudah difahami sejak awal pembangunan. Contoh model bagi kaedah ini seperti LR dan DT di mana model-model ini yang membolehkan pengguna memahami keputusan tanpa memerlukan alat tambahan (Marino et al. 2025). Sebaliknya, kaedah pasca-hoc digunakan untuk menjelaskan model kompleks seperti ANN yang sering dianggap sebagai "kotak hitam." Teknik seperti SHAP dan *Local Interpretable Model-agnostic Explanations* (LIME) membantu memberikan penjelasan tentang bagaimana setiap fitur input mempengaruhi keputusan model. Teknik SHAP menggunakan pendekatan teori permainan untuk menganalisis sumbangan fitur-fitur input (Lundberg & Lee 2017) manakala teknik LIME membina model lokal yang lebih sederhana untuk memberikan penjelasan pada tahap keputusan individu (Ribeiro et al. 2016).

Skop penjelasan dalam XAI pula dibahagikan kepada penjelasan global dan penjelasan tempatan. Penjelasan global menerangkan bagaimana model berfungsi secara keseluruhan. Ia melihat kepada seluruh data dan tunjukkan fitur-fitur mana yang paling penting untuk hasil model. Sebaliknya, penjelasan tempatan pula menerangkan keputusan untuk satu kes sahaja. Dalam imej perubatan, penjelasan tempatan juga boleh menunjukkan bahagian imej yang menyebabkan model mengesan tumor dalam MRI (Van der Velden et al. 2022).

Kaedah XAI juga boleh dikategorikan mengikut jenisnya iaitu model-spesifik dan model-agnostik (Adadi & Berrada 2018). Kaedah penjelasan model-spesifik hanya boleh digunakan untuk jenis model pembelajaran tertentu sahaja. Sebagai contoh, teknik peta salien digunakan khas untuk model CNN sahaja bagi memberi gambaran visual terhadap bahagian imej yang menyumbang kepada keputusan model (Van der Velden et al. 2022). Sebaliknya, kaedah model-agnostik tidak terikat kepada sebarang jenis model pembelajaran mesin. Ia memisahkan proses ramalan daripada proses penjelasan dan biasanya digunakan selepas model dibina. Contoh kaedah model-agnostik seperti SHAP dan LIME (Ladbury et al. 2022).

2.5.2 Penggunaan Model XAI dalam Penyelidikan Kanser Payudara

Penggunaan model XAI dalam penyelidikan kanser payudara semakin berkembang seiring dengan keperluan kepada sistem AI yang telus dan dipercayai. Kemampuan untuk memahami dan mentafsir keputusan AI amat penting bagi menyokong rawatan serta membina keyakinan dalam kalangan pengamal kesihatan.

Kajian oleh Srinivasu et al. (2024) telah membangunkan gabungan model CatBoost dan MLP bagi meningkatkan keupayaan diagnosis awal kanser payudara. Kajian ini menggunakan analisis varians iaitu ANOVA bagi pemilihan fitur manakala model SHAP untuk menjelaskan sumbangan setiap fitur terhadap keputusan. Keputusan menunjukkan peningkatan ketara dari segi ketepatan dan kefahaman model serta sesuai untuk digunakan dalam persekitaran klinikal sebenar. Selain itu, Amoroso et al. (2021) telah membangunkan rangka kerja XAI berasaskan pengurangan dimensi adaptif bagi mengenal pasti fitur klinikal paling signifikan yang membezakan kelompok pesakit kanser payudara. Kajian mendapati bahawa subjenis molekul adalah penentu utama dalam pengelompokan dan hasil ini sejajar dengan garis panduan terapi sedia ada menunjukkan bahawa XAI boleh menghasilkan profil rawatan berdasarkan data malah tetap konsisten dengan amalan klinikal.

Penggunaan model XAI juga telah dikaji dalam bidang penyelidikan omik seperti kajian Chakraborty et al. (2021) yang meramalkan prognosis pesakit kanser payudara. Kajian ini mendedahkan hubungan antara faktor biologi dan hasil rawatan yang lebih baik. Kajian Silva-Aravena et al. (2023) pula membangunkan satu sistem

sokongan keputusan klinikal berasaskan pembelajaran mesin dan XAI bagi membantu keadaan selepas pandemik COVID-19 yang telah menjejaskan proses rawatan pesakit kanser payudara. Sistem ini membantu dalam mengenal pasti pesakit yang berisiko, menentukan faktor-faktor risiko utama serta memberikan makluman awal pada setiap pesakit. Model XGB menunjukkan 81.3% ACC dalam ramalan manakala model SHAP digunakan untuk menjelaskan faktor-faktor yang mempengaruhi keputusan tersebut. Bai et al. (2024) pula mengulas pelbagai aplikasi XAI seperti LIME, SHAP dan Grad-CAM dalam gabungan dengan model pembelajaran mesin dan pembelajaran dalaman untuk pengelasan imej perubatan seperti mamografi dan ultrabunyi. Kajian ini membuktikan bahawa XAI memperkukuh kepercayaan pihak pengamal perubatan terhadap sistem AI dengan menyediakan justifikasi kepada setiap keputusan yang dihasilkan.

Secara keseluruhannya, kajian kesusasteraan menunjukkan trend yang jelas ke arah penggunaan model XAI untuk meningkatkan kebolehpercayaan dalam penyelidikan kanser payudara. Walau bagaimanapun, cabaran utama masih wujud kerana XAI belum disepadukan secara meluas dengan teknik pemilihan fitur seperti ARM untuk mengesahkan fitur yang berinteraksi yang sangat penting bagi meningkatkan prestasi model pengelasan serta penerimaan pihak klinikal.

2.5.3 Rumusan Isu Model XAI

Penggunaan XAI semakin menjadi keperluan utama dalam sistem pembelajaran mesin yang digunakan dalam bidang klinikal khususnya dalam diagnosis dan rawatan penyakit seperti kanser payudara. Kajian-kajian terdahulu jelas menunjukkan bahawa XAI telah berkembang pesat dalam pelbagai aplikasi kesihatan termasuk diagnosis dan prognosis kanser payudara. Walaupun teknik seperti SHAP, LME, *Permutation Importance* dan Grad-CAM mampu meningkatkan kefahaman serta keyakinan klinikal terhadap model AI, kebanyakannya masih terhad kepada aspek interpretasi selepas keputusan dihasilkan.

Jurang utama yang dikenal pasti ialah ketiadaan integrasi mendalam antara XAI dan proses pemilihan fitur atau ARM. Integrasi ini berpotensi bukan sahaja memperjelas keputusan model, tetapi juga menambah baik pemilihan fitur penting dan

peraturan diagnostik yang lebih bermakna. Justeru itu, penyelidikan lanjut diperlukan untuk meneroka bagaimana XAI boleh digabungkan dengan ARM dan pemilihan fitur bagi menghasilkan model yang lebih telus, boleh dipercayai dan diterima pakai dalam amalan klinikal sebenar. Dalam kajian ini, SHAP dipilih sebagai teknik XAI kerana ia menyediakan penjelasan yang lebih menyeluruh serta mampu menilai interaksi antara fitur melalui nilai interaksi *Shapley*. Berbeza dengan LIME yang hanya memberikan penjelasan secara lokal bagi satu sampel pada satu masa (Ribeiro et al. 2016) dan *Permutation Importance* yang boleh menghasilkan keputusan tidak stabil apabila wujud kebergantungan tinggi antara fitur (Hooker et al. 2021), SHAP menyediakan interpretasi yang konsisten di peringkat lokal dan global.

Tambahan pula, asas teori permainan (*game theory*) yang digunakan dalam nilai *Shapley* memastikan pembahagian sumbangan fitur yang lebih adil, stabil dan boleh dijelaskan, yakni fitur yang penting dalam sistem pengelasan perubatan. Oleh itu, SHAP dipilih untuk memastikan proses pengesahan fitur berinteraksi dalam kajian ini dapat dijalankan dengan lebih telus dan berasaskan penjelasan kuantitatif yang kukuh.

Walau bagaimanapun, terdapat juga kekangan tertentu. Kaedah SHAP boleh menghasilkan interpretasi yang tidak stabil apabila digunakan pada subset data yang berbeza dan hasilnya boleh berubah-ubah bergantung kepada model yang digunakan. Nilai SHAP bagi kes fitur yang berinteraksi boleh berubah bergantung kepada kehadiran atau ketiadaan fitur lain menjadikan proses interpretasi lebih mencabar. Ini menunjukkan keperluan terhadap kerangka pengesahan tambahan bagi memastikan kebolehpercayaan keputusan terutamanya dalam konteks klinikal (Sobhan & Mondal 2023). Justeru, kajian ini menggunakan kaedah majoriti berwajaran berdasarkan beberapa model pengelasan bagi memastikan kestabilan sesuatu keputusan.

2.6 PENGOPTIMUMAN HIPERPARAMETER

Pengoptimuman hiperparameter adalah proses mengoptimumkan hiperparameter yang ditetapkan sebelum proses latihan model pembelajaran mesin. Proses ini penting untuk memastikan model dapat mencapai prestasi terbaik. Berbeza dengan parameter model yang diperoleh dalam proses latihan, hiperparameter merupakan parameter yang tidak

boleh diubah secara automatik oleh model dan perlu ditetapkan terlebih dahulu sebelum proses latihan bermula. Ia menentukan struktur serta strategi pembelajaran model termasuk bagaimana algoritma mempelajari data dan menghasilkan keputusan.

Sebagai contoh, model DT mempunyai hiperparameter seperti kedalaman pokok. Sekiranya nilai kedalaman pokok ditetapkan terlalu tinggi, model berisiko mengalami "terlalu padan" di mana model terlalu sesuai dengan data latihan tetapi gagal berfungsi dengan baik pada data baharu. Sebaliknya, kedalaman yang terlalu rendah menyebabkan model tidak mampu mengenal pasti pola yang kompleks dalam data dan mengakibatkan "terkurang padan" (Elgeldawi et al., 2021).

2.6.1 Pendekatan Pengoptimuman Hiperparameter

Terdapat beberapa kaedah untuk menetapkan hiperparameter bagi set data tertentu. Salah satunya adalah dengan menetapkan secara manual dan mengira ketepatan model bagi setiap percubaan. Selepas itu, nilai lain boleh diuji dan disesuaikan secara berulang melalui proses cuba-dan-ralat. Walaupun kaedah ini boleh menghasilkan keputusan yang memuaskan, ia adalah proses yang memakan masa dan memerlukan usaha yang besar terutamanya apabila melibatkan model yang kompleks dengan ruang parameter yang luas (Elgeldawi et al., 2021). Kaedah alternatif adalah menggunakan nilai lalai iaitu *default* yang disyorkan oleh perisian ML seperti *scikit-learn* atau *TensorFlow* (Simon et al., 2023). Nilai ini biasanya disarankan berdasarkan kajian terdahulu serta pengalaman pakar dalam bidang tersebut. Kadangkala, nilai lalai ini berfungsi dengan baik untuk sesetengah set data tetapi ini tidak bermakna ia boleh memberikan ketepatan terbaik untuk semua bidang (Bischl et al., 2023).

Pemilihan kaedah penalaan hiperparameter yang sesuai bergantung kepada saiz ruang hiperparameter, sumber yang tersedia dan keperluan penggunaan. Antara kaedah tradisional yang sering digunakan bagi penalaan hiperparameter termasuk *grid search* (GS) dan *random search* (RS). GS melibatkan ujian menyeluruh ke atas semua kombinasi parameter dalam ruang pencarian. Dalam pencarian ini, grid nilai hiperparameter ditetapkan, model dilatih dan skor model dikira berdasarkan data ujian untuk setiap kombinasi parameter. Walaupun GS mampu menjamin penemuan kombinasi hiperparameter terbaik, ia boleh menjadi sangat tidak efisien dari segi masa

dan kapasiti pemprosesan (Bischl et al., 2023). Sebagai alternatif, RS lebih efisien dengan memilih kombinasi parameter secara rawak (Bergstra & Y. Bengio, 2012). Teknik ini tidak meneroka semua kemungkinan kombinasi tetapi memilih sejumlah sampel rawak daripada ruang parameter berdasarkan pengedaran kebarangkalian yang telah ditetapkan. Walau bagaimanapun, kelemahan utama RS ialah ia tidak menggunakan maklumat daripada percubaan sebelumnya untuk memilih set hiperparameter yang lebih baik bagi percubaan seterusnya dan ia tidak mempunyai strategi untuk meramalkan hasil percubaan berikutnya (Bischl et al., 2023).

Selain kaedah tradisional, pengoptimuman Bayesian (BO) merupakan kaedah yang lebih cekap dan sistematik dalam penalaan hiperparameter. Ia menggunakan model probabilistik seperti proses Gaussian untuk memberi fokus kepada kawasan hiperparameter yang berpotensi memberikan hasil terbaik sekali gus mengurangkan bilangan percubaan dan kos penilaian dalam model yang kompleks (Nguyen, 2019; Snoek et al., 2012). Tidak seperti GS yang memilih kombinasi secara rawak, BO mempelajari hasil daripada percubaan sebelumnya untuk menentukan percubaan seterusnya. Proses ini bermula dengan beberapa kombinasi hiperparameter dipilih secara rawak untuk membina model awal fungsi objektif. Berdasarkan model ini, algoritma boleh memilih sama ada untuk meneroka kawasan berhampiran kombinasi terbaik yang diperoleh atau mencari kawasan lain yang berpotensi. Kaedah ini lebih cekap berbanding GS yang memerlukan semua kombinasi diuji dan lebih strategik daripada RS yang tidak menggunakan maklumat daripada percubaan terdahulu (Bischl et al., 2023).

Penalaan hiperparameter menggunakan kaedah tradisional melibatkan penerokaan subset parameter tertentu atau julat parameter yang telah ditetapkan. Dalam ruang parameter yang terhad, kaedah ini boleh menjadi cekap untuk mencari penyelesaian global. Namun, kaedah tradisional sering mengalami kesukaran kerana kekangan dalam kriteria pencarian apabila melibatkan pelbagai objektif (Ghaemi et al., 2022). Kelemahan ini dapat diatasi melalui algoritma metaheuristik yang menyelesaikan masalah yang mempunyai banyak pemboleh ubah dan keupayaannya untuk menyelesaikan kedua-dua masalah linear dan tidak linear (Mumtahina et al., 2024). Ia juga boleh digunakan untuk pengoptimuman tunggal dan berbilang objektif

(Silveira et al., 2021). Algoritma metaheuristik juga mempunyai keupayaan untuk menangani ruang pencarian yang sangat besar termasuk ruang parameter yang kompleks secara struktur melalui penggunaan operator seperti mutasi dan penyeberangan. Prestasi algoritma ini dianggap berada di tengah-tengah antara BO yang kompleks namun cekap dari segi sampel dan RS yang mudah tetapi kurang cekap dalam pengurusan sumber (Bischl et al., 2023).

2.6.2 Pengoptimuman Hiperparameter dalam Perlombongan Peraturan Kesatuan

Kaedah ARM bergantung kepada tetapan hiperparameter seperti *minSupp* dan *minConf* untuk mengenal pasti pola penting dalam data. Pemilihan parameter yang tidak sesuai boleh menghasilkan sama ada terlalu banyak peraturan yang tidak berguna atau kegagalan mengenal pasti pola yang signifikan. Justeru itu, penyelidik telah mencuba pelbagai pendekatan bagi mengoptimumkan hiperparameter ini khususnya melalui penggunaan algoritma metaheuristik.

Beberapa kajian menunjukkan keberkesanan gabungan pendekatan heuristik dan metaheuristik dalam meningkatkan kualiti peraturan. Sebagai contoh, Kliegr & Kuchař (2019) telah menggabungkan kaedah heuristik dan SA untuk pengoptimuman *minSupp* dan *minConf*, manakala Indira & Kanmani (2015) menerapkan adaptif PSO untuk pengoptimuman pelbagai hiperparameter ARM. Walaupun kaedah ini meningkatkan kecekapan perlombongan, ia berdepan dengan isu kos pengiraan yang tinggi bagi set data berskala besar.

Selain itu, Triantali et al. (2022) menggunakan PSO berwajaran untuk rantaian bekalan, manakala Su et al. (2019) menggabungkan PSO bersama algoritma *FP-growth* dalam analisis keselamatan sosial. Kajian lain memperkenalkan algoritma hibrid baru seperti Pengoptimuman Dingo (DO) dan *Billiards-Inspired Optimization* (BIO) (Kishor & Porika 2023) atau tabu search (TS) dengan algoritma *Apriori* (Beausoleil 2020). Walaupun kaedah ini berjaya meningkatkan kualiti peraturan, namun kebanyakannya hanya diuji pada set data piawai UCI dan masih terhad penggunaannya pada set data dunia sebenar berskala besar. Kajian berasaskan GA (Kanani & Mishra, 2015) dan ACO (Olmo et al., 2013) menekankan kelebihan dalam memilih peraturan yang lebih relevan.

Dalam sektor industri, Zhou et al. (2019) menggunakan pengoptimuman ARM untuk meningkatkan prestasi sistem pengurusan tenaga di bangunan komersial khususnya pada sistem HVAC (*Heating, Ventilation, and Air Conditioning*) manakala Tunali & Bayrak (2021) menggunakan ARM masa nyata dalam sektor makanan segera. Kajian Connolly et al. (2024) pula menunjukkan pengagregatan masa dan pengelompokan kategori produk berkesan mengatasi masalah data transaksi yang jarang.

Isu dimensi data yang kompleks turut ditangani oleh Siswanto et al. (2024) dengan memperkenalkan kaedah *Set Difference FP-Growth* (SDFP-growth) bagi mengurangkan dimensi data sebanyak 69% tanpa menjejaskan kualiti peraturan. Kajian lain pula mengaplikasikan ARM bersama GIS untuk keselamatan trafik (Jiang et al. 2020) manakala Ogedengbe et al. (2024) dan Wang et al. (2019) menekankan keperluan pengoptimuman dinamik dalam data yang sering berubah melalui algoritma adaptif dan *Incremental Fuzzy* ARM.

Secara keseluruhannya, kajian kesusasteraan menunjukkan trend yang jelas ke arah penggunaan algoritma metaheuristik untuk mengoptimumkan parameter ARM dalam pelbagai domain. Walau bagaimanapun, cabaran utama masih kekal pada isu penalaan hiperparameter ARM secara automatik bagi proses pemilihan dan pengesahan fitur berinteraksi. Jadual 2.5 menunjukkan ringkasan kajian kesusasteraan bagi kaedah pengoptimuman hiperparameter ARM.

Jadual 2.5 Ringkasan kajian kesusasteraan bagi kaedah pengoptimuman hiperparameter ARM

Rujukan	Domain	Kaedah ARM	Kaedah Pengoptimuman	Metrik	Hasil Kajian	Limitasi Kajian
Kliegr & Kuchař (2019)	Pelbagai	CBA	Heuristik, SA	Supp, Conf	Prestasi pengelasan meningkat	Hanya metrik asas digunakan
Indira & Kanmani (2015)	Pelbagai	ARM	Adaptif PSO	Supp, Conf, Lift, Conv, Lev, Laplace	Prestasi meningkat dan masa cekap	Hanya membandingkan algoritma asas PSO
Triantali et al. (2022)	Ekonomi	ARM	PSO	Supp, Conf	Prestasi meningkat	Hanya metrik asas digunakan
Su et al. (2019)	Keselamatan sosial	FP-Growth	PSO, Entropi	Supp, Conf	Prestasi meningkat	Hanya metrik asas digunakan
Kishor & Porika (2023)	Data teks	NARMA	DO + BIO	Supp, Conf, Lift, Korelasi	Conf=81%	Domain data terhad
Beausoleil (2020)	Pelbagai	Apriori	TS	Supp, Conf	Ambang diperoleh secara automatik	Hanya metrik asas digunakan
Kanani & K. Mishra (2015)	Data berangka	ARM	GA	Supp, Conf, Comp, Int	Peningkatan kualiti peraturan	Tiada pengesahan statistik
Olmo et al. (2013)	Pelbagai	ARM	ACO	Supp, Conf,	Prestasi lebih baik	Hanya metrik asas digunakan
Zhou et al. (2019)	Perindustrian	ARM	Pembahagian data	Supp, Conf	Pengurangan penggunaan tenaga	Set data terhad
Tunali & Bayrak (2021)	Makanan segera	ARM	Kemaskini masa nyata	Supp, Conf, Lift	Peningkatan skalabiliti	Masa pengiraan tinggi
Mohapatra et al. (2021)	Peruncitan	Apriori	ECLAT	Supp, Conf, Lift	Pengenalpastian peraturan penting	Set data terhad
Connolly et al. (2024)	E-dagang	FP-Growth	Pengagregatan masa dan kategori	Supp	Dimensi data berkurang, masa cekap	Keperluan pra-pemprosesan lanjutan
Siswanto et al. (2024)	Pelbagai	SDFP-Growth	Set teori	Supp, Conf, Lift	Dimensi data berkurang	Keperluan menggunakan ML
Li et al. (2024)	Sistem maklumat	Apriori	GA	Supp	Prestasi meningkat dan masa cekap	Hanya metrik asas digunakan
Jiang, Yuen, et al. (2020)	Keselamatan trafik	ARM	GS	Supp, Conf	Prestasi pengelasan lebih baik	Pengesahan dunia sebenar
Ogedengbe et al. (2024)	Pendidikan	ARM	Adaptif minSupp	Supp, Conf, Lift	Peraturan lebih signifikan	Tiada perbandingan kaedah lain
Wang et al. (2019)	Pelbagai	ARM, ECLAT	Set kabur, tetingkap gelongsor	Supp, Conf	Prestasi pengelasan meningkat, masa cekap	Kebergantungan pada pengalaman pengguna

Berdasarkan Jadual 2.5, pelbagai domain telah mengaplikasikan pengoptimuman hiperparameter pada ARM seperti ekonomi, keselamatan sosial dan trafik, perindustrian, peruncitan dan lain-lain. Kaedah ARM digunakan secara meluas dengan gabungan kaedah lain seperti SA, PSO, ACO, GA dan TS bagi mempercepatkan proses penalaan hiperparamater dan meningkatkan prestasi ramalan. Kebanyakan metrik yang dikaji adalah seperti *Supp, Conf* dan *Lift*. Hasil kajian-kajian ini menunjukkan bahawa peningkatan prestasi model, masa pemprosesan yang lebih cepat serta peraturan dijana lebih berkualiti telah tercapai. Walau bagaimanapun, terdapat beberapa batasan kajian seperti hanya metrik asas digunakan serta kebergantungan pada pengalaman pengguna yang memerlukan kajian lebih lanjut. Justeru, jurang ini mengukuhkan objektif kajian ini untuk membangunkan model pemilihan fitur berasaskan interaksi dengan pengoptimuman hiperparameter secara automatik bagi pelbagai metrik bagi mengekstrak peraturan yang lebih kukuh, stabil dan dapat dijelaskan.

2.6.3 Rumusan Isu Pengoptimuman Hiperparameter

Pengoptimuman hiperparameter memainkan peranan penting dalam memastikan model pembelajaran mesin berfungsi dengan lebih cekap dan tepat terutamanya dalam domain penjagaan kesihatan seperti pengelasan kanser dan sistem sokongan keputusan klinikal. Hiperparameter yang dipilih dengan baik bukan sahaja dapat meningkatkan prestasi model tetapi juga membantu dalam mengurangkan masalah "terlalu padan". Namun, cabaran utama ialah ruang pencarian parameter yang besar dan berdimensi tinggi menjadikan proses penalaan hiperparameter sesuatu yang memakan masa dan sumber pengiraan yang tinggi (Bischl et al., 2023).

Kaedah tradisional seperti *Grid Search* (GS) dan *Random Search* (RS) sering digunakan untuk mencari kombinasi hiperparameter terbaik. Namun begitu, kaedah ini kurang cekap kerana ia meneroka pelbagai kombinasi secara tetap atau rawak yang seterusnya melibatkan kos pengiraan yang sangat tinggi dari segi masa dan sumber. Selain itu, kebanyakan kaedah pengoptimuman metaheuristik hanya menumpukan kepada prestasi pengelasan semata-mata. Penggunaan teknik metaheuristik pada ARM bagi pengoptimuman hiperparameter turut digunakan, namun kebanyakannya masih bergantung pada metrik asas seperti *Supp* dan *Conf* sahaja sebagai panduan pencarian

peraturan. Berbeza dengan pendekatan tersebut, kajian ini menggunakan algoritma SA dengan memfokuskan kepada peningkatan prestasi pengelasan melalui fungsi objektif berasaskan AUC serta memastikan bahawa peraturan yang dihasilkan lebih kukuh menggunakan beberapa kriteria tambahan seperti *minSupp, minConf, minLift, minConv, minCF, minLev*. Kajian ini memilih algoritma SA kerana sifatnya yang lebih sesuai untuk ruang carian hiperparameter ARM yang tidak linear dan mudah terperangkap dalam lokal optimum. Tidak seperti GA atau PSO yang memerlukan populasi besar serta pelbagai parameter kawalan seperti saiz populasi, kadar mutasi, kadar pembelajaran dan momentum, algoritma SA meneroka penyelesaian dengan lebih fleksibel melalui proses penyejukan berperingkat yang membenarkan penerimaan penyelesaian kurang baik pada peringkat awal. Ini meningkatkan keupayaannya untuk melarikan diri daripada lokal optimum dengan kos pengiraan yang lebih rendah dan lebih praktikal bagi penalaan hiperparameter CARM.

Di samping itu, proses penalaan hiperparameter sering dilakukan secara berasingan daripada proses pemilihan fitur sedangkan kedua-dua aspek ini saling berkait rapat. Pendekatan berasingan ini bukan sahaja membazir sumber pengiraan tetapi juga mengabaikan potensi hubungan antara fitur dan tetapan nilai hiperparameter. Kajian oleh (Binder et al., 2020) menyokong bahawa penalaan hiperparameter dan pemilihan fitur seharusnya digabungkan dalam satu kerangka pelbagai objektif untuk mengekalkan prestasi model.

2.7 KESIMPULAN

Bab ini membincangkan kepentingan pemilihan fitur, *Association Rule Mining* (ARM), *eXplainable AI* (XAI) dan pengoptimuman hiperparameter dalam pembangunan model pengelasan kanser payudara. Kajian kesusasteraan menunjukkan ARM membantu memilih fitur yang relevan, mengesan interaksi penting dan mengurangkan pertindihan antara fitur bagi meningkatkan ketepatan model pengelasan manakala XAI dapat serta kebolehpercayaan model dan pengoptimuman hiperparameter menambah kecekapan model. Walau bagaimanapun, masih wujud jurang dalam pengenalpastian interaksi fitur yang signifikan secara automatik dan penyepaduan ARM dengan XAI bagi menghasilkan diagnosis yang lebih dipercayai.

BAB III

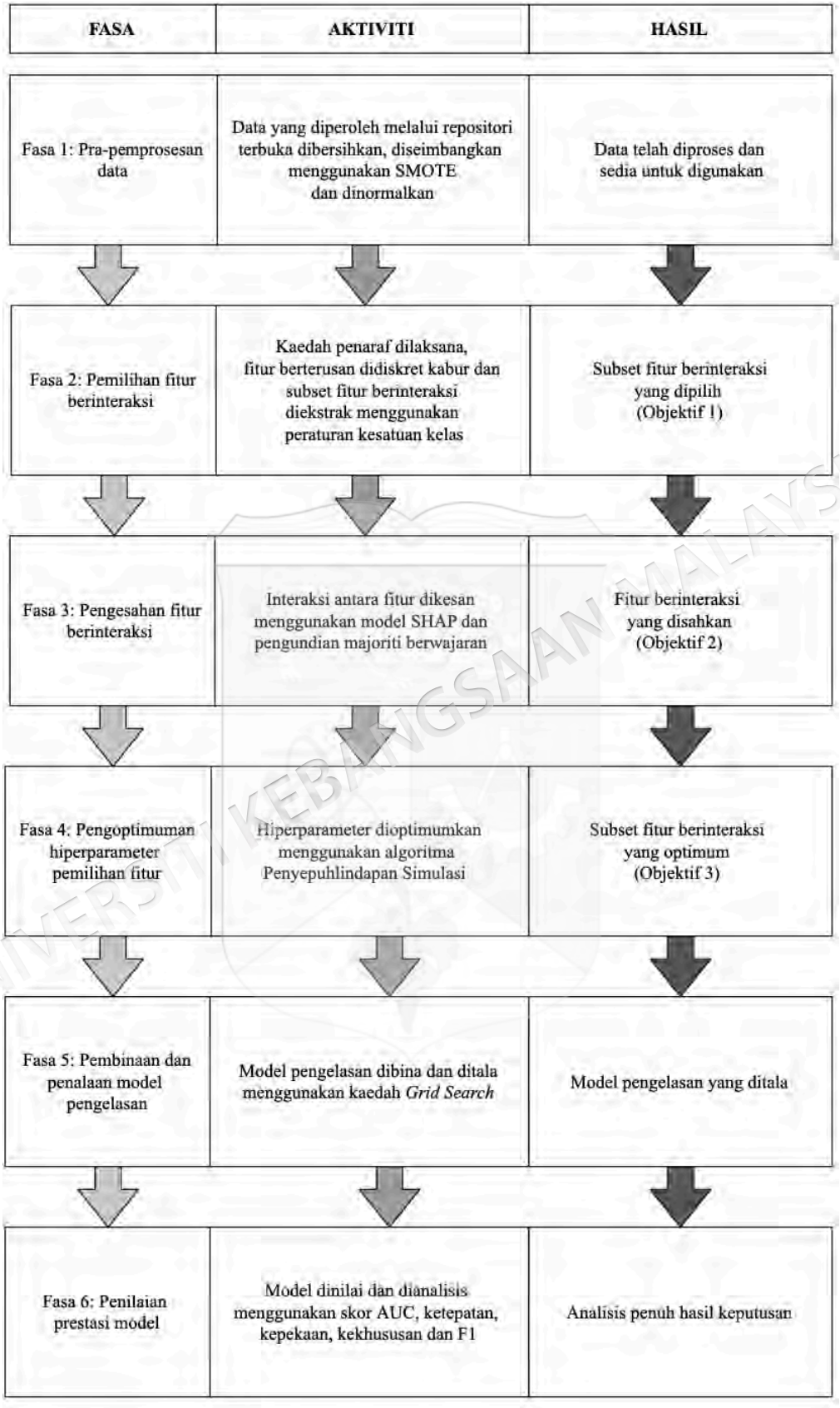
METODOLOGI KAJIAN

3.1 PENGENALAN

Bab ini menerangkan metodologi kajian yang dibangunkan secara menyeluruh untuk mencapai objektif kajian khususnya dalam konteks pengelasan kanser payudara. Kajian ini memperkenalkan satu aliran kerja pemilihan fitur berasaskan interaksi dan boleh dijelaskan yang merangkumi beberapa teknik terpilih yang digabungkan secara strategik bagi meningkatkan keberkesanan pemilihan dan pengesahan fitur. Enam fasa kajian direka bentuk bagi menangani isu-isu utama yang dikenal pasti dalam kenyataan masalah seperti pengabaian interaksi antara fitur, kekurangan kebolehjelasan model dan kesukaran penalaan hiperparameter secara manual.

3.2 REKA BENTUK KAEDAH KAJIAN

Kajian ini berpandukan kepada pendekatan kuantitatif berasaskan eksperimen yang disusun dalam bentuk aliran kerja dengan menumpukan kepada pembangunan, penilaian dan pengesahan pemilihan fitur berdasarkan interaksi bagi pengelasan kanser payudara. Reka bentuk ini melibatkan enam fasa utama seperti yang diringkaskan dalam Rajah 3.1.



Rajah 3.1 Aliran metodologi kajian

Fasa 1 melibatkan pra-pemprosesan data yang melibatkan penyediaan data melalui proses seperti pembersihan, penormalan dan penyeimbangan. Pembersihan data melibatkan penghapusan nilai *NaN* (*Not a Number*) iaitu entri yang tidak mempunyai sebarang nilai bagi memastikan tiada kehilangan maklumat yang boleh mempengaruhi analisis. Bagi menangani ketidakseimbangan data, teknik *Synthetic Minority Over-sampling Techniques* (SMOTE) digunakan bagi memastikan taburan kelas adalah seimbang. Proses ini penting bagi memastikan data bersedia untuk digunakan dalam eksperimen seterusnya. Penormalan data pula dilakukan menggunakan kaedah Penskalaan Piawai (*Standard Scaler*) bagi memastikan data berada dalam skala yang seragam dan mengelakkan kesan dominasi oleh fitur dengan nilai lebih besar.

Fasa 2 melibatkan pemilihan fitur berasaskan interaksi yang akan digunakan pada fasa pengelasan. Proses ini dimulakan dengan pengenalpastian calon subset fitur menggunakan tiga kaedah penaraf iaitu (i) *Pearson Correlation Coefficient* (PCC), (ii) *Mutual Information* (MI) dan (iii) *Feature Importance* (FI). Ambang Sisihan Piawai (SD) digunakan pada setiap kaedah tersebut bagi menyingkirkan fitur yang tidak menyumbang kepada keberkesanan model. Seterusnya, proses pendiskretan dilakukan menggunakan model Logik Kabur bagi mengubah fitur berterusan kepada fitur berkategori yang lebih sesuai untuk analisis peraturan. Setelah itu, pengekstrakan fitur yang relevan dan interaksi dilakukan menggunakan CAR manakala fitur bertindih disingkir menggunakan AAR. Hasil daripada fasa ini ialah subset fitur yang relevan, berinteraksi dan tidak bertindih.

Fasa 3 mengesahkan interaksi antara fitur yang telah dipilih menggunakan model XAI. Kaedah ini memastikan model pemilihan fitur yang dicadangkan benar-benar mempertimbangkan fitur berinteraksi. Proses ini bermula dengan pengenalpastian fitur yang kurang relevan menggunakan nilai SHAP dan pengundian majoriti berwajaran. Nilai SHAP dihasilkan daripada model RF, GB dan XGB manakala pemberat bagi pengundian majoriti berwajaran ialah skor AUC bagi setiap model. Fitur-fitur tersebut kemudian disusun secara menurun berdasarkan nilai SHAP dan fitur kurang relevan diekstrak menggunakan ambang kuantil ke-75. Setelah itu, interaksi antara fitur yang kurang relevan tersebut dianalisis menggunakan nilai interaksi SHAP. Hasil daripada fasa ini ialah fitur yang kurang relevan tetapi berinteraksi yang akan

disahkan dengan subset fitur yang diekstrak sebelum penyingkiran fitur bertindih yang diperoleh daripada fasa 2.

Fasa 4 melibatkan prestasi setiap model yang ditingkatkan melalui penalaan hiperparameter menggunakan algoritma SA. Dalam konteks kajian ini, algoritma SA digunakan untuk menambah baik proses penalaan hiperparameter bagi kaedah CARM. Hiperparameter yang ditala merangkumi nilai ambang seperti sokongan minimum (*minSupp*), keyakinan minimum (*mincConf*), angkat minimum (*minLift*), sabitan minimum (*minConv*),), faktor kepastian minimum (*minCF*) dan keumpilan minimum (*minLev*). Setelah kombinasi hiperparameter tersebut diperoleh, CAR dan AAR akan diekstrak bagi mengenal pasti subset fitur. Prestasi subset fitur yang diperoleh akan diuji berdasarkan purata skor AUC bagi lima model pengelasan seperti DT, RF, NB, LR dan SVM. Proses ini dijalankan sebanyak 20 lelaran untuk mencari purata skor AUC terbaik. Hasil akhir daripada fasa ini ialah subset fitur optimum yang menghasilkan prestasi skor AUC tertinggi.

Fasa 5 melibatkan pembinaan model pengelasan menggunakan subset fitur yang telah dipilih dan disahkan dalam fasa sebelumnya. Lima model pembelajaran mesin yang berbeza telah digunakan iaitu DT, RF, NB, LR dan SVM bagi menilai keupayaan generalisasi subset fitur yang diperoleh melalui kaedah FD-CARI. Setiap model dibina menggunakan data latihan dan diuji menggunakan data pengujian yang telah dibahagikan berdasarkan nisbah 2:1 berpandukan kajian oleh Wang et al. (2021). Kaedah *Grid Search* (GS) dengan pengesahan silang sebanyak 5 kali digunakan untuk penalaan hiperparameter setiap model pengelasan. Hasil daripada fasa ini ialah pengelasan kanser berdasarkan setiap set data.

Fasa 6 melibatkan penilaian dan perbandingan prestasi model pengelasan yang dibina menggunakan subset fitur daripada kaedah pemilihan fitur yang dicadangkan dengan kaedah pemilihan fitur piawai. Penilaian ini merangkumi tiga aspek utama iaitu:

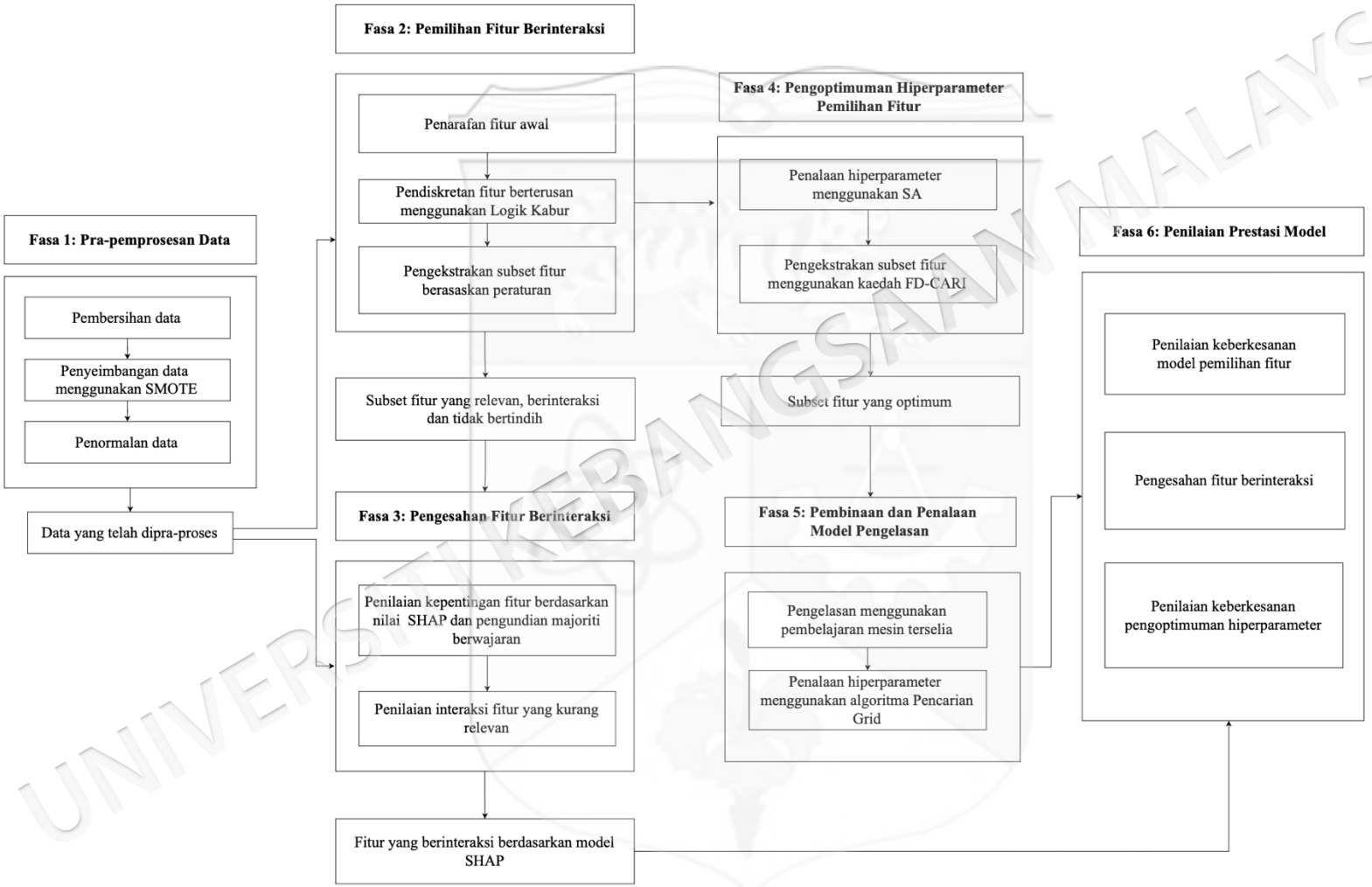
a. Perbandingan kaedah pemilihan fitur iaitu prestasi kaedah FD-CARI dibandingkan dengan lima kaedah pemilihan fitur piawai seperti CFS, FCBF,

Consistency, Relief-F dan mRMR. Penilaian prestasi ini menggunakan pelbagai metrik iaitu skor AUC, ACC, SEN, SPE dan F1.

b. Penilaian kerangka pengesahan fitur berinteraksi di mana keberkesanan kerangka pengesahan fitur berinteraksi (menggunakan SHAP dan pengundian majoriti berwajaran) dilaksanakan melalui perbandingan antara fitur yang diperoleh daripada fasa 3 dengan subset fitur yang diekstrak sebelum penyingkiran fitur bertindih yang diperoleh daripada fasa 2. Maklumat terperinci mengenai fasa ini dibentangkan dalam Seksyen 3.10.

c. Perbandingan kaedah pemilihan fitur iaitu prestasi algoritma FD-CARI+SA dibandingkan dengan lima teknik pemilihan fitur piawai tersebut. Penilaian prestasi ini menggunakan metrik AUC.

Rajah 3.2 menunjukkan reka bentuk kaedah kajian secara keseluruhan.



Rajah 3.2 Reka bentuk kaedah kajian

3.3 PERISIAN DAN PLATFORM

Pelaksanaan keseluruhan kajian ini menggunakan pelbagai perisian dan platform analitik yang menyokong pembangunan algoritma, pemprosesan data, latihan model pembelajaran mesin serta analisis statistik. Senarai perisian dan platform yang digunakan adalah seperti berikut:

a. *Python* (versi 3.11.11) digunakan sebagai bahasa pengaturcaraan utama untuk menjalankan pendiskretan kabur, membangunkan algoritma FD-CARI, penalaan hiperparameter menggunakan SA serta pengujian model pengelasan.

b. Platform *Spyder* versi 5.4.3 digunakan sebagai Persekitaran Pembangunan Bersepadu (IDE) untuk memudahkan penulisan dan pengujian kod secara interaktif. Perisian ini dipilih kerana fleksibilitinya dalam pengendalian data berskala besar serta sokongan luas terhadap pustaka seperti:

   - *NumPy*, *Pandas* untuk analisis data.
   - *StandardScaler* untuk penormalan.
   - *SMOTE* untuk menangani ketidakseimbangan data.
   - *scikit-learn* untuk membantu dalam pemprosesan data.
   - *mlxtend* untuk mengekstrak dan menganalisis peraturan kesatuan.
   - *SHAP* untuk menilai kepentingan fitur dan mengenal pasti interaksi fitur.
   - *matplotlib* yang digunakan sebagai asas bagi pembangunan pustaka pengvisualan lain seperti *seaborn* dan *ggplot* untuk menghasilkan grafik yang lebih informatif.

c. Perisian MATLAB R2024b digunakan bagi pendiskretan fitur berterusan menggunakan model Logik Kabur sebelum digunakan pada kaedah CARM. MATLAB dipilih kerana ia menyediakan fungsi terbina dalaman yang memudahkan analisis. Selain itu, fungsi terbina dalam bagi teknik FS piawai seperti mRMR dan Relief-F dapat dijalankan dengan lebih cekap.

d. Perisian WEKA versi 3.8.6 pula digunakan bagi teknik piawai lain seperti CFS, FCBF dan Consistency. WEKA merupakan perisian sumber terbuka yang

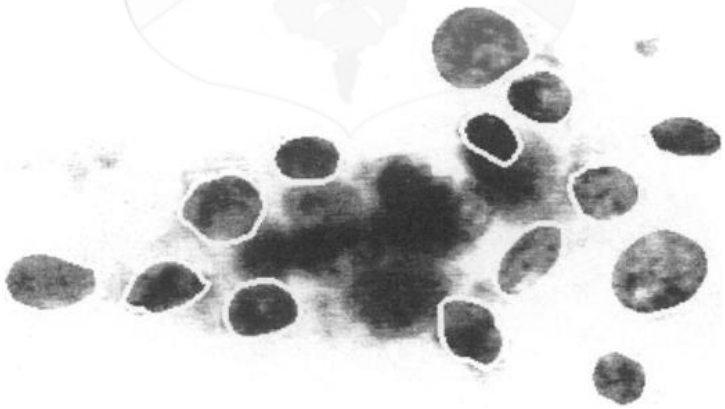
menyokong pelbagai kaedah pemilihan fitur dan dipilih kerana kemudahannya dalam menjalankan eksperimen.

e. Perisian *Microsoft Excel* versi 16.94 tahun 2025 digunakan bagi penyediaan dan analisis data mentah kerana ia menyediakan pelbagai kemudahan dalam pengiraan, unjuran, analisis statistik serta pengurusan data secara sistematik. Perisian ini juga digunakan bagi melakukan analisis data sebelum dan selepas diproses.

3.4 SET DATA KAJIAN

Kajian ini menggunakan tiga set data kanser payudara yang diperoleh daripada sumber terbuka iaitu Kanser Payudara Diagnostik Wisconsin (WDBC) (Wolberg, Mangasarian, et al., 1995), Kanser Payudara Prognostik Wisconsin (WPBC) (Wolberg, Street, et al., 1995) dan Kanser Payudara Coimbra (BCC) (Patrcio et al., 2018). Setiap set data ini mewakili variasi dari segi bilangan fitur, jenis fitur seperti klinikal, biometrik dan morfologi serta beberapa bentuk pengelasan seperti diagnosis, prognosis dan risiko.

Fitur-fitur pada set data WDBC dan WPBC diekstrak daripada imej Biopsi Aspirasi Jarum Halus Digital (FNA) seperti ditunjukkan dalam Rajah 3.3 manakala set data BCC merangkumi fitur-fitur yang diperoleh daripada analisis darah.



Rajah 3.3 Contoh imej FNA (Mangasarian et al., 2008)

3.4.1 Kanser Payudara Diagnostik Wisconsin (WDBC)

Set data WDBC terdiri daripada 569 sampel dengan 30 fitur merangkumi ukuran morfologi nukleus sel yang diekstrak daripada FNA. Terdapat 10 jenis fitur asas dalam set data ini termasuk jejari, tekstur, perimeter, luas, kelicinan, kekompakan, kecekungan, titik cekung, simetri dan dimensi fraktal. Setiap 10 fitur ini mempunyai tiga variasi berdasarkan nilai purata, ralat piawai (SE) dan maksimum yang menjadikan jumlah keseluruhan fitur ialah 30 fitur. Jadual 3.1 dan 3.2 menunjukkan senarai fitur serta julat bagi setiap fitur dan kelas sasaran masing-masing bagi set data WDBC.

Jadual 3.1 Senarai fitur bagi set data WDBC

Nama Fitur	Penerangan Fitur	Julat Purata (m*ean*, μ)	Julat Ralat Piawai (SE)	Julat Maximum (*worst*)
Jejari (*radius*)	Purata jarak dari pusat nukleus sel ke perimeter nukleus	6.98 – 28.11	0.11 – 2.87	7.93 – 36.04
Tekstur (*texture*)	Variasi intensiti aras kelabu dalam nukleus	9.71 – 39.28	0.36 – 4.89	12.02 – 49.54
Perimeter (*perimeter*)	Jumlah panjang keseluruhan sempadan nukleus	43.79 – 188.50	0.76 – 21.98	50.41 – 251.20
Luas (*area*)	Keluasan kawasan nukleus	143.50 – 2501	6.80 – 542.20	185.20 – 4254.00
Kelicinan (*smoothness*)	Perbezaan antara panjang jejari individu dan purata semua jejari di sekeliling nukleus	0.05 – 0.16	0.00 – 0.03	0.07 – 0.22
Kekompakan (*compactness*)	Ukuran kebundaran nukleus	0.02 – 0.35	0.00 – 0.14	0.02 – 1.06
Kecekungan (*concavity*)	Tahap kedalaman cekungan pada sempadan nukleus	0.00 – 0.43	0.00 – 0.40	0.00 – 1.25
Titik cekung (*concave points*)	Bilangan cekungan pada perimeter nukleus	0.00 – 0.20	0.00 – 0.05	0.00 – 0.29
Simetri (*symmetry*)	Tahap kesimetrian bentuk nukleus	0.11 – 0.30	0.01 – 0.08	0.16 – 0.66
Dimensi fraktal (*fractal dimension)*	Kerumitan geometri sempadan nukleus	0.05 – 0.10	0.00 – 0.03	0.06 – 0.21

Jadual 3.2 Kelas sasaran bagi set data WDBC

Diagnosis	Jumlah sampel	Kelas
Benigna (*benign*)	357	B
Malignan (*malignant*)	212	M

3.4.2 Kanser Payudara Prognostik Wisconsin (WPBC)

Set data WPBC terdiri daripada 198 sampel dengan 33 fitur. Jumlah 30 fitur dalam set data ini mengandungi maklumat ciri sel daripada 10 fitur dengan tiga variasi berdasarkan nilai purata (*1*), ralat piawai (SE) (*2*) dan maksimum (*3*) setiap fitur dengan tambahan 3 fitur iaitu masa tindak balas, saiz tumor dan status nod limfa. Jadual 3.3 dan 3.4 menunjukkan senarai fitur serta julat bagi setiap fitur dan kelas sasaran masing-masing bagi set data WPBC.

Jadual 3.3 Senarai fitur bagi set data WPBC

Nama Fitur	**Penerangan Fitur**	**Julat Purata (m*ean*, μ)**	**Julat Ralat Piawai (SE)**	**Julat Maximum (*worst*)**
Jejari (*radius*)	Purata jarak dari pusat nukleus sel ke perimeter nukleus	10.95 – 27.22	0.19 – 1.82	12.84 – 35.13
Tekstur (*texture*)	Variasi intensiti aras kelabu dalam nukleus	10.38 – 39.28	0.36 – 3.50	16.67 – 49.54
Perimeter (*perimeter*)	Jumlah panjang keseluruhan sempadan nukleus	71.90 – 182.10	1.15 – 13.28	85.10 – 232.20
Luas (*area*)	Keluasan kawasan nukleus	361.60 – 2250	13.99 – 316.00	508.10 - 3903
Kelicinan (*smoothness*)	Perbezaan antara panjang jejari individu dan purata semua jejari di sekeliling nukleus	0.07 – 0.15	0.00 – 0.04	0.08 – 0.23
Kekompakan (*compactness*)	Ukuran kebundaran nukleus	0.04 – 0.32	0.00 – 0.14	0.05 –1.06
Kecekungan (*concavity*)	Tahap kedalaman cekungan pada sempadan nukleus	0.01 – 0.43	0.01 – 0.15	0.02 – 1.17
Titik cekung (*concave points*)	Bilangan cekungan pada perimeter nukleus	0.02 – 0.21	0.01 – 0.04	0.03 – 0.30
Simetri (*symmetry*)	Tahap kesimetrian bentuk nukleus	0.13 – 0.31	0.00 – 0.07	0.15 – 0.67
Dimensi fraktal (*fractal dimension)*	Kerumitan geometri sempadan nukleus	0.05 – 0.10	0.00 – 0.02	0.05 – 0.21
Masa (*Time*)	Tempoh tindakan lanjut pesakit selepas pembedahan (bulan)		1- 125	
Saiz tumor (*tumor_size*)	Diameter tumor yang dikeluarkan (mm)		0.40 – 10.00	
Status Nod Limfa (*lymph_node_status*)	Bilangan nod limfa aksilari yang dikesan semasa pembedahan		0 – 27	

Jadual 3.4 Kelas sasaran bagi set data WPBC

Keputusan	Jumlah Sampel	Kelas
Berulang (recurrence)	47	R
Tidak berulang (non-recurrence)	151	N

3.4.3 Kanser Payudara Coimbra (BCC)

Set data BCC mengandungi 116 sampel dengan 9 fitur yang terdiri daripada data antropometrik dan parameter yang diperoleh daripada sampel darah. Fitur-fitur dalam set data ini termasuk umur, indeks jisim badan (BMI), glukosa, Homeostasis Model Assessment (HOMA), Insulin dan jenis hormon seperti Leptin, Adiponektin, Resistin serta *Monocyte Chemoattractant Protein-1* (MCP.1). Jadual 3.5 dan 3.6 menunjukkan senarai fitur serta julat bagi setiap fitur dan kelas sasaran masing-masing bagi set data BCC.

Jadual 3.5 Senarai fitur bagi set data BCC

Nama Fitur	Penerangan Fitur	Julat Nilai
Umur (*Age*)	Umur pesakit (tahun)	24 – 89
BMI (*BMI)*	Indeks jisim badan (kg/m^2)	18.37 – 38.58
Glukosa (*Glucose*)	Paras glukosa (mg/dL)	60.00 – 201.00
Insulin (*Insulin*)	Paras insulin (μU/mL)	2.43 – 58.46
HOMA (*HOMA*)	Indeks rintangan insulin	0.47 – 25.06
Leptin	Hormon berkaitan dengan metabolisma lemak (ng/mL)	4.31 – 90.28
Adiponectin	Hormon berkaitan sensitiviti insulin (μU/mL)	1.65 – 38.04
Resistin	Hormon berkaitan keradangan dan rintangan insulin (ng/mL)	3.21 – 82.10
MCP.1	Protein penanda keradangan	45.84 – 1698.44

Jadual 3.6 Kelas sasaran bagi set data BCC

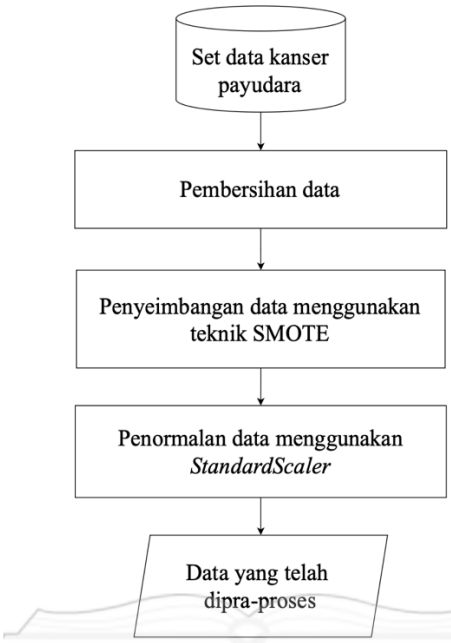
Pengelasan	Jumlah sampel	Kelas
Kawalan sihat (H*ealthy controls*)	52	1
Pesakit (*Patients*)	64	2

Ketiga-tiga set data tersebut telah digunakan secara meluas dalam kajian terdahulu berkaitan analisis kanser payudara seperti dalam kajian Elkorany et al. (2022), Ayoola & Ogunfunmi (2022), Murugesan et al. (2021) dan Dalwinder et al. (2020). Selain itu, set data ini juga bersifat pelbagai dimensi dan mencerminkan ciri-ciri biologi yang relevan dalam pengelasan tumor.

Ketiga-tiga set data ini juga dipilih kerana (i) akses kepada set data ini adalah terbuka dan mempunyai ketersediaan secara umum dan membolehkan pengesahan kajian oleh penyelidik lain, (ii) kepelbagaian fitur dan kelas sasaran merangkumi aspek diagnosis seperti ramalan pengulangan dan penilaian risiko kanser dan (iii) set data ini memberi peluang untuk menguji keupayaan kerangka cadangan dalam pelbagai senario pengelasan.

3.5 FASA 1: PRA-PEMPROSESAN DATA

Pra-pemprosesan data merupakan langkah awal yang penting bagi memastikan kualiti data adalah sesuai untuk analisis lanjut. Proses ini bertujuan untuk mengatasi isu seperti ketidakseimbangan kelas, nilai hilang dan variasi skala yang boleh menjejaskan keberkesanan pemilihan fitur dan model pengelasan. Ketiga-tiga set data menggunakan prosedur pra-pemprosesan yang sama bagi memastikan konsistensi dan mengelakkan ketidakadilan dalam analisis, memandangkan kebanyakan algoritma pembelajaran mesin sensitif terhadap variasi skala, taburan dan julat nilai fitur (Ahsan et al. 2021; Theng & Bhoyar 2024). Pendekatan seragam ini juga membolehkan perbandingan prestasi model dilakukan secara adil merentasi set data yang berbeza. Di samping itu, penggunaan prosedur pra-pemprosesan yang konsisten adalah selaras dengan kajian-kajian terdahulu yang telah membuktikan keberkesanan kaedah ini apabila diaplikasikan pada ketiga-tiga set data tersebut (Alsabry et al. 2023; Patial et al. 2025; S. Wang et al. 2021). Prosedur pra-pemprosesan yang dilaksanakan dalam kajian ini merangkumi beberapa langkah berikut seperti yang ditunjukkan dalam Rajah 3.4.



Rajah 3.4 Proses pra-pemprosesan data

3.5.1 Pembersihan Data

Setiap set data diperiksa bagi mengenal pasti fitur yang tidak memberikan maklumat bermakna dan nilai hilang yang boleh menjejaskan analisis statistik dan latihan model. Bagi set data WPBC, hanya rekod lengkap digunakan kerana terdapat 4 sampel yang mempunyai nilai hilang dalam fitur *lymph_node_status.* Pendekatan ini dipilih kerana jumlah sampel yang terjejas adalah kecil dan penyingkirannya tidak memberi kesan besar terhadap keseimbangan data (Hasan & Shafi, 2023; Mohammed et al., 2020; Zeid et al., 2022). Manakala bagi set data WDBC dan BCC, tiada fitur hilang dikesan. Proses pembersihan ini digunakan bagi mengekalkan integriti data.

3.5.2 Penyeimbangan Data

Ketidakseimbangan antara bilangan kelas (seperti *malignant* vs *benign* atau *recurrence* vs *non-recurrence*) berpotensi menyebabkan model pembelajaran mesin lebih cenderung terhadap kelas majoriti dan mengabaikan kelas minoriti. Untuk mengatasi isu ini, teknik *Synthetic Minority Over-sampling Techniques* (SMOTE) digunakan pada set data yang tidak seimbang (Chawla et al., 2002). SMOTE menjana sampel sintetik baharu, x_{baru} menggunakan formula seperti berikut:

$$x_{baru} = x + \delta(x_{jiran} - x) \quad \ldots(3.1)$$

di mana δ ialah nombor rawak antara 0 dan 1. SMOTE dilakukan berdasarkan kelas minoriti tanpa meniru rekod asal dan membolehkan pengagihan kelas yang lebih seimbang. Proses ini hanya diaplikasikan ke atas set latihan bagi mengelakkan kebocoran maklumat (*data leakage*) kepada set pengujian (Demircioğlu, 2024; Kabane, 2024; Kapoor & Narayanan, 2023; Rosenblatt et al., 2024).

Sebagai contoh, taburan kelas dalam set data WPBC adalah tidak seimbang sebelum SMOTE digunakan di mana kelas minoriti (R) yang hanya mempunyai 46 sampel berbanding 148 sampel dalam kelas majoriti (N) selepas pembersihan data dilakukan. Namun, selepas SMOTE digunakan, kelas minoriti telah diseimbangkan dengan menambahkan sampel sintetik sehingga mencapai jumlah yang sama dengan kelas majoriti. Jadual 3.7 menunjukkan contoh dua sampel dalam fitur *radius1* daripada kelas minoriti tersebut dalam set data WPBC.

Jadual 3.7 Contoh sampel fitur *radius1* bagi set data WPBC

Sampel	**Nilai *radius1***	**Kelas**
A	17.02	R
B	19.27	R

Pengiraan teknik ini diperincikan dengan mencari jiran terdekat dan menjana sampel sintetik menggunakan interpolasi linear. Pengiraan bermula dengan pemilihan sampel B iaitu nilai 19.27 sebagai jiran terdekat kepada sampel asal A dengan nilai 17.02. Nilai bagi δ dipilih secara rawak sebagai 0.6. Seterusnya, pengiraan sampel sintetik menggunakan persamaan ...(3.1) seperti berikut:

$$x_{baru} = 17.02 + 0.6(19.27 - 17.02) = 18.07 \quad \ldots(3.2)$$

Daripada pengiraan di atas, nilai sampel sintetik yang dihasilkan ialah 18.07 di mana ia berada di antara sampel asal A (17.02) dan jiran terdekatnya B (19.27). Proses interpolasi ini memastikan bahawa sampel sintetik yang dijana masih mengekalkan corak taburan data asal tanpa menyalin semula data sebenar.

3.5.3 Penormalan Data

Untuk memastikan setiap fitur menyumbang secara seimbang kepada model pembelajaran mesin, penormalan skala dilakukan menggunakan teknik skor Z (*Z-score normalization*). Teknik ini sesuai untuk data yang mempunyai taburan berbeza-beza dan tidak terhad kepada julat tertentu. Ia dapat mengelakkan fitur berskala besar daripada mendominasi model. Teknik skor Z ini juga dikenali sebagai *StandardScaler* dalam perisian *Python* dengan menormalkan setiap nilai fitur x kepada nilai piawai x' berdasarkan formula berikut:

$$x' = \frac{x - \mu}{\sigma} \quad \ldots(3.3)$$

di mana μ ialah purata bagi fitur tersebut dan σ ialah sisihan piawai bagi fitur tersebut. Setiap fitur berangka akan ditukarkan kepada skala baharu dengan purata sifar dan sisihan piawai satu. Ini memastikan semua atribut berada pada skala yang setara semasa latihan model. Jadual 3.8 menunjukkan contoh beberapa sampel dalam fitur *radius1* dalam set data WPBC.

Jadual 3.8 Contoh sampel fitur *radius1* bagi set data WPBC

Nombor sampel	Nilai asal
1	18.02
2	17.99
3	21.37
4	11.42
5	20.29

Pengiraan skor Z diperincikan bagi lima sampel pertama fitur tersebut sebagai contoh. Sebelum itu, nilai μ dan σ bagi fitur *radius1* dikira terlebih dahulu. Purata dikira dengan mengambil jumlah keseluruhan nilai dalam set data dan membahagikannya dengan bilangan sampel seperti berikut:

$$\mu = \frac{18.02 + 17.99 + 21.37 + 11.42 + 20.29}{5} \quad \ldots(3.4)$$

$$\mu = \frac{89.0}{5} = 17.818$$

Kemudian, SD dikira menggunakan formula berikut:

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+(x_3-\mu)^2+(x_4-\mu)^2+(x_5-\mu)^2}{N}} \quad \ldots(3.5)$$

$$\sigma = \sqrt{\frac{(18.02-17.818)^2+(17.99-17.818)^2+(21.37-17.818)^2 \\ +(11.42-17.818)^2+(20.29-17.818)^2}{5}}$$

$$\sigma = \sqrt{\frac{59.7341}{5}} = 3.456$$

Selepas mendapatkan nilai μ=17.818 dan σ=3.456, pengiraan nilai skor bagi setiap sampel seperti berikut:

$$Z_1 = \frac{18.02-17.818}{3.456} = 0.058 \quad \ldots(3.6)$$
$$Z_2 = \frac{17.99-17.818}{3.456} = 0.050$$
$$Z_3 = \frac{21.37-17.818}{3.456} = 1.028$$
$$Z_4 = \frac{11.42-17.818}{3.456} = -1.851$$
$$Z_5 = \frac{20.29-17.818}{3.456} = 0.715$$

Oleh itu, nilai selepas dinormalkan bagi setiap sampel adalah seperti Jadual 3.9:

Jadual 3.9 Nilai fitur *radius1* selepas penormalan

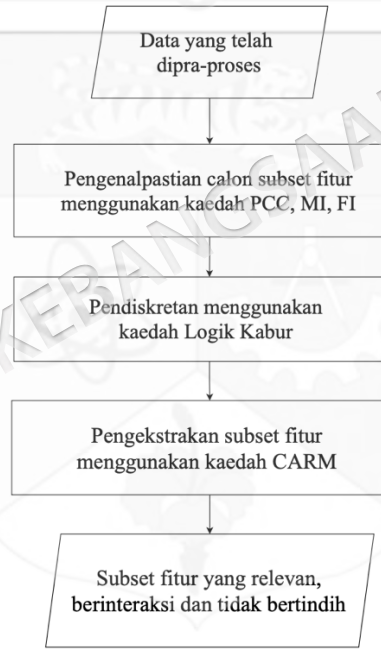
Nombor sampel	Nilai asal	Nilai Selepas Penormalan
1	18.02	0.0058
2	17.99	0.0050
3	21.37	1.028
4	11.42	-1.851
5	20.29	0.715

Berdasarkan hasil pengiraan pada Jadual 3.9, nilai skor Z positif seperti sampel 3 dan 5 menunjukkan bahawa sampel berada di atas purata, manakala nilai skor Z negatif seperti sampel 4 menandakan bahawa sampel berada di bawah purata. Nilai skor Z seperti sampel 1 dan 2 yang hampir kepada sifar pula menunjukkan bahawa sampel berada hampir dengan purata. Penggunaan skor ini memastikan bahawa setiap fitur mempunyai pengagihan data yang lebih seragam. Setiap set data ini kemudiannya

digunakan dalam fasa seterusnya bagi pemilihan fitur berasaskan interaksi menggunakan kaedah FD-CARI.

3.6 FASA 2: PEMILIHAN FITUR BERINTERAKSI

Fasa ini melibatkan pemilihan fitur berasaskan interaksi yang menggabungkan proses pengenalpastian calon subset fitur melalui penarafan fitur awal, pendiskretan kabur serta perlombongan peraturan menggunakan kaedah CARM. Tujuan utama adalah untuk mengenal pasti subset fitur yang relevan, tidak bertindih dan berinteraksi bagi meningkatkan keberkesanan model pengelasan. Fasa ini dijalankan melalui beberapa proses seperti yang ditunjukkan dalam Rajah 3.5.



Rajah 3.5 Proses pemilihan fitur

3.6.1 Pengenalpastian Calon Subset Fitur Melalui Penarafan Fitur Awal

Proses pertama dalam fasa ini adalah pengenalpastian calon subset fitur menggunakan kaedah penaraf di mana setiap fitur dinilai berdasarkan kepentingannya terhadap kelas sasaran. Terdapat tiga kaedah pemarkahan iaitu: (i) *Pearson Correlation Coefficient* (PCC), (ii) *Mutual Information* (MI) dan (iii) *Feature Importance* (FI).

Ketiga-tiga kaedah tersebut dipilih kerana masing-masing mewakili perspektif penilaian fitur yang berbeza tetapi saling melengkapi supaya dapat menghasilkan kedudukan yang lebih stabil dan adil serta sesuai untuk data perubatan (Jia Li et al. 2023; Minnoor & Baths 2023; Pacheco 2022).

PCC mengukur hubungan linear antara pasangan fitur di mana fitur yang mempunyai korelasi tinggi dengan pemboleh ubah sasaran sering dianggap lebih relevan (Benesty et al., 2009). Nilai PCC fitur X dan kelas Y dinilai menggunakan formula berikut:

$$\mathrm{PCC}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \quad \ldots(3.7)$$

di mana x_i dan y_i ialah nilai bagi setiap sampel dalam set data manakala $\overline{x}$ dan $\overline{y}$ ialah purata bagi fitur X dan kelas Y masing-masing.

MI pula digunakan untuk mengukur kebergantungan antara setiap fitur dan kelas sasaran di mana nilai yang lebih tinggi menunjukkan bahawa fitur tersebut mengandungi lebih banyak maklumat berguna untuk pengelasan (Battiti, 1994). Formula untuk menilai MI seperti berikut:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad \ldots(3.8)$$

di mana $p(x,y)$ ialah kebarangkalian bersama bagi X dan Y manakala $p(x)$ dan $p(y)$ ialah kebarangkalian marginal bagi X dan Y masing-masing.

FI pula digunakan untuk mengukur impak relatif setiap fitur dalam proses pengelasan di mana nilai kepentingan yang lebih tinggi menunjukkan fitur yang lebih signifikan dalam menghasilkan keputusan model (Breiman, 2001). FI dikira berdasarkan jumlah pengurangan bendasing Gini bagi semua pembahagian yang menggunakan fitur tersebut. GI bagi sesuatu nod t dengan kelas K dinyatakan seperti berikut:

$$G(t) = 1 - \sum_{i=1}^{K} p(i)^2 \quad \ldots(3.9)$$

di mana $p(i)$ ialah kebarangkalian pemilihan sampel secara rawak yang tergolong dalam kelas i dalam nod t. Oleh itu, kepentingan sesuatu fitur X dalam RF dianggarkan berdasarkan jumlah pengurangan bendasing yang dihasilkan oleh fitur tersebut dalam semua pokok dalam hutan yang boleh dinilai seperti berikut:

$$FI(X) = \sum_{t \in T} p_t \cdot \Delta\text{Gini}_t \quad \ldots(3.10)$$

di mana T ialah set semua nod pembahagian dalam pokok keputusan yang menggunakan fitur X manakala p_t ialah kebarangkalian nod t dan ΔGini_t ialah perubahan bendasing Gini selepas pembahagian pada nod t. ΔGini_t dikira sebagai:

$$\Delta\text{Gini}_t = G(t) - \left[\frac{N_L}{N} G(t_L) + \frac{N_R}{N} G(t_R)\right] \quad \ldots(3.11)$$

di mana N_L dan N_R ialah bilangan sampel dalam nod anak kiri dan kanan masing-masing manakala N ialah jumlah sampel di nod asal t. $G(t)$ pula ialah bendasing Gini sebelum pembahagian manakala $G(t_L)$ dan $G(t_R)$ ialah bendasing Gini bagi nod anak kiri dan kanan selepas pembahagian masing-masing.

Setiap kaedah ini menghasilkan senarai penarafan fitur berdasarkan kepentingan fitur yang dinilai melalui kaedah masing-masing. Walau bagaimanapun, tidak semua fitur dalam subset ini memberikan sumbangan yang signifikan kepada pengelasan. Oleh itu, satu proses penapisan fitur dilakukan untuk membuang fitur yang kurang bermakna atau tidak relevan menggunakan ambang sisihan piawai (SD) (Prasetiyowati et al., 2021)**.** Ambang ini boleh dikira seperti formula berikut:

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} \quad \ldots(3.12)$$

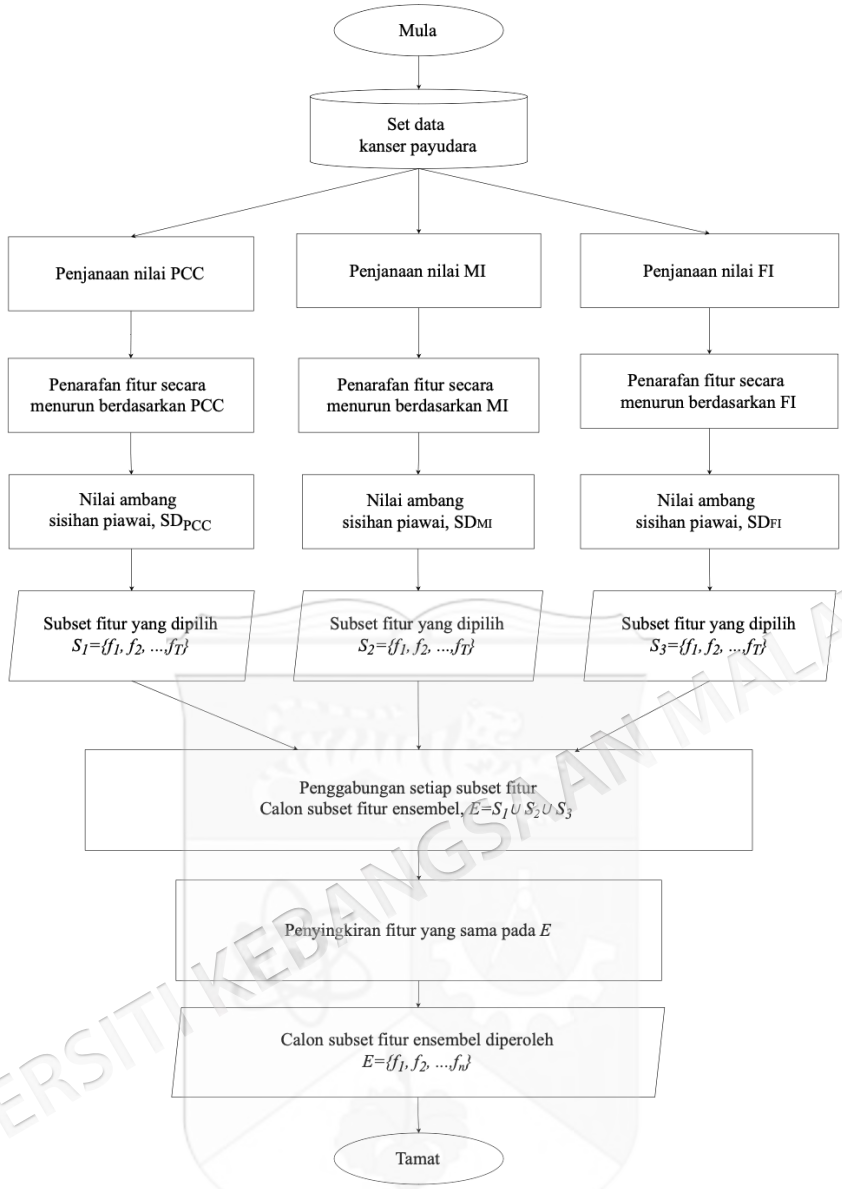
di mana x_i ialah nilai individu bagi setiap fitur i, $\overline{x}$ ialah jumlah purata bagi semua fitur dan n ialah jumlah fitur dalam set data.

Setelah semua fitur dinilai, calon subset fitur ensembel, E dibentuk. Calon subset ini menggabungkan semua subset fitur calon daripada ketiga-tiga kaedah penaraf menggunakan teori kesatuan yang dikenali sebagai *Union*. Fitur yang dipilih oleh sekurang-kurangnya satu kaedah penaraf akan dikekalkan dalam calon subset ini.

Pembentukan calon subset fitur ensembel daripada gabungan beberapa subset fitur calon $S_1, S_2, \ldots, S_N$ dinyatakan seperti persamaan berikut:

$$E = S_1 \cup S_2 \cup \ldots \cup S_N \quad \ldots(3.13)$$

Subset calon fitur ensembel, E yang terhasil dapat menangkap hubungan antara fitur yang paling relevan berdasarkan pelbagai perspektif penarafan fitur. Walau bagaimanapun, terdapat kemungkinan fitur yang dipilih oleh lebih daripada satu kaedah. Oleh itu, penyingkiran fitur yang sama dilakukan dengan hanya mengekalkan satu bagi setiap fitur yang dipilih untuk memastikan bahawa setiap fitur hanya muncul sekali dalam calon subset fitur akhir. Proses pertama ini boleh dilihat secara keseluruhan pada carta alir pada Rajah 3.6.



Rajah 3.6 Carta alir bagi proses pengenalpastian calon subset fitur melalui penarafan fitur awal

3.6.2 Pendiskretan Fitur Berterusan

Proses pendiskretan menjadi keperluan kepada kaedah CARM kerana ia bergantung kepada fitur berkategori. Namun, kaedah pendiskretan tradisional mempunyai masalah yang dikenali sebagai "*sharp boundary problem*" di mana nilai yang terletak berhampiran sempadan selang diperlakukan sama seperti nilai yang berada di tengah julat yang boleh mengakibatkan kehilangan maklumat penting dan penurunan ketepatan pengelasan (Kianmehr et al., 2010). Dalam kajian ini, pendiskretan kabur (*fuzzy discretization*) digunakan pada fitur berterusan (*continuous features*) bagi meningkatkan keupayaan penghasilan peraturan yang mudah ditafsir.

3.6.3 Set Kabur dan Fungsi Keahlian

Konsep set kabur berasal daripada kaedah Logik Kabur (*Fuzzy Logic*) yang diperkenalkan oleh Zadeh (1965) bagi mewakili ketidakpastian dan kekaburan dalam data. Dalam teori set klasik, sesuatu elemen ditentukan sama ada ahli kepada sesuatu set atau bukan ahli set tersebut yang boleh ditulis sebagai $x \in A$ atau $x \notin A$ masing-masing. Set ini dikenali sebagai set tegas. Namun begitu, keahlian sesuatu elemen dalam set kabur tidak ditentukan secara mutlak, tetapi diberikan darjah keahlian yang dinyatakan sebagai berikut:

$$A = \{(x, \mu_A(x)) | \, x \in X\} \qquad \ldots(3.14)$$

di mana X ialah set semesta dan $\mu_A(x)$: $X \rightarrow [0,1]$ ialah fungsi keahlian yang memetakan tahap keahlian setiap elemen x dalam set X kepada nilai antara 0 hingga 1. Jika $\mu_A(x) = 0$, ini menunjukkan bahawa x bukan ahli dalam A manakala jika $\mu_A(x) = 1$, x dianggap sebagai ahli penuh dalam A. Sebaliknya, jika nilai keahlian berada antara 0 dan 1, elemen tersebut mempunyai keahlian separa dalam set kabur.

Proses pengaburan dilakukan dengan menggunakan fungsi keahlian segi tiga kerana ia mudah dan efektif. Fungsi keahlian segi tiga digunakan untuk menentukan darjah keahlian sesuatu nilai dalam julat kategori tertentu dan diberikan oleh persamaan berikut:

$$\mu_A(X;\ a, b, c) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a \leq x \leq b \\ \dfrac{c-x}{c-b} & b \leq x \leq c \\ 0 & x \geq c \end{cases} \qquad \ldots(3.15)$$

dan juga boleh dirumuskan dengan menggunakan operator *min* dan *max* seperti persamaan berikut:

$$\mu_A(X;\ a, b, c) = \max\left[\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right] \qquad \ldots(3.16)$$

di mana x ialah elemen dalam set kabur X, a dan c masing-masing mewakili sempadan bawah dan atas bagi set kabur X, manakala b ialah pusat set kabur yang mempunyai tahap keahlian maksimum.

3.6.4 Prosedur Pendiskretan Kabur

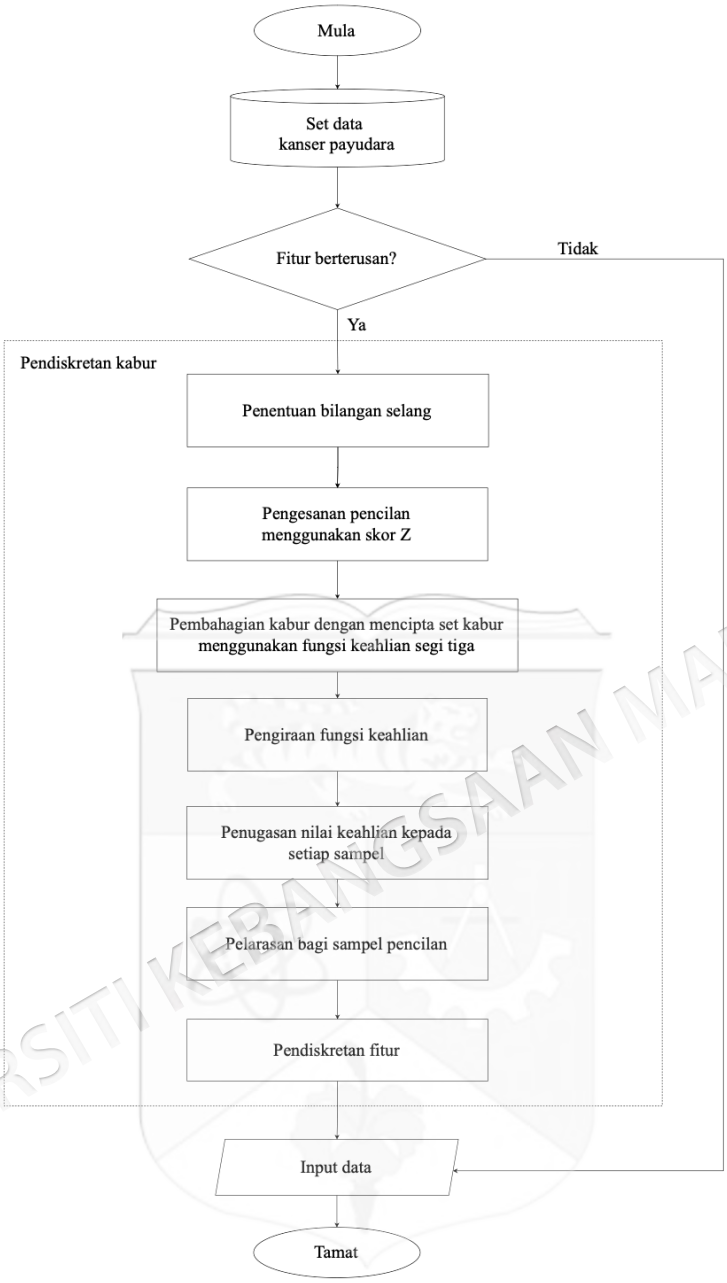
Proses pendiskretan kabur dalam kajian ini dilakukan dalam tiga langkah utama:

a. Pengenalpastian fitur berterusan dan pengesanan pencilan
   - Semua fitur berterusan dalam set data dikenal pasti untuk disediakan bagi pendiskretan kabur.
   - Pengesanan pencilan dilakukan menggunakan skor Z sebelum julat kategori ditentukan. Ini penting kerana proses *equal width division* boleh terganggu oleh nilai ekstrem.

b. Pembentukan set kabur dan fungsi keahlian
   - Set kabur dijana berdasarkan bilangan selang yang ditetapkan iaitu Rendah (*Low*), Sederhana (*Medium*) dan Tinggi (*High*).
   - Fungsi keahlian segi tiga digunakan untuk menilai tahap keahlian setiap nilai dalam julat kategori tersebut. Setiap nilai diberikan tahap keahlian bagi setiap kategori.

c. Penentuan nilai kategori akhir
   - Jika satu nilai fitur tergolong dalam lebih daripada satu kategori, kategori dengan tahap keahlian maksimum dipilih sebagai nilai akhir bagi sampel tersebut. Proses ini diulang untuk semua fitur berterusan sehingga seluruh set data telah didiskretkan sepenuhnya.

Keseluruhan proses pendiskretan kabur digambarkan dalam carta alir pada Rajah 3.7.

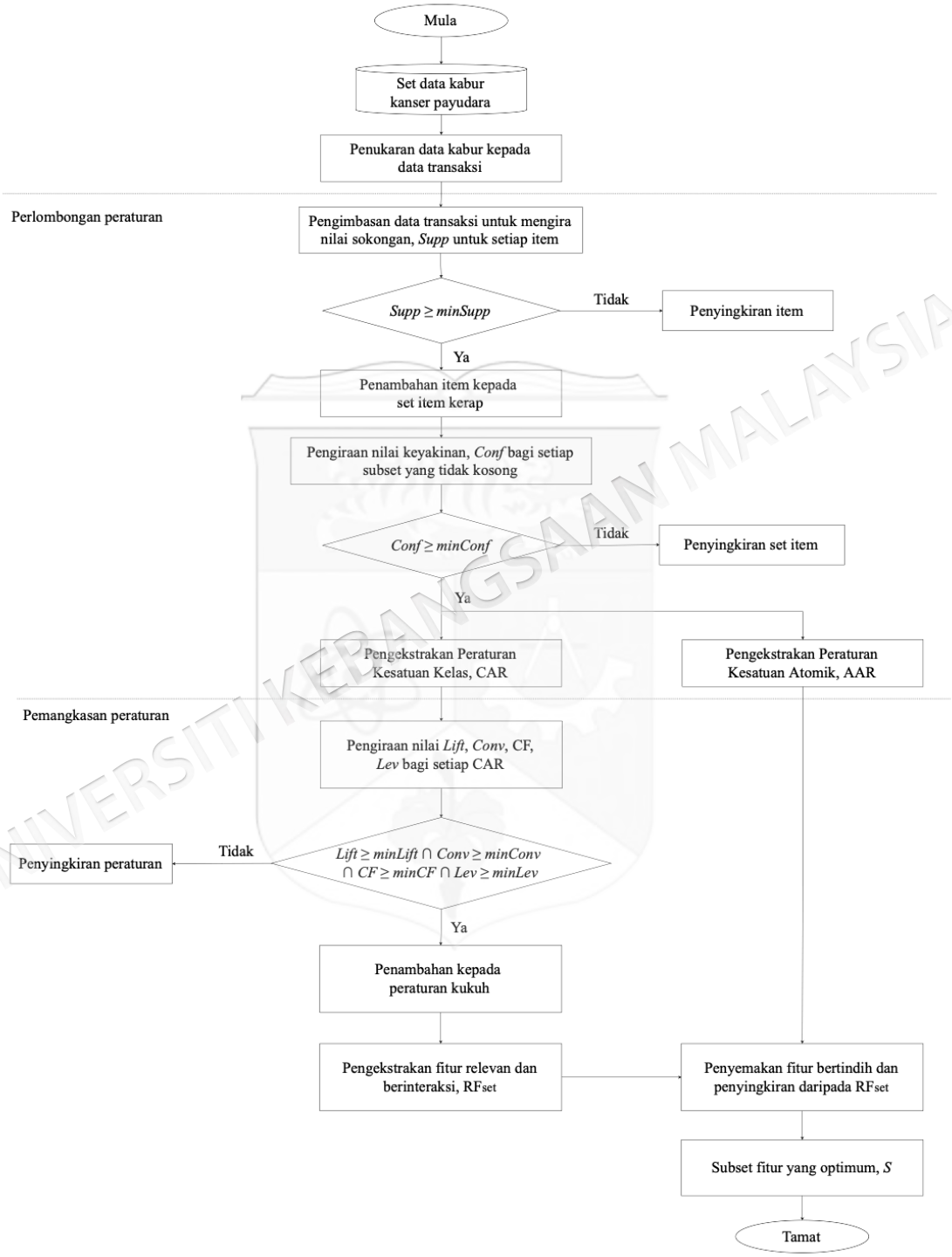


Rajah 3.7 Carta alir proses pendiskretan kabur

3.6.5 Pengekstrakan Subset Fitur Berasaskan Peraturan

Proses pengekstrakan subset fitur berasaskan peraturan merupakan langkah utama dalam fasa pemilihan fitur berinteraksi. Proses ini menggunakan kaedah CARM bagi mengenal pasti gabungan fitur yang mempunyai hubungan signifikan dengan kelas sasaran. Tujuannya adalah untuk menghasilkan subset fitur yang relevan, berinteraksi dan tidak bertindih yang dapat meningkatkan prestasi dan kebolehjelasan model

pengelasan. Proses ini dilaksanakan melalui empat langkah utama seperti berikut (lihat Rajah 3.8).



Rajah 3.8 Carta alir proses pengekstrakan subset fitur

Sebelum proses ini dijalankan, suatu langkah penting yang perlu dilakukan ialah transformasi data ke dalam format transaksi di mana setiap sampel dalam set data ditukar kepada binari iaitu nilai “1” atau “0”. Nilai "1" diberikan jika satu sampel mengandungi nilai fitur tertentu manakala nilai "0" jika sebaliknya.

a. Perlombongan Peraturan

Perlombongan peraturan bertujuan untuk mengenal pasti corak hubungan antara fitur. Algoritma *Apriori* dipilih dalam kajian ini kerana ia lebih sesuai dengan keperluan pembinaan peraturan kesatuan kelas yang memerlukan pemetaan eksplisit antara kombinasi fitur kepada kelas sasaran. *Apriori* menghasilkan set item kerap secara iteratif dan membolehkan setiap calon peraturan difahami dengan jelas dalam bentuk “Jika–Maka” iaitu struktur yang penting untuk interpretasi dalam konteks pengelasan perubatan. Berbanding algoritma *FP-Growth* yang menggunakan struktur *FP-tree* yang lebih kompleks serta tidak menghasilkan pemetaan fitur–kelas secara langsung, *Apriori* memberikan kawalan yang lebih telus terhadap proses pembentukan calon peraturan.

Selain itu, ketiga-tiga set data yang digunakan dalam kajian ini (WDBC, WPBC dan BCC) berada dalam kategori bersaiz kecil hingga sederhana serta mempunyai bilangan fitur yang masih praktikal untuk diproses secara cekap menggunakan algoritma *Apriori.* Dalam keadaan ini, kelebihan algoritma *FP-Growth* dari segi penjimatan memori tidak memberikan manfaat yang signifikan, manakala sifat iteratif *Apriori* membolehkan penjanaan peraturan yang lebih stabil dan mudah ditafsir. Terdapat beberapa kajian terdahulu dalam domain pengelasan kanser turut menunjukkan bahawa *Apriori* memberikan prestasi yang konsisten dan peraturan yang lebih bermakna berbanding alternatif lain dalam eksperimen berskala kecil sehingga sederhana (Alam et al. 2021; Keyvanpour et al. 2021; Kharya et al. 2022). Oleh itu, pemilihan *Apriori* dalam kajian ini adalah selaras dengan keperluan interpretabiliti, kestabilan peraturan dan sifat taburan data dalam set data kanser payudara.

Proses perlombongan peraturan ini bermula dengan imbasan pangkalan data transaksi bagi mengira tahap *Supp* setiap item dalam set data. Item yang mempunyai nilai *Supp* lebih tinggi atau sama dengan nilai *minSupp* akan ditambahkan ke dalam senarai set item kerap, manakala item yang tidak memenuhi syarat ini disingkir. Setelah

itu, set item kerap yang dikenal pasti digunakan untuk membina subset yang lebih besar dan bagi setiap subset yang tidak kosong, nilai *Conf* dikira. Hanya set item yang mempunyai nilai *Conf* sekurang-kurangnya nilai ambang *minConf* dikekalkan, manakala set item yang tidak memenuhi syarat ini disingkir.

Melalui proses ini juga, dua jenis peraturan kesatuan diekstrak iaitu peraturan kesatuan kelas (CAR) dan peraturan kesatuan atomik (AAR). CAR ialah peraturan yang mempunyai konsekuen hanya kepada kelas sasaran yang digunakan bagi mengekstrak fitur yang relevan dan berinteraksi manakala AAR merupakan peraturan yang mempunyai hanya satu item pada kedua-dua anteseden dan konsekuen yang digunakan bagi mengekstrak dan menyingkir fitur yang bertindih.

b. Pemangkasan Peraturan

Bagi mengelakkan kualiti peraturan yang rendah, proses pemangkasan CAR dijalankan menggunakan nilai angkat (*Lift*), sabitan (*Conv*), faktor kepastian (*CF*) dan keumpilan (*Lev*). Kriteria-kriteria tersebut dibincangkan seperti berikut (Akbas et al. 2022):

Nilai angkat mengukur kekerapan X (anteseden) dan Y (konsekuen) yang berlaku bersama-sama daripada yang dijangkakan jika ia bebas dari segi statistik. Angkat boleh ditakrifkan sebagai:

$$Lift(\mathrm{X}\Rightarrow\mathrm{Y}) = \frac{Conf(\mathrm{X}\Rightarrow\mathrm{Y})}{Supp(\mathrm{Y})} \qquad \ldots(3.17)$$

Jika nilai angkat > 1, ia menunjukkan korelasi positif antara. Jika nilai angkat < 1, ia menunjukkan korelasi negatif antara X dan Y. Jika nilai angkat =1, X dan Y tidak berkorelasi antara satu sama lain.

Sabitan pula ialah ukuran kekuatan implikasi peraturan $X\Rightarrow Y$. Ia mengukur kekerapan X membawa kepada Y berbanding jika Y berlaku secara bebas daripada X. Sabitan boleh dinyatakan sebagai:

$$Conv(\mathrm{X}\Rightarrow\mathrm{Y}) = \frac{1 - Supp(\mathrm{Y})}{1 - Conf(\mathrm{X}\Rightarrow\mathrm{Y})} \quad \ldots(3.18)$$

Jika nilai sabitan > 1, ia menunjukkan bahawa *X* sangat berkait rapat dengan *Y*.

Faktor kepastian pula ialah ukuran untuk menilai sejauh mana kejadian X meningkatkan kemungkinan Y berbanding kebarangkalian garis dasar Y. Ia membantu mengukur kekuatan perkaitan antara *X* dan Y yang boleh dikira seperti berikut:

$$CF(X \Rightarrow Y) = \begin{cases} \dfrac{Conf(\mathrm{X}\Rightarrow\mathrm{Y}) - Supp(\mathrm{Y})}{1 - Supp(\mathrm{Y})} & Conf(\mathrm{X}\Rightarrow\mathrm{Y}) > Supp(\mathrm{Y}) \\ \dfrac{Conf(\mathrm{X}\Rightarrow\mathrm{Y}) - Supp(\mathrm{Y})}{Supp(\mathrm{Y})} & Conf(\mathrm{X}\Rightarrow\mathrm{Y}) < Supp(\mathrm{Y}) \\ 0 & \text{sebaliknya} \end{cases} \quad \ldots(3.19)$$

Keumpilan pula mengukur perbezaan antara kekerapan sebenar *X* dan *Y* yang berlaku bersama-sama dan kekerapan yang dijangkakan jika *X* dan *Y* adalah bebas. Formula keumpilan ialah seperti berikut:

$$Lev(\mathrm{X}\Rightarrow\mathrm{Y}) = Supp(X \cup Y) - [Supp(X) \times Supp(Y)] \quad \ldots(3.20)$$

Jika nilai keumpilan = 0, tiada perkaitan antara *X* dan *Y*.

Peraturan yang tidak memenuhi ambang minimum *minLift, minConv, minCF, minLev* disingkirkan. Jadual 3.10 menunjukkan kriteria pemangkasan bagi CAR.

Jadual 3.10 Kriteria pemangkasan bagi peraturan kesatuan kelas

Set Data	***minLift***	***minConv***	***minCF***	***minLev***
WDBC	1.5	50	0.8	0.1
WPBC	1.0	2	0.5	0.03
BCC	1.5	*inf*	1.0	0.1

c. Pengekstrakan Fitur Berinteraksi

Setelah pemangkasan peraturan selesai, fitur berinteraksi yang relevan diekstrak daripada senarai peraturan CAR yang kukuh yang telah ditapis dan membentuk set fitur berinteraksi, RFset.

d. Penyingkiran Fitur Bertindih

Langkah terakhir dalam proses ini ialah penyingkiran fitur bertindih pada RFset. Proses ini dilakukan dengan menganalisis AAR. Fitur yang berada di bahagian konsekuen AAR dianggap sebagai bertindih dan disingkir daripada RFset. Setelah semua langkah ini selesai, subset fitur yang optimum, S diperoleh yang terdiri hanya daripada fitur-fitur relevan, berinteraksi dan tidak bertindih. Subset fitur ini akan digunakan dalam fasa seterusnya untuk proses pengelasan.

Kaedah FS yang dicadangkan ini dikenali sebagai ***F**uzzy **D**iscretization-**C**lass **A**ssociation **R**ule Mining based on **I**nteraction* (FD-CARI). Pseudokod bagi kaedah ini dirujuk pada Jadual 3.11.

Jadual 3.11 Pseudokod kaedah FD-CARI

Algoritma 3.1 Kaedah FD-CARI

Fungsi FD-CARI (*D, minSupp, minConf, minLift, minConv, minCF, minLev*):

Input: D: set sampel latihan; minSupp: ambang nilai sokongan minimum; minConf: ambang nilai keyakinan minimum; minLift: ambang nilai angkat minimum minConv: ambang nilai sabitan minimum; minCF: ambang nilai faktor kepastian minimum; minLev: ambang nilai keumpilan

Output: *S*: Subset fitur optimum

Algoritma

//Bahagian I: Pendiskretan kabur

1: $S = \emptyset$; $\mathrm{RF}_{\mathrm{set}} = \emptyset$;

2: diskretkabur(D);

//Bahagian II: Perlombongan Peraturan Kesatuan

4: $[\mathrm{CAR}, \mathrm{AAR}] = \mathit{Apriori}\ (D, \mathrm{minSupp}, \mathrm{minConf})$;

//Bahagian III: Pemangkasan dan Pengekstrakan Fitur Relevan Berinteraksi

5: **bagi** setiap peraturan $r \in \mathrm{CAR}$ **lakukan**

6: **jika** Lift < minLift ∩ Conv < minConv ∩ CF < minCF ∩ Lev < minLev **maka**

7: buang r daripada CAR;

8: $\mathrm{RF}_{\mathrm{set}} = \mathrm{RF}_{\mathrm{set}} \cup r \cdot$ Anteseden;

9: **tamat jika**

10: **tamat bagi**

//Bahagian IV: Penyingkiran Fitur Bertindih

11: menyusun(AAR);

12: **sementara** AAR ≠ ∅ **lakukan**

13: ambil peraturan pertama di mana r = n_peraturan_pertama(AAR);

14: AAR = AAR − {r};

15: **jika** $r \cdot$ Anteseden ⊂ $\mathrm{RF}_{\mathrm{set}}$;

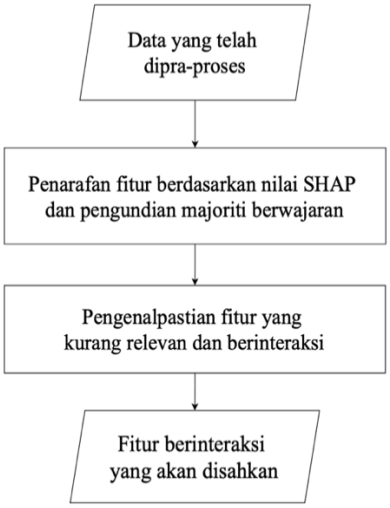
16: $\mathrm{RF}_{\mathrm{set}}$=$\mathrm{RF}_{\mathrm{set}} - r \cdot$ Konsekuen;

```
17:              bagi setiap r′ ∈ AAR lakukan
18:                  jika  r′ · Anteseden == r · Konsekuen maka
19:                  AAR = AAR − {r′}
20                   tamat jika
21:              tamat bagi
21:          tamat jika
22:    tamat sementara
//Bahagian V: Pengenalpastian Subset Fitur Optimum
23:    bagi setiap nilai fitur lakukan
24:       jika nilai ∈ RF_set maka
25:            S = S ∪ {F};
26:       tamat jika
27:    tamat bagi
28:    kembalikan S;
```

Jadual 3.11 menunjukkan pseudokod kaedah FD-CARI yang terdiri daripada lima bahagian utama. Baris 1 hingga 3 memulakan proses dengan penetapan pemboleh ubah dan pendiskretan kabur terhadap data. Seterusnya, baris 4 menjalankan proses perlombongan peraturan kesatuan menggunakan algoritma *Apriori* untuk menghasilkan set peraturan CAR dan AAR. Pada baris 5 hingga 9, setiap CAR ditapis berdasarkan nilai ambang bagi metrik *Lift*, *Conv*, *CF* dan *Lev*. Fitur yang berada dalam anteseden CAR yang memenuhi syarat akan dimasukkan ke dalam RFset. Baris 10 hingga 22 pula melibatkan penyingkiran fitur bertindih dengan membandingkan antara anteseden dan konsekuen AAR. Fitur yang berada pada konsekuen dianggap bertindih dan disingkir daripada RFset. Akhirnya, baris 23 hingga 28 mengekstrak subset fitur akhir S yang akan digunakan dalam fasa pembinaan model pengelasan.

3.7 FASA 3: PENGESAHAN FITUR BERINTERAKSI

Fasa ini memfokuskan kepada proses pengesahan fitur-fitur yang telah dipilih dalam fasa sebelumnya bagi memastikan bahawa fitur tersebut benar-benar berinteraksi secara signifikan dan menyumbang kepada prestasi pengelasan yang signifikan dalam proses pengelasan. Kaedah yang digunakan dalam fasa ini adalah berdasarkan nilai SHAP dan strategi pengundian majoriti berwajaran seperti dalam Rajah 3.9.



Rajah 3.9 Proses pengesahan fitur berinteraksi menggunakan nilai SHAP dan pengundian majoriti berwajaran

3.7.1 Penilaian Kepentingan Fitur

Langkah pertama melibatkan pengiraan nilai SHAP di mana ia mengukur kesan marginal setiap fitur terhadap keputusan model pengelasan. Kaedah ini membolehkan pemetaan yang lebih terperinci terhadap pengaruh individu dan gabungan fitur. Nilai SHAP dikira menggunakan formula berikut:

$$\varphi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} [f(S \cup \{i\}) - f(S)] \quad \ldots(3.21)$$

di mana φ_i ialah nilai SHAP bagi fitur i, F ialah set penuh fitur, S ialah subset fitur yang tidak termasuk i dan $f(S)$ ialah fungsi keputusan model apabila subset S digunakan.

Tiga model pembelajaran mesin digunakan untuk menjana nilai SHAP iaitu *RF, GB* dan XGB. Model-model ini dipilih kerana kestabilan dan prestasi konsisten yang sering dilaporkan dalam kajian terdahulu melibatkan data kesihatan (Ansyari et al., 2023; Dutta et al., 2024).

Model RF ialah model ensembel yang terdiri daripada pelbagai DT secara rawak di mana setiap DT membuat ramalan secara individu dan keputusan akhir ditentukan melalui majoriti mod keputusan yang dibuat oleh pokok-pokok tersebut (Breiman,

2001). RF membantu mengurangkan masalah "terlalu padan" dengan memilih subset rawak daripada fitur yang tersedia. Ramalan bagi RF seperti berikut:

$$RF(x) = \text{mod}(\{f_1(x), f_2(x), \ldots, f_T(x)\}) \quad \ldots(3.22)$$

di mana T ialah bilangan pokok keputusan dalam RF dan $f_T(x)$ ialah ramalan bagi pokok keputusan ke-T bagi data input x.

Model GB pula menggunakan formula seperti berikut:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} -(y(F_{m-1}(x_i) + \gamma) - \log\left(1 + e^{F_{m-1}(x_i)+\gamma}\right)) \quad \ldots(3.23)$$

di mana γ_{jm} ialah nilai γ yang meminimumkan fungsi kehilangan bagi kawasan ke-j, R_{jm} pada pokok ke-m manakala F_{m-1} ialah ramalan model ensembel sehingga lelaran ke-$m-1$ yang menggabungkan ramalan awal serta sumbangan daripada semua pembelajaran asas sebelumnya.

Model XGB merupakan pengoptimuman kepada model GB dengan menambah baik teknik kelaziman dan pengendalian sampel hilang. Persamaan kehilangan tersebut boleh dinilai menggunakan formula berikut:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad \ldots(3.24)$$

di mana $L^{(t)}$ ialah fungsi kehilangan pada lelaran ke-t, $l(y_i, \hat{y}_i^{(t-1)})$ ialah fungsi kehilangan individu dan $\Omega(f_t)$ ialah fungsi kelaziman yang mengawal kerumitan model.

Bagi setiap fitur, nilai SHAP daripada setiap model digabungkan menggunakan kaedah pengundian majoriti berwajaran, di mana nilai AUC model digunakan sebagai pemberat. Formula pengiraan adalah seperti berikut:

$$W_j = \sum_{i=1}^{n} (\mathrm{AUC}_i \times \mathrm{SHAP}_{i,j}) \quad \ldots(3.25)$$

di mana W_j ialah skor akhir kepentingan fitur j, AUC_i ialah skor AUC selepas penormalan bagi model i dan $\mathrm{SHAP}_{i,j}$ ialah nilai SHAP bagi fitur j dalam model i.

Fitur disusun mengikut skor W_j secara menurun dan fitur-fitur yang berada dalam kuantil ke-75 ke atas dianggap sebagai sangat relevan.

3.7.2 Penilaian Interaksi Fitur yang Kurang Relevan

Fitur yang tidak mencapai ambang ke-75 tidak terus disingkir, sebaliknya dianalisis semula berdasarkan potensi interaksi dengan fitur penting lain, selaras dengan cadangan kajian-kajian terdahulu (Engel et al., 2023; Koemel et al., 2024; Sylvester et al., 2024). Analisis ini menggunakan nilai interaksi SHAP dua arah yang dinilai seperti berikut:

$$\varphi_{ij} = f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S) \quad \ldots(3.26)$$

di mana φ_{ij} ialah nilai interaksi SHAP antara fitur i dan j, S ialah subset fitur tanpa fitur i dan j manakala $f(s)$ ialah fungsi keputusan model bagi subset S.

Fitur yang menunjukkan nilai interaksi SHAP yang tinggi dengan fitur utama atau fitur kurang relevan lain dianggap berpotensi menyumbang secara tidak langsung melalui interaksi. Fitur-fitur ini akan dipertimbangkan semula dalam proses analisis lanjut pada fasa seterusnya.

3.8 FASA 4: PENGOPTIMUMAN HIPERPARAMETER PEMILIHAN FITUR

Fasa ini memfokuskan kepada proses penalaan hiperparameter bagi kaedah FD-CARI yang telah dibina dalam Fasa 2 dengan tujuan untuk meningkatkan prestasi model pengelasan.

3.8.1 Penalaan Hiperparameter Kaedah FD-CARI+SA

Bagi model FD-CARI yang dibangunkan dalam kajian ini, proses penalaan menggunakan algoritma SA melibatkan pencarian nilai optimum bagi hiperparameter berikut:

- *minSupp*
- *minConf*
- *minLift*
- *minConv*
- *minCF*
- *minLev*

3.8.2 Prosedur Algoritma Penyepuhlindapan Simulasi

Algoritma SA dimulakan dengan penetapan satu penyelesaian awal, S_0 dan suhu awal, T_0 yang ditetapkan pada nilai yang tinggi. Penetapan suhu tinggi pada peringkat awal ini membolehkan algoritma meneroka ruang pencarian secara meluas dan tidak terhad kepada penyelesaian setempat bagi meningkatkan peluang untuk menemui kombinasi hiperparameter yang memberikan skor AUC yang lebih baik.

Seterusnya, satu penyelesaian baharu iaitu S_{baru} dijana dengan mengubah nilai hiperparameter daripada julat yang telah ditetapkan. Hiperparameter yang terlibat dalam proses ini merangkumi *minSupp*, *minConf*, *minLift*, *minConv*, *minCF* dan *minLev* dalam kaedah FD-CARI. Selepas penyelesaian baharu dihasilkan, kedua-dua penyelesaian S_{lama} dan S_{baru} akan dinilai menggunakan fungsi objektif berdasarkan skor AUC yang diperoleh daripada purata lima model pengelasan iaitu DT, RF, NB, LR dan SVM menggunakan subset fitur dihasilkan oleh kaedah FD-CARI menggunakan kombinasi hiperparameter semasa. Jika S_{baru} menghasilkan skor AUC yang lebih tinggi daripada S_{lama}, maka penyelesaian tersebut akan diterima serta-merta sebagai penyelesaian baharu. Namun demikian, sekiranya S_{baru} menghasilkan skor AUC yang lebih rendah, ia masih berpeluang untuk diterima berdasarkan kebarangkalian tertentu yang dikira menggunakan peraturan Metropolis. Kebarangkalian ini diberikan oleh formula berikut:

$$P = e^{\frac{\Delta f}{kT}} \quad \ldots(3.27)$$

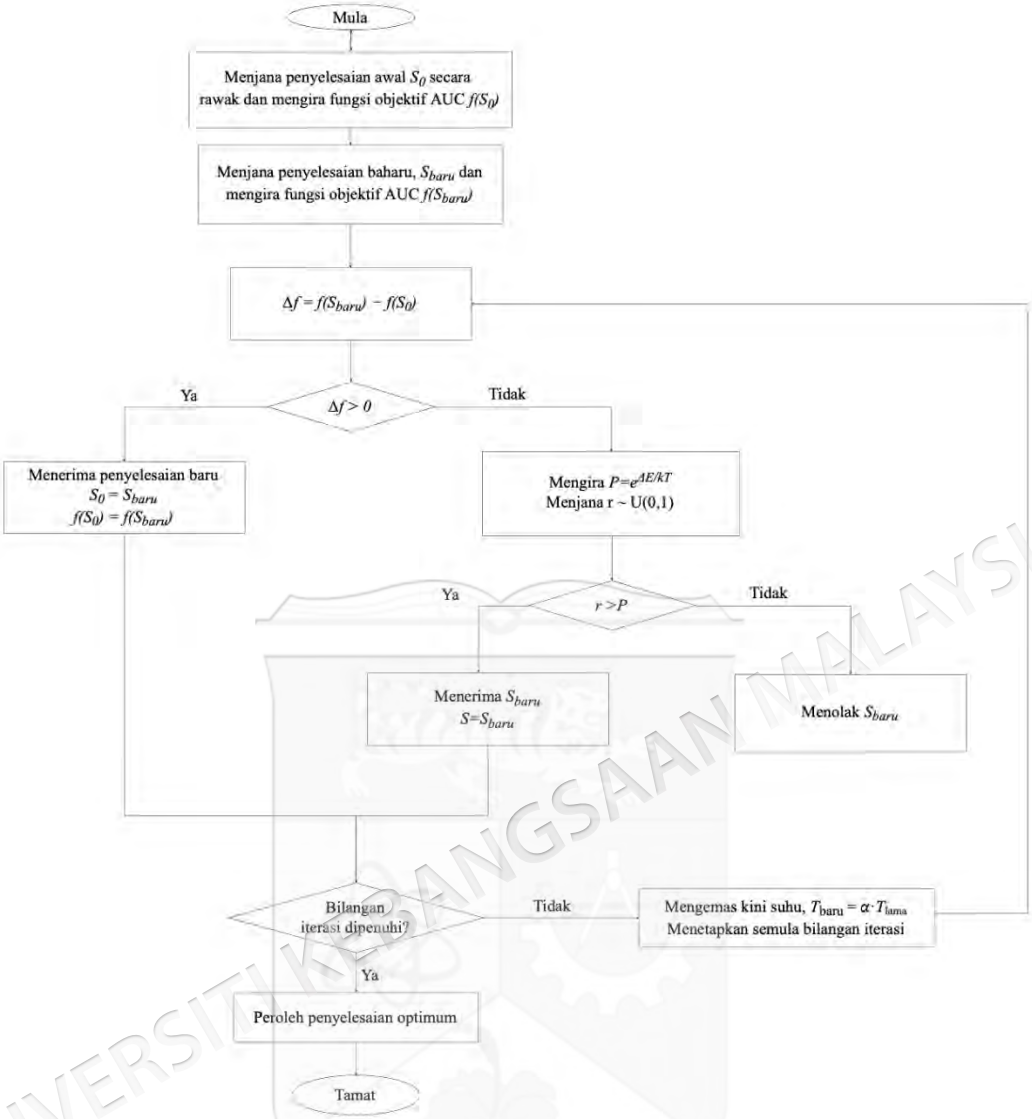
di mana P ialah kebarangkalian untuk menerima penyelesaian yang lebih buruk, $\Delta f = f(S_{\text{baru}}) - f(S_{\text{lama}})$ mewakili perubahan dalam nilai fungsi objektif skor AUC, k ialah pemalar dan T ialah suhu semasa. Mekanisme ini membolehkan algoritma keluar daripada perangkap optimum setempat dan meneruskan pencarian ke arah penyelesaian yang lebih baik secara global.

Selepas setiap lelaran selesai, suhu dikemas kini berdasarkan jadual penyejukan yang telah ditentukan. Kaedah penyejukan eksponen digunakan dalam kajian ini yang dirumuskan seperti berikut:

$$T_{\text{baru}} = \alpha \cdot T_{\text{lama}} \quad \ldots(3.28)$$

di mana α ialah faktor penyejukan dalam julat (0,1). Penurunan suhu secara eksponen ini memastikan bahawa kebarangkalian menerima penyelesaian yang lebih rendah akan berkurang secara beransur-ansur apabila suhu semakin menurun. Proses penyejukan ini diteruskan secara berperingkat sehingga mencapai bilangan maksimum lelaran yang telah ditetapkan. Setelah lelaran maksimum dicapai, proses pengoptimuman akan dihentikan dan kombinasi hiperparameter dengan skor AUC tertinggi akan dipilih sebagai penyelesaian akhir.

Rajah 3.10 menunjukkan carta alir proses penalan hiperparameter menggunakan algoritma SA dalam kajian ini.



Rajah 3.10 Carta alir algoritma penyepuhlindapan simulasi

Jadual 3.12 menunjukkan pseudokod kaedah FD-CARI dan algoritma SA secara keseluruhan.

Jadual 3.12 Pseudokod kaedah FD-CARI+SA

Algoritma 3.2 Kaedah FD-CARI+SA

Input:

D: set data latihan; S_0: penyelesaian awal; T: suhu permulaan ; k: pemalar Boltzmann; α: kadar penyejukan, *max_iter*: bilangan maksimum lelaran

Output:

S_{opt}: penyelesaian optimum; AUC_{opt}: skor AUC tertinggi

Algoritma

1: $T = T_0,\ S = S_0,\ \text{subset} = \text{FD-CARI}(D, S)$;

2: f = AUC (model(subset));
3: $S_{\text{opt}} = S$, $\text{AUC}_{\text{opt}} = f$;
4: **bagi** *i*=1 hingga *max_iter* **lakukan**
5: jana S_{baru} dengan mengubah semua hiperparameter secara rawak
6: $\text{subset}_{\text{baru}}$=FD-CARI($D, S_{\text{baru}}$);
7: f_{baru}=$\text{AUC}_{\text{purata}}$(model($\text{subset}_{\text{baru}}$));
8: **jika** $f_{\text{baru}} > f$ **maka**
9: $S = S_{\text{baru}}, f = f_{\text{baru}}$;
10: **jika** $f_{\text{baru}} > \text{AUC}_{\text{opt}}$ **maka**
11: $S_{\text{opt}} = S_{\text{baru}}, \text{AUC}_{\text{opt}} = f_{\text{baru}}$;
12: **tamat jika**
13: **jika tidak**
14: $\Delta f = f_{\text{baru}} - f$;
15: r = nombor rawak daripada (0,1);
16: **jika** $r < e^{\frac{\Delta f}{kT}}$ **maka**
17: $S = S_{\text{baru}}, f = f_{\text{baru}}$;
18: **tamat jika**
19: **tamat jika**
20: $T = \alpha \cdot T$;
21: **tamat bagi**
22: **kembalikan** S_{opt} dan AUC_{opt};

Pseudokod dalam Jadual 3.12 memperincikan gabungan kaedah FD-CARI dengan algoritma SA sebagai pengoptimuman hiperparameter. Pada baris 1 hingga 4, proses dimulakan penyelesaian awal hiperparameter S_0, suhu awal T_0 dan pengiraan skor AUC awal yang diperoleh berdasarkan purata prestasi daripada lima model pengelasan menggunakan subset fitur daripada kaedah FD-CARI. Dalam baris 5 hingga 21, algoritma menjalankan pencarian penyelesaian baharu secara ulangan. Bagi setiap lelaran, satu set penyelesaian baharu dijana secara rawak dan digunakan dalam FD-CARI untuk menghasilkan subset fitur. Sekiranya AUC baharu lebih tinggi daripada sebelumnya, penyelesaian akan dikemaskini. Jika prestasinya juga lebih baik daripada penyelesaian terbaik keseluruhan, maka nilai tersebut disimpan sebagai optimum. Jika sebaliknya, penyelesaian yang kurang baik masih boleh diterima berdasarkan kebarangkalian Metropolis. Proses diteruskan sehingga bilangan maksimum lelaran dipenuhi sambil suhu dikurangkan secara berperingkat mengikut kadar penyejukan. Pada akhir algoritma, kombinasi parameter terbaik serta purata skor AUC tertinggi akan dikembalikan (Baris 22) sebagai hasil akhir proses pengoptimuman.

3.9 FASA 5: PEMBINAAN DAN PENALAAN MODEL PENGELASAN

Fasa ini menerangkan proses pembinaan model pengelasan bagi pengelasan tumor kanser payudara menggunakan subset fitur yang telah dipilih dan disahkan dalam fasa sebelumnya. Lima model pembelajaran mesin terselia digunakan iaitu DT, RF, NB, LR dan SVM. Pemilihan model-model ini adalah berasaskan keberkesanan yang telah dibuktikan secara konsisten dalam pengelasan binari, khususnya dalam diagnosis kanser payudara menggunakan set data tabular (Islam et al., 2024; Mahesh et al., 2024; Uddin et al., 2023). Selain itu, pemilihan lima model ini merangkumi pelbagai paradigma pembelajaran mesin iaitu: (i) model linear: LR, (ii) model bukan linear: SVM, (iii) model berasaskan peraturan: DT, (iv) model *ensemble* berasaskan pokok: RF, dan (v) model berasaskan kebarangkalian: NB yang sekali gus memberikan penilaian lebih menyeluruh terhadap prestasi subset fitur yang dipilih merentasi algoritma yang berbeza. Pendekatan ini penting kerana data perubatan lazimnya bersifat heterogen dan tidak linear, maka penilaian silang menggunakan model daripada paradigma berbeza dapat meningkatkan kebolehpercayaan dan kekukuhan dapatan kajian.

3.9.1 Penyediaan Data bagi Pengujian

Bagi memastikan prestasi model dapat dinilai secara objektif, setiap set data dibahagikan kepada dua bahagian iaitu dua pertiga (2/3) untuk data latihan dan satu pertiga (1/3) untuk data pengujian. Pendekatan ini membolehkan model dilatih dengan set data yang mencukupi dan diuji menggunakan data yang tidak pernah dilihat sebelumnya. Proses pengujian ini diulang sebanyak sepuluh kali di mana model dilatih dan diuji semula dalam setiap pusingan. Hasil daripada setiap pusingan dianalisis bagi memperoleh purata prestasi yang lebih stabil dan boleh dipercayai (Wang et al., 2021).

3.9.2 Pengelasan Menggunakan Pembelajaran Mesin Terselia

Setiap model yang digunakan dalam kajian ini mempunyai sistem pengelasan yang berbeza. DT ialah model pembelajaran yang tidak berparameter, yang membina struktur hierarki bagi memisahkan data kepada kelas tertentu berdasarkan pengiraan *Information Gain* (IG) dan Entropi (Quinlan, 1986). Nilai Entropi dikira menggunakan formula berikut:

$$H(S) = -\sum_{i=1}^{K} p_i \log_2(p_i) \quad \ldots(3.29)$$

di mana p_i ialah perkadaran sampel dalam kelas i dalam set S.

Model RF pula dapat dinilai menggunakan formula pada persamaan ...(3.22). Selain itu, model NB ialah model kebarangkalian yang berdasarkan Teorem Bayes dengan andaian bahawa semua fitur adalah bebas secara bersyarat (Lewis, 1998). Model ini sesuai digunakan apabila terdapat banyak fitur dalam set data. Formula bagi Teorem Bayes seperti berikut:

$$P(C|X) = \frac{P(X|C) \times P(C)}{P(X)} \quad \ldots(3.30)$$

di mana $P(C|X)$ ialah kebarangkalian posterior bagi kelas C dengan fitur X dan $P(X|C)$ ialah kebarangkalian pemerhatian fitur X dalam kelas C. $P(C)$ ialah kebarangkalian kelas C dan $P(X)$ ialah kebarangkalian pemerhatian fitur X.

Model LR pula ialah model regresi logistik yang kerap digunakan untuk pengelasan binari. Fungsi utama LR ialah menentukan sempadan keputusan dalam ruang fitur yang dapat memisahkan dua kelas secara berkesan menggunakan fungsi logistik seperti berikut (Le Cessie & Van Houwelingen, 1992):

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T X}} \quad \ldots(3.31)$$

di mana dengan persamaan linear bagi pemboleh ubah bebas:

$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \quad \ldots(3.32)$$

di mana $h_\theta(x)$ ialah kebarangkalian ramalan bagi kelas positif manakala $\theta^T X$ ialah jumlah wajaran fitur.

Akhir sekali, model SVM digunakan untuk mencari hiperplan optimum yang dapat memisahkan dua kelas dengan margin maksimum dalam ruang dimensi tinggi (Cortes & Vapnik, 1995). Jika data tidak boleh dipisahkan secara linear, SVM menggunakan fungsi kernel untuk menukar data kepada dimensi yang lebih tinggi yang membolehkan pengelasan dilakukan dengan lebih tepat. Fungsi linear bagi SVM boleh diukur dengan persamaan berikut:

$$f(x) = sign(w^T x + b) \quad \ldots(3.33)$$

di mana x ialah vektor input, w^T ialah transposisi bagi vektor berat yang berserenjang dengan hiperplan, b ialah pemalar dan fungsi *sign* menentukan kelas label.

3.9.3 Penalaan Hiperparameter

Bagi meningkatkan prestasi model pengelasan, setiap model ditala menggunakan kaedah *Grid Search* (GS) iaitu kaedah konvensional yang mencuba semua kombinasi parameter dalam julat yang ditentukan. Proses penalaan hiperparameter adalah berbeza bagi setiap model GS digunakan untuk mencari kombinasi parameter tersebut kerana ia mudah dilaksanakan (Alibrahim & Ludwig, 2021). Pemilihan nilai hiperparameter bagi setiap model pengelasan dilakukan berdasarkan rujukan daripada kajian terdahulu (Di Carlo et al., 2023; Kusumaningrum et al., 2021) .

Bagi model DT, parameter *criterion* yang mengukur ketidakpastian dalam pemilihan atribut diuji menggunakan GI atau Entropi, manakala *max_depth* dan *min_samples_split* dipelbagaikan bagi menentukan had terbaik dalam pembinaan pokok keputusan. Model RF pula memerlukan penetapan *n_estimators* iaitu bilangan pokok dalam RF serta *max_features* yang menentukan bilangan fitur yang digunakan dalam setiap pokok keputusan. Bagi model NB pula, parameter *var_smoothing* diuji dalam beberapa skala bagi mengelakkan nilai kebarangkalian sifar yang boleh mengganggu kestabilan model. Selain itu, LR memerlukan pemilihan parameter *C* untuk kelaziman serta *penalty* untuk menentukan jenis kelaziman manakala bagi model SVM, pemilihan *kernel* menjadi aspek penting kerana ia menentukan bagaimana model memetakan data ke dalam ruang berdimensi tinggi dan parameter *C* dan *gamma* turut

diuji dalam pelbagai nilai untuk mendapatkan gabungan yang paling sesuai dalam memisahkan kelas.

Nilai hiperparameter yang diuji bagi setiap model dalam kajian ialah berdasarkan kajian Di Carlo et al. (2023) dan Kusumaningrum et al. (2021) yang disenaraikan dalam Jadual 3.13.

Jadual 3.13 Senarai hiperparameter bagi setiap model pengelasan

Model Pengelasan	Hiperparameter	Julat Nilai Diuji
DT	*criterion*	gini, entropy
	max_features	none, sqrt, log2
	min_samples_spilt	2, 4, 8, 12, 16, 20
	min_samples_leaf	1, 4, 8, 12, 16, 20
RF	*criterion*	gini, entropy
	max_features	none, sqrt, log2
	min_samples_spilt	2, 4, 8, 12, 16, 20
	min_samples_leaf	1, 4, 8, 12, 16, 20
NB	*var_smoothing*	logspace (0, -9, 100)
LR	*penalty*	L2, none
	C	0.001, 0.01, 0.1, 1, 10, 100, 1000
SVM	*kernel*	linear, poly, rbf, sigmoid
	C	0.1, 1, 10, 100, 1000
	gamma	1, 0.1, 0.01, 0.001, 0.0001

3.10 FASA 6: PENILAIAN PRESTASI MODEL

Fasa keenam ini merupakan peringkat akhir kajian yang bertujuan untuk menilai keberkesanan keseluruhan pendekatan yang dicadangkan merangkumi kaedah pemilihan fitur (FD-CARI), pengesahan fitur berinteraksi dan pengoptimuman hiperparameter (FD-CARI+SA). Penilaian prestasi dilakukan secara berperingkat berdasarkan struktur metodologi.

3.10.1 Penilaian Keberkesanan Model Pemilihan Fitur

Penilaian ini bertujuan untuk mengukur keberkesanan kaedah FD-CARI dalam mengenal pasti subset fitur yang relevan, berinteraksi dan tidak bertindih bagi meningkatkan prestasi pengelasan. Subset fitur yang dipilih melalui kaedah FD-CARI dibandingkan dengan lima kaedah pemilihan fitur piawai seperti CFS, FCBF,

Consistency, Relief-F dan mRMR. Setiap subset diuji menggunakan lima model pembelajaran mesin terselia iaitu DT, RF, NB, LR dan SVM.

Proses pengujian dijalankan menggunakan data latihan (2/3) dan data ujian (1/3) dengan 10 kali pengulangan rawak untuk mendapatkan purata prestasi yang stabil. Metrik penilaian prestasi yang digunakan seperti matriks kekeliruan (CM), luas di bawah kawasan lengkung (AUC), ketepatan (ACC), kepekaan (SEN), kekhususan (SPE) dan F1 (Tharwat, 2018).

a. Matriks Kekeliruan

Matriks ini digunakan untuk menilai bilangan pengelasan yang betul dan salah. Ia terdiri daripada empat komponen utama iaitu:

- Positif benar (TP): Bilangan sampel yang dikelaskan dengan betul sebagai positif. Sebagai contoh, tumor malignan yang dikenal pasti sebagai malignan.
- Negatif benar (TN): Bilangan sampel yang dikelaskan dengan betul sebagai negatif. Sebagai contoh, tumor benigna yang dikenal pasti sebagai benigna.
- Positif palsu (FP): Bilangan sampel yang dikelaskan secara salah sebagai positif (kesilapan Jenis I). Sebagai contoh, sampel benigna dikelaskan sebagai malignan.
- Negatif palsu (FN): Bilangan sampel yang dikelaskan secara salah sebagai negatif (kesilapan Jenis II). Sebagai contoh, sampel malignan dikelaskan sebagai benigna.

Kesalahan FN boleh membawa kesan kritikal dalam diagnosis kanser kerana ia menyebabkan pesakit tidak mendapat rawatan segera, kanser boleh merebak dan menjadi lebih sukar dirawat dan risiko maut meningkat. Manakala, kesalahan FP boleh menyebabkan risiko berlebihan seperti pesakit menjalani rawatan yang tidak perlu seperti biopsi, pembedahan dan kemoterapi serta tekanan emosi dan kewangan. CM ini memberikan gambaran jelas tentang jumlah kes yang dikelaskan dengan betul dan salah yang boleh dilihat pada Jadual 3.14.

Jadual 3.14 Matriks kekeliruan

		Kelas sebenar	
		Benigna	**Malignan**
Keputusan Pengelasan	**Benigna**	TN	FP
	Malignan	FN	TP

CM digunakan sebagai asas untuk mengira pelbagai metrik prestasi model seperti skor AUC, ACC, SEN, SPE dan F1.

b. Luas Kawasan di Bawah Lengkung ROC

AUC mengukur keupayaan model untuk membezakan antara dua kelas berdasarkan lengkung *Receiver Operating Characteristic* (ROC) yang dibina daripada *True Positive Rate* (TPR) dan *False Positive Rate* (FPR). TPR mengukur prestasi model dalam mengesan kes positif sebenar seperti pengesanan tumor malignan pada pesakit yang menghadapinya manakala FPR mengukur kadar kesilapan model apabila ia mengelaskan kes negatif (tumor benigna) sebagai positif (tumor malignan). Formula bagi TPR dan FPR adalah seperti berikut:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad \ldots(3.34)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \qquad \ldots(3.35)$$

AUC menilai model berdasarkan luas kawasan yang terletak di bawah lengkung ROC. Skor AUC sentiasa berada antara '0' dan '1' di mana model dengan nilai AUC yang lebih tinggi menunjukkan prestasi pengelasan yang lebih baik.

c. Ketepatan

Ketepatan (ACC) mengukur peratusan ramalan yang betul oleh model daripada jumlah keseluruhan kes yang diuji. ACC ditakrifkan seperti berikut:

$$\mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \qquad \ldots(3.36)$$

d. Kepekaan

Kepekaan (SEN) mengukur keupayaan model untuk mengesan kes positif sebenar. SEN ditakrifkan seperti berikut:

$$\text{SEN} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad \ldots(3.37)$$

e. Kekhususan

Kekhususan (SPE) mengukur keupayaan model untuk mengesan kes negatif sebenar. SPE ditakrifkan seperti berikut:

$$\text{SPE} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad \ldots(3.38)$$

f. Skor F1

Skor F1 adalah satu ukuran prestasi model yang mengambil kira kedua-dua keupayaan model mengesan kes positif dengan betul serta kebolehan model untuk mengurangkan kesilapan FP dan FN. Skor ini amat berguna apabila terdapat ketidakseimbangan kelas pada data. Skor F1 ditakrifkan seperti berikut:

$$\text{Skor F1} = \frac{2 \times \text{TP}}{2 \times \text{TP}+\text{FP}+\text{FN}} \quad \ldots(3.39)$$

g. Ujian Signifikan

Untuk menilai sama ada perbezaan prestasi antara kaedah FD-CARI dan kaedah pemilihan fitur piawai yang lain adalah signifikan, ujian *Shapiro-Wilk* dijalankan terlebih dahulu sebelum ujian statistik seterusnya. Ujian ini menguji sama ada perbezaan skor AUC bertaburan normal. Jika taburan adalah normal ($p > 0.05$), maka ujian *t* berpasangan digunakan. Jika sebaliknya ($p < 0.05$), ujian *Wilcoxon* digunakan kerana ia lebih sesuai untuk data yang tidak normal. Semua analisis dijalankan pada aras keyakinan 5% ($\alpha = 0.05$).

3.10.2 Pengesahan Fitur Berinteraksi

Pengesahan ini dilaksanakan dengan membandingkan subset fitur yang dipilih oleh FD-CARI dengan model interaksi berasaskan SHAP bagi menilai sejauh mana fitur terpilih mengandungi interaksi yang bermakna. Bagi memastikan liputan penilaian yang menyeluruh, semakan turut melibatkan semua fitur yang dipilih dan kurang relevan berdasarkan nilai SHAP sebelum proses penyingkiran fitur bertindih dijalankan. Setiap fitur ini diperiksa sama ada ia terlibat sekurang-kurangnya dalam satu pasangan interaksi yang berada dalam julat kuantil ke-75 atau ke-50 tertinggi berdasarkan nilai interaksi SHAP. Julat ini berkemungkinan mengandungi interaksi yang memberi kesan ketara terhadap prestasi model selaras dengan saranan beberapa kajian terdahulu (Castronuovo et al. 2023; Engel et al. 2023) .

Kehadiran interaksi yang signifikan bagi fitur yang kurang relevan ini menyokong bahawa kaedah FD-CARI tidak hanya menilai kekuatan individu fitur secara tersendiri, malah turut mengambil kira sumbangan interaktif antara kombinasi fitur.

3.10.3 Penilaian Keberkesanan Pengoptimuman Hiperparameter

Perbandingan prestasi model sebelum dan selepas pengoptimuman hiperparameter menggunakan algoritma SA dilakukan bagi menilai keberkesanan proses pengoptimuman terhadap kaedah FD-CARI. Terdapat dua metrik utama digunakan dalam perbandingan ini iaitu bilangan fitur yang dipilih dan purata skor AUC bagi lima model pengelasan yang diuji. Perbandingan ini bertujuan untuk menentukan sama ada penggunaan hiperparameter yang dioptimumkan melalui SA dapat menghasilkan subset fitur yang lebih kecil dan meningkatkan ketepatan pengelasan. Skor AUC bagi setiap model pengelasan yang menggunakan kaedah FD-CARI+SA juga dibandingkan dengan kaedah pemilihan fitur piawai yang lain.

Bagi menilai sama ada peningkatan prestasi selepas pengoptimuman adalah signifikan, ujian statistik turut dilaksanakan dengan membandingkan prestasi kaedah FD-CARI+SA dengan kaedah pemilihan fitur piawai.

3.11 KESIMPULAN

Bab ini telah membentangkan metodologi menyeluruh yang menjadi asas kepada pelaksanaan kajian. Rangka kerja metodologi yang dicadangkan merangkumi enam fasa utama iaitu pra-pemprosesan data, pemilihan fitur berinteraksi, pengesahan fitur berinteraksi, pengoptimuman hiperparameter pemilihan fitur, pembinaan dan penalaan model pengelasan dan penilaian prestasi model. Fasa pra-pemprosesan data melibatkan pembersihan, penyeimbangan dan penormalan set data bagi memastikan kualiti input untuk proses pemilihan fitur. Fasa seterusnya menumpukan kepada pemilihan fitur berasaskan interaksi menggunakan pendiskretan kabur dan kaedah CARM bagi mengekstrak subset fitur yang relevan, berinteraksi dan tidak bertindih. Proses ini disusuli oleh fasa penilaian interaksi fitur menggunakan gabungan model SHAP dan kaedah pengundian majoriti berwajaran bagi mengesahkan sumbangan interaksi fitur terhadap prestasi model. Seterusnya, pengoptimuman kaedah CARM dilaksanakan melalui penggunaan algoritma SA bagi mendapatkan konfigurasi hiperparameter terbaik yang dapat meningkatkan kecekapan model. Akhir sekali, prosedur penilaian prestasi keseluruhan model yang dibangunkan telah dihuraikan secara terperinci termasuk perbandingan dengan kaedah pemilihan fitur piawai, penggunaan pelbagai metrik prestasi dan ujian statistik bagi menilai keberkesanan kaedah yang dicadangkan secara empirikal.

BAB IV

PEMILIHAN FITUR YANG RELEVAN, BERINTERAKSI DAN TIDAK BERTINDIH MENGGUNAKAN KAEDAH PERATURAN KESATUAN KELAS

4.1 PENGENALAN

Bab ini membentangkan hasil kajian bagi objektif pertama iaitu pemilihan fitur berdasarkan interaksi menggunakan gabungan pendiskretan kabur dan Perlombongan Peraturan Kesatuan Kelas yang dikenali sebagai FD-CARI. Fokus utama adalah untuk mengenal pasti subset fitur yang relevan, berinteraksi dan tidak bertindih yang mampu meningkatkan prestasi model pengelasan kanser payudara. Kaedah FD-CARI melibatkan beberapa komponen penting seperti pengenalpastian calon fitur melalui penarafan awal fitur, pendiskretan kabur bagi menangani fitur berterusan dan pengekstrakan peraturan berasaskan CAR serta AAR. Hasil akhir daripada proses ini adalah subset fitur yang dirumus untuk digunakan dalam perbandingan lima model pengelasan. Perbandingan juga dilakukan dengan kaedah pemilihan fitur piawai berdasarkan beberapa metrik penilaian prestasi.

4.2 HASIL PRA-PEMPROSESAN DATA DAN PENGEKSTRAKAN SUBSET FITUR

4.2.1 Pembersihan Data

Pembersihan data dalam kajian ini dilakukan dengan menyingkirkan sampel yang mengandungi nilai hilang. Jadual 4.1 menunjukkan bahawa (i) set data WDBC dan BCC tidak mempunyai nilai hilang manakala (ii) set data WPBC kehilangan 4 rekod, namun masih cukup besar untuk analisis selanjutnya.

Jadual 4.1 Bilangan sampel sebelum dan selepas proses pembersihan data

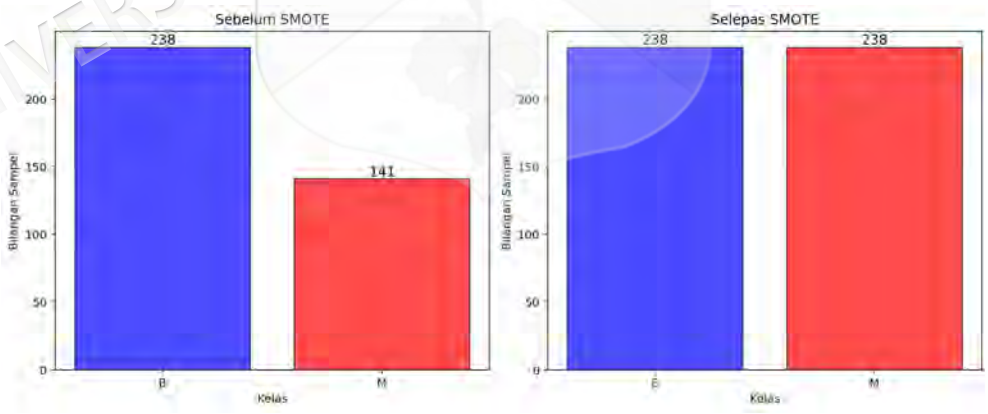
Set Data	Sebelum Pembersihan	Selepas Pembersihan
WDBC	569	569
WPBC	198	194
BCC	116	116

4.2.2 Penyeimbangan Data

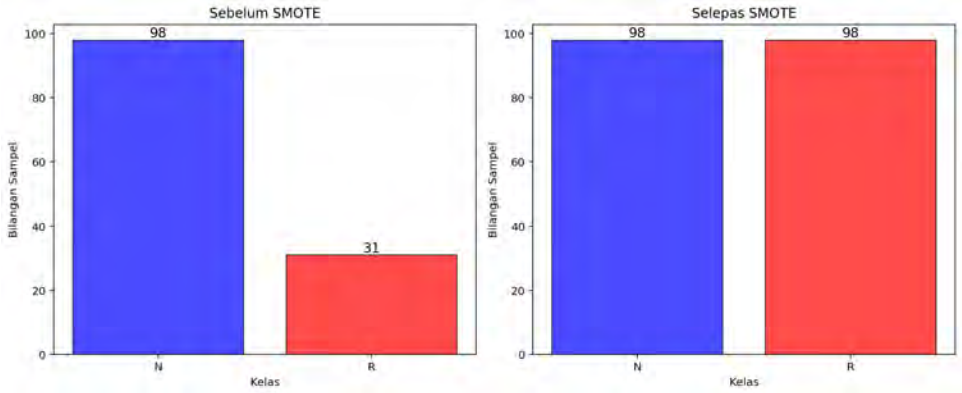
Ketidakseimbangan kelas ditangani dengan teknik SMOTE. Jadual 4.2 dan Rajah 4.1 menggambarkan kesan SMOTE yang menambah jumlah sampel bagi kelas minoriti sehingga seimbang dengan kelas majoriti. Pendekatan ini bertujuan memastikan tiada bias dalam proses seterusnya.

Jadual 4.2 Bilangan sampel latihan sebelum dan selepas proses penyeimbangan data

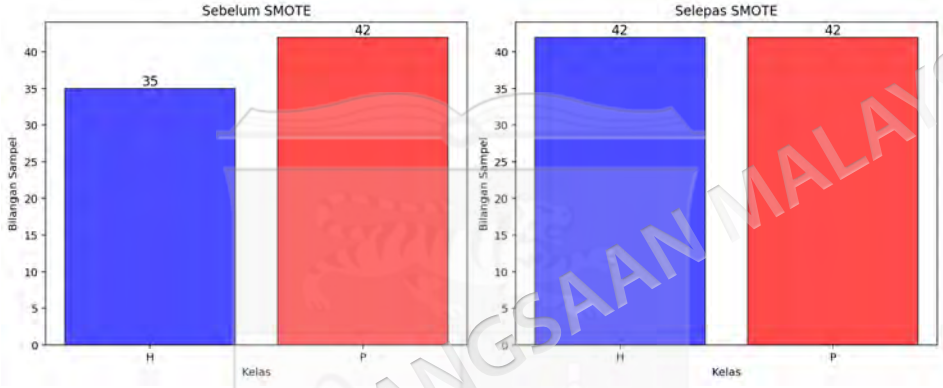
Set Data	Kelas	Sebelum Penyeimbangan	Selepas Penyeimbangan
WDBC	Benigna (B)	238	238
	Malignan (M)	141	238
WPBC	Tidak berulang (N)	98	98
	Berulang (R)	31	98
BCC	Kawalan Sihat (H)	35	42
	Pesakit (P)	42	42



(a)



(b)



(c)

Rajah 4.1 Set data (a) WDBC (b) WPBC (c) BCC sebelum dan selepas penyeimbangan data menggunakan SMOTE

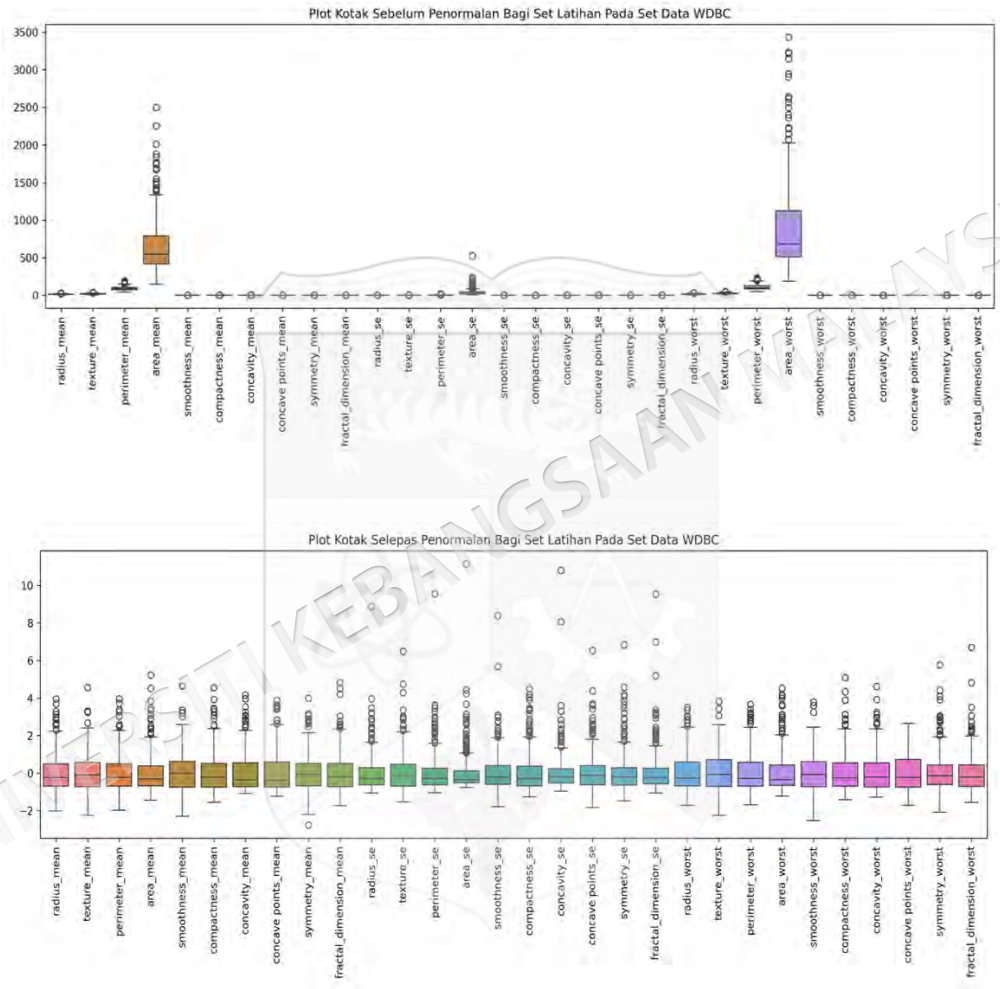
4.2.3 Penormalan Data

Penormalan data dilakukan menggunakan teknik skor Z bagi memastikan setiap fitur menyumbang secara seimbang. Jadual 4.3 dan Rajah 4.2 menunjukkan bagaimana fitur yang berbeza skala menjadi lebih setara selepas penormalan. Ini penting bagi mengelakkan dominasi fitur berskala besar ke atas prestasi model.

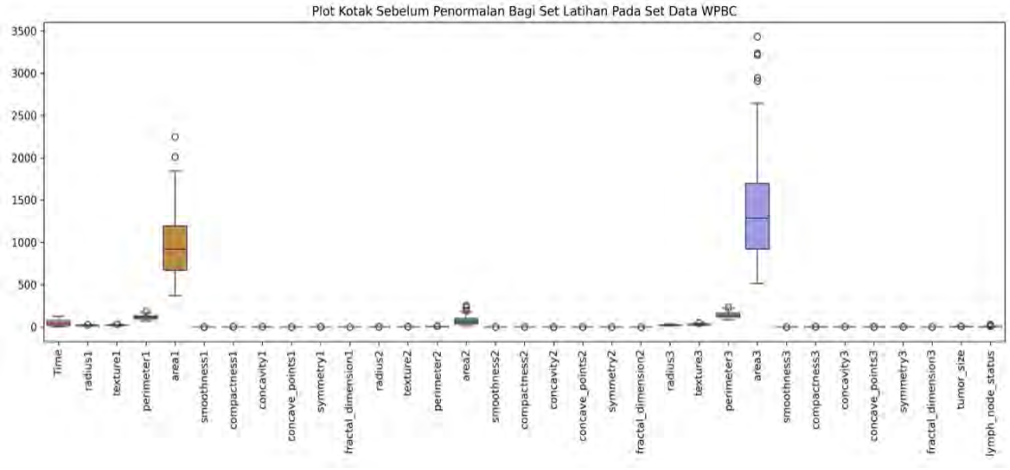
Jadual 4.3 Nilai beberapa fitur terpilih sebelum dan selepas proses penormalan data bagi set data WDBC, WPBC dan BCC

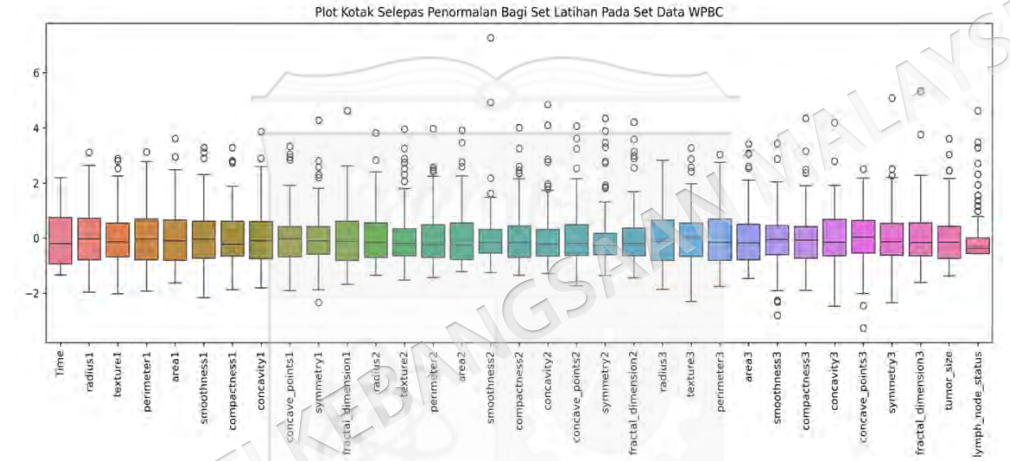
Set Data	Nama Fitur	Sebelum Penormalan	Selepas Penormalan
WDBC	*radius_mean*	11.6	-0.7116
	texture_mean	18.36	-0.2605
WPBC	*Time*	74	0.7406
	lymph_node_status	1	-0.3745
BCC	*Glucose*	84	-0.6655
	Insulin	2.869	-0.8077

Jadual 4.3 menunjukan julat nilai fitur yang berbeza sebelum penormalan. Sebagai contoh, fitur *Glucose* bernilai 84 berbanding *Insulin* yang hanya 2.869 dalam set data BCC. Namun selepas penormalan, *Glucose* bertukar kepada -0.6655 dan *Insulin* kepada -0.8077 di mana kedua-dua fitur menunjukkan kedudukan nilai yang lebih rendah daripada purata dalam set data tersebut.

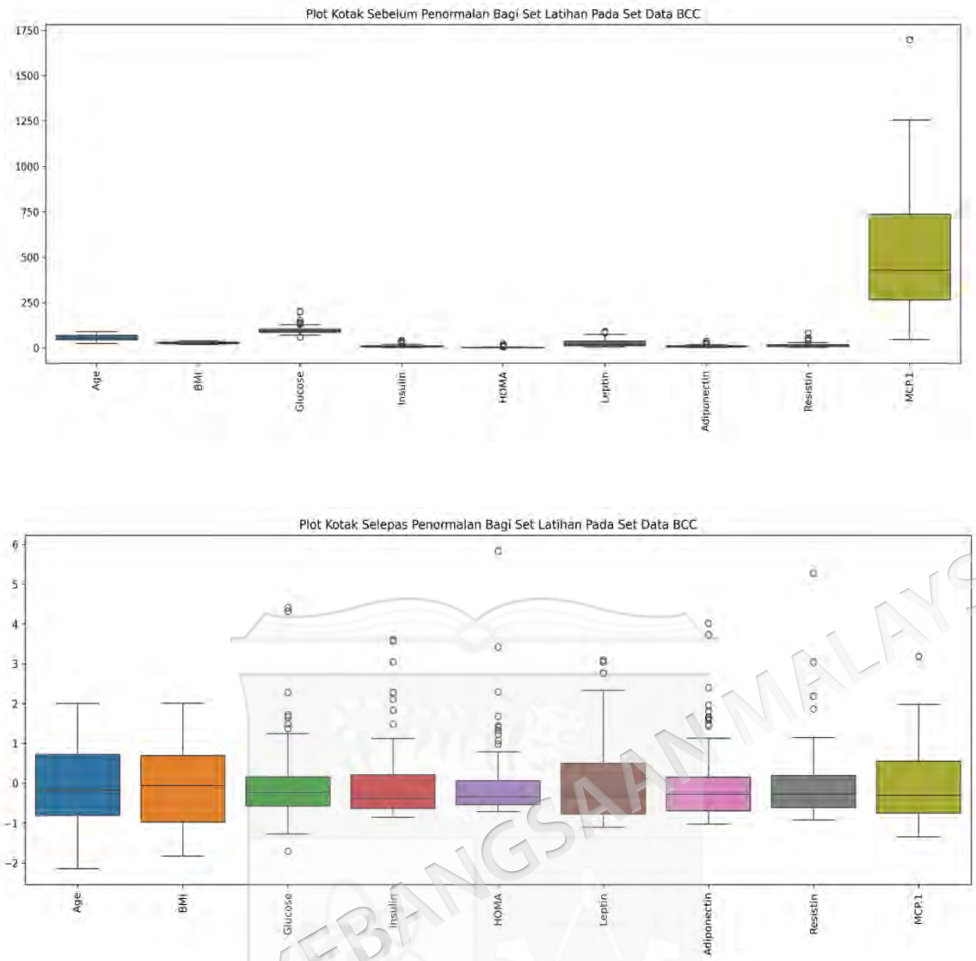


(a)





(b)



(c)

Rajah 4.2 Perbandingan plot kotak bagi nilai fitur sebelum dan selepas penormalan bagi set data (a) WDBC (b) WPBC (c) BCC

4.2.4 Pengenalpastian Calon Subset Fitur Menggunakan Kaedah Penaraf

Selepas data disediakan, calon subset fitur dikenal pasti berdasarkan nilai sisihan piawai (SD) bagi setiap kaedah penaraf PCC, MI dan FI bagi memastikan hanya fitur yang relevan digunakan dalam analisis. Jadual 4.4 sehingga Jadual 4.6 menyenaraikan fitur terpilih untuk setiap kaedah penaraf bagi setiap set data. Strategi ensembel digunakan di mana fitur yang melepasi ambang bagi sekurang-kurangnya satu kaedah dikekalkan sebagai calon subset.

Jadual 4.4 Calon fitur terpilih berdasarkan ambang SD bagi set data WDBC

PCC		MI		FI	
Fitur Terpilih (Ambang SD= 0.2674)	**Nilai PCC**	**Fitur Terpilih (Ambang SD= 0.1626)**	**Nilai MI**	**Fitur Terpilih (Ambang SD= 0.0401)**	**Nilai FI**
concave points_worst	0.7936	*perimeter_worst*	0.4788	*area_worst*	0.1394
perimeter_worst	0.7829	*area_worst*	0.4637	*concave points_worst*	0.1322
concave points_mean	0.7766	*radius_worst*	0.4514	*concave points_mean*	0.1070
radius_worst	0.7765	*concave points_worst*	0.4384	*radius_worst*	0.0828
perimeter_mean	0.7426	*concave points_mean*	0.4374	*perimeter_worst*	0.0808
area_worst	0.7338	*perimeter_mean*	0.4050	*perimeter_mean*	0.0680
radius_mean	0.7300	*concavity_mean*	0.3731	*concavity_mean*	0.0669
area_mean	0.7090	*radius_mean*	0.3650	*area_mean*	0.0605
concavity_mean	0.6964	*area_mean*	0.3593		
concavity_worst	0.6596	*area_se*	0.3408		
compactness_mean	0.5965	*concavity_worst*	0.3145		
compactness_worst	0.5910	*perimeter_se*	0.2726		
radius_se	0.5671	*radius_se*	0.2484		
perimeter_se	0.5561	*compactness_worst*	0.2255		
area_se	0.5482	*compactness_mean*	0.2129		
texture_worst	0.4569				
smoothness_worst	0.4215				
symmetry_worst	0.4163				
texture_mean	0.4152				
concave points_se	0.4080				
smoothness_mean	0.3586				
symmetry_mean	0.3305				
fractal_dimension_worst	0.3239				
compactness_se	0.2930				

Jadual 4.4 menunjukkan bahawa ambang SD yang digunakan pada set data WDBC ialah 0.2674 bagi kaedah PCC, 0.1626 bagi MI dan 0.0401 bagi FI. Sebanyak 24 fitur telah dipilih menggunakan kaedah PCC, 15 fitur menggunakan kaedah MI dan 8 fitur menggunakan kaedah FI. Kaedah PCC memilih fitur utama seperti *concave points_worst*, *perimeter_worst*, *concave points_mean* dan *radius_worst*. Kaedah MI dan FI turut mengenal pasti beberapa fitur yang sama termasuk *area_worst* dan *radius_worst*. Sebaliknya, 6 fitur seperti *fractal_dimension_mean*, *texture_se*, *smoothness_se*, *concavity_se, symmetry_se* dan *fractal_dimension_se* tidak melepasi ambang bagi ketiga-tiga kaedah dan telah dikeluarkan daripada calon subset kerana dianggap kurang relevan untuk tugas pengelasan.

Jadual 4.5 Calon fitur terpilih berdasarkan ambang SD bagi set data WPBC

PCC		MI		FI	
Fitur Terpilih (Ambang SD= 0.1267)	**Nilai PCC**	**Fitur Terpilih (Ambang SD= 0.0193)**	**Nilai MI**	**Fitur Terpilih (Ambang SD= 0.0120)**	**Nilai FI**
area3	0.2228	*compactness1*	0.0587	*Time*	0.0881
radius3	0.2211	*concavity2*	0.0532	*smoothness3*	0.0464
perimeter3	0.2171	*symmetry1*	0.0511	*lymph_node_status*	0.0417
area1	0.1787	*Time*	0.0492	*fractal_dimension1*	0.0375
tumor_size	0.1731	*radius2*	0.0479	*concave_points2*	0.0373
lymph_node_status	0.1692	*texture3*	0.0405	*texture2*	0.0360
perimeter1	0.1647	*smoothness2*	0.0351	*texture3*	0.0344
radius1	0.1636	*lymph_node_status*	0.0327	*tumor_size*	0.0336
area2	0.1359	*tumor_size*	0.0298	*concavity2*	0.0310
		concavity3	0.0252	*concavity1*	0.0306
		texture1	0.0216	*perimeter2*	0.0298
		area3	0.0214	*area3*	0.0296
				radius3	0.0294
				concave_points3	0.0290
				fractal_dimension2	0.0282
				radius1	0.0279
				smoothness2	0.0279
				compactness2	0.0270
				symmetry3	0.0269
				fractal_dimension3	0.0248
				area1	0.0247
				compactness3	0.0247
				symmetry2	0.0245
				texture1	0.0245
				perimeter3	0.0243
				concavity3	0.0241
				concave_points1	0.0241
				radius2	0.0236
				symmetry1	0.0231
				perimeter1	0.0230
				smoothness1	0.0223
				area2	0.0212
				compactness1	0.0189

Jadual 4.5 menunjukkan bahawa ambang SD yang digunakan ialah 0.1267 bagi kaedah PCC, 0.0193 bagi kaedah MI dan 0.0120 bagi kaedah FI untuk set data WPBC. Sebanyak 9 fitur telah dipilih melalui kaedah PCC, 12 fitur melalui kaedah MI dan semua fitur iaitu 33 fitur melalui kaedah FI. Kaedah PCC memilih fitur utama termasuk

area3, *radius3*, *perimeter3*, *area1* dan *tumor_size* manakala kaedah MI memilih fitur *compactness1*, *concavity2*, *symmetry1* dan *Time*. Sementara itu, kaedah FI mengenal pasti fitur penting seperti *Time*, *smoothness3*, *lymph_node_status* dan *fractal_dimension1*. Oleh disebabkan subset akhir ditentukan berdasarkan gabungan subset seperti dalam persamaan ...(3.12), fitur yang dipilih oleh sekurang-kurangnya satu kaedah dikekalkan. Maka, tiada fitur dikeluarkan dalam proses ini kerana kaedah FI memilih kesemua 33 fitur.

Jadual 4.6 Calon fitur terpilih berdasarkan ambang SD bagi set data BCC

PCC		**MI**		**FI**	
Fitur Terpilih (Ambang SD= 0.1797)	**Nilai PCC**	**Fitur Terpilih (Ambang SD= 0.0488)**	**Nilai MI**	**Fitur Terpilih (Ambang SD= 0.0456)**	**Nilai FI**
Glucose	0.3843	*Age*	0.4788	*Glucose*	0.2095
HOMA	0.2840	*Glucose*	0.4637	*Resistin*	0.1301
Insulin	0.2768	*Resistin*	0.4514	*BMI*	0.1281
Resistin	0.2273			*HOMA*	0.1206
				Age	0.1160
				Insulin	0.0900
				Leptin	0.0800
				Adiponectin	0.0712
				MCP.1	0.0545

Jadual 4.6 menunjukkan fitur yang dipilih berdasarkan ambang SD iaitu 0.1797 untuk PCC, 0.0488 untuk MI dan 0.0456 untuk FI. Sebanyak 4 fitur telah dipilih menggunakan kaedah PCC, 3 fitur menggunakan kaedah MI dan 9 fitur menggunakan kaedah FI. Kaedah PCC memilih fitur termasuk *Glucose*, *HOMA*, *Insulin*, dan *Resistin* manakala kaedah MI pula mengenal pasti *Age*, *Glucose*, dan *Resistin* sebagai fitur paling penting. Sementara itu, kaedah FI memilih kesemua fitur dalam set data BCC sebagai penting. Oleh disebabkan kaedah FI mengekalkan semua fitur, tiada fitur dikecualikan dalam pemilihan calon subset fitur.

Jadual 4.7 menyenaraikan senarai lengkap calon subset fitur ensembel bagi setiap set data.

Jadual 4.7 Bilangan fitur sebelum dan selepas pengenalpastian calon subset fitur

Set Data	Bilangan Fitur (Sebelum)	Bilangan Fitur (Selepas)	Senarai Calon Subset Fitur Ensembel
WDBC	30	24	*radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, radius_se, perimeter_se, area_se, compactness_se, concave points_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst*
WPBC	32	32	*Time, radius1, texture, perimeter1, area1,smoothness1, compactness1, concavity1, concave_points1, symmetry1, fractal_dimension1, radius2, texture2, perimeter2, area2, smoothness2, compactness2, concavity2, concave_points2, symmetry2, fractal_dimension2, radius3, texture3, perimeter3, area3, smoothness3, compactness3, concavity3, concave_points3, symmetry3, fractal_dimension3, tumor_size, lymph_node_status*
BCC	9	9	*Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP.1*

Proses penarafan fitur berjaya menghasilkan subset awal yang lebih relevan. Sebanyak 24 daripada 30 fitur dipilih bagi set data WDBC manakala semua fitur dikekalkan bagi WPBC dan BCC kerana semua fitur dipilih sekurang-kurangnya oleh satu daripada tiga kaedah penaraf. Keputusan ini dipengaruhi oleh kaedah FI yang cenderung menilai semua fitur sebagai penting, sekali gus menyokong objektif kajian untuk mengekalkan sebanyak mungkin fitur berpotensi interaktif bagi fasa perlombongan peraturan dan pengesahan seterusnya. Pendekatan ini mengelakkan penyingkiran awal fitur yang mungkin kurang relevan secara individu tetapi penting dalam kombinasi interaksi.

4.2.4 Pendiskretan Kabur untuk Penyediaan Fitur dalam Perlombongan Peraturan

Pendiskretan kabur bertujuan untuk menukarkan fitur berterusan kepada kategori kabur yang lebih sesuai digunakan dan lebih efektif dalam kaedah berasaskan peraturan seperti CARM. Pendiskretan kabur dipilih untuk digunakan dalam kajian ini kerana ia mampu mengatasi masalah sempadan keras (*sharp boundary problem)* yang wujud dalam kaedah pendiskretan tradisional seperti *equal-width*, *equal-frequency* dan kaedah berasaskan entropi termasuk C4.5. Walaupun kaedah berasaskan entropi seperti C4.5 lazim digunakan dan berkesan untuk mengenal pasti titik pemotongan optimum bagi

keputusan dua kelas, ia masih menghasilkan selang (*interval*) diskret yang bersifat keras (*crisp)* di mana setiap nilai fitur hanya tergolong dalam satu kategori sahaja. Pendekatan ini boleh menyebabkan kehilangan maklumat, terutama apabila nilai sebenar berada hampir dengan sempadan kategori.

Sebaliknya, pendiskretan kabur lebih sesuai untuk data kanser payudara kerana kebanyakan fitur morfologi dan biokimia (contohnya *radius, texture, concavity* dan *perimeter*) berubah secara beransur-ansur dan tidak mempunyai sempadan semula jadi. Sifat kabur yang membenarkan nilai fitur mempunyai lebih daripada satu darjah keanggotaan (contohnya Rendah–Sederhana, Sederhana–Tinggi) menghasilkan perwakilan yang lebih realistik terhadap variasi biologi sebenar. Ini bukan sahaja mengurangkan kehilangan maklumat, tetapi juga menghasilkan peraturan yang lebih stabil, lebih bermakna secara klinikal dan kurang sensitif kepada hingar berbanding pendiskretan keras.

Tambahan pula, pendiskretan kabur terbukti menghasilkan peraturan yang lebih mudah dijelaskan dan lebih konsisten dalam perlombongan peraturan berasaskan ARM terutamanya apabila berurusan dengan hubungan tidak linear dan interaksi fitur. Beberapa kajian terdahulu turut melaporkan bahawa pendiskretan kabur mampu meningkatkan ketepatan pengelasan dan kestabilan corak yang ditemui dalam data perubatan berbanding kaedah entropi tradisional, sekali gus menyokong pemilihan pendekatan ini sebagai langkah pra-pemprosesan dalam kajian ini.

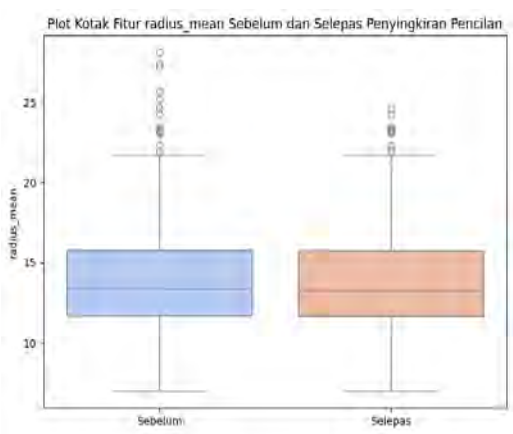
Perlaksanaan pendiskretan ini melibatkan beberapa langkah seperti berikut:

a. Pengesanan Pencilan

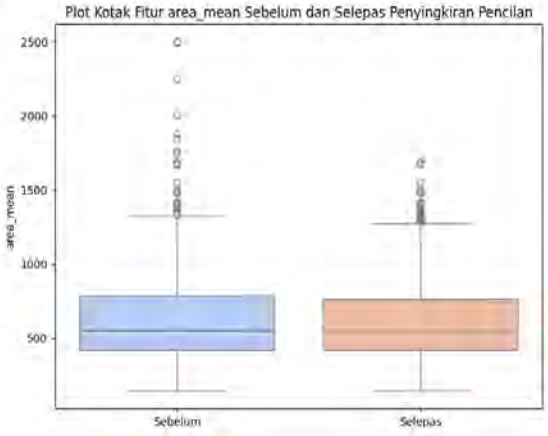
Proses pengesanan dan penyingkiran pencilan menggunakan teknik skor-Z adalah berdasarkan julat ambang $-3<Z<3$. Jadual 4.8 menunjukkan julat nilai bagi beberapa fitur terpilih sebelum dan selepas penyingkiran pencilan dalam setiap set data dan digambarkan dalam Rajah 4.3. Dalam set data WDBC, fitur *radius_mean* mempunyai 5 pencilan manakala *area_mean* mencatatkan 8 pencilan. Bagi set data WPBC pula, dua fitur iaitu *symmetry1* dan *fractal_dimension1* masing-masing mempunyai 1 fitur terpencil. Sementara itu, dalam set data BCC, *Glucose* dan *Insulin* masing-masing mempunyai 3 dan 4 pencilan. Penyingkiran pencilan ini penting untuk memastikan proses pendiskretan kabur tidak terkesan oleh nilai luar jangka yang boleh memesongkan sempadan fungsi keahlian.

Jadual 4.8 Julat nilai fitur terpilih sebelum dan selepas penyingkiran pencilan

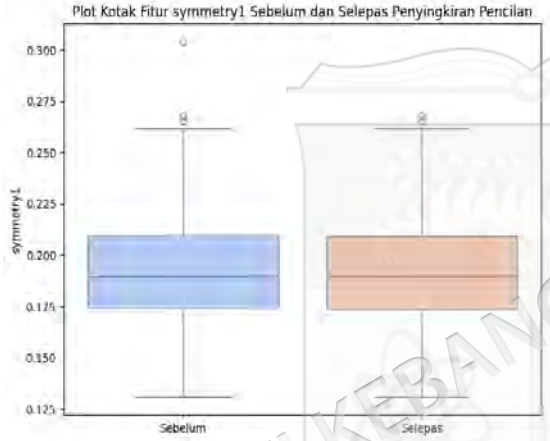
Set Data	Contoh Fitur	Julat Sebelum Pencilan	Julat Selepas Pencilan	Bilangan Pencilan
WDBC	*radius_mean*	[6.981, 28.11]	[6.981, 24.63]	5
	area_mean	[143.5, 2501]	[143.5, 1686.0]	8
WPBC	*symmetry1*	[0.13, 0.30]	[0.13, 0.27]	1
	fractal_dimension1	[0.05, 0.10]	[0.05, 0.08]	1
BCC	*Glucose*	[60, 201]	[60, 152]	3
	Insulin	[2.43, 58.46]	[2.43, 36.94]	4



(a)

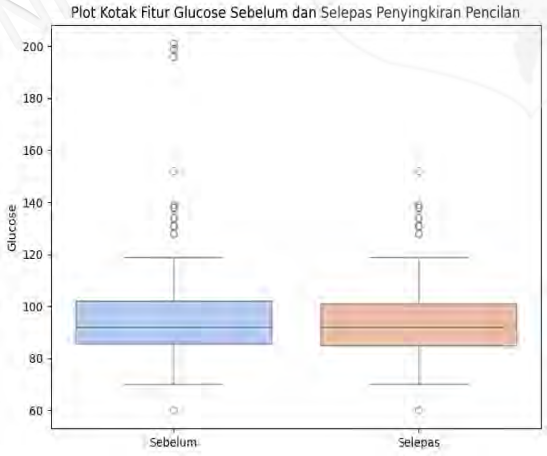


(b)

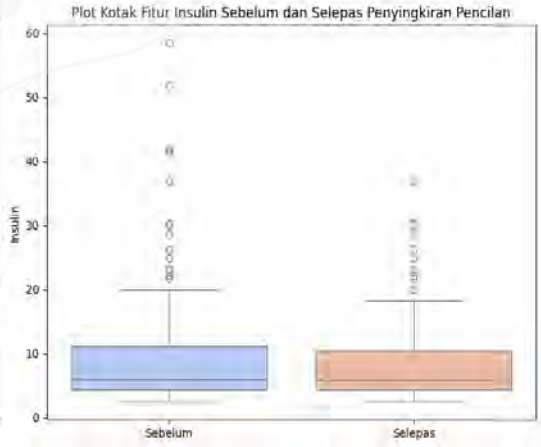


(c)

(d)



(e)



(f)

Rajah 4.3 Plot kotak sebelum dan selepas penyingkiran pencilan bagi fitur (a) *radius_mean* (b) *area_mean* (c) *symmetry1* (d) *fractal_dimension1* (e) *Glucose* (f) *Insulin*

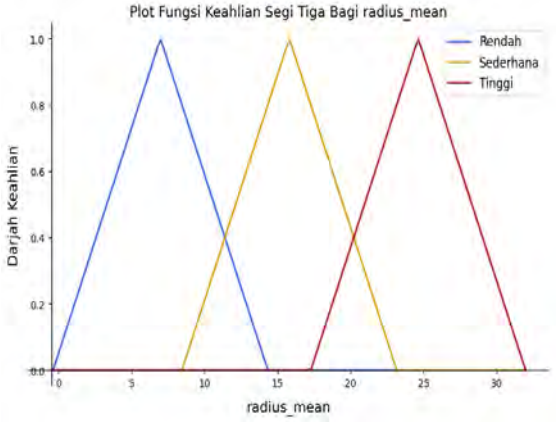
b. Pembahagian kabur berdasarkan fungsi keahlian segi tiga

Setiap fitur berangka dibahagikan kepada beberapa kelas kabur menggunakan fungsi keahlian segi tiga (*triangular membership function*). Bilangan pembahagian dan sempadan bagi setiap kelas ditentukan berdasarkan nilai minimum, maksimum dan purata bagi setiap fitur. Setiap fitur dibahagikan kepada tiga kategori: Rendah (*Low*), Sederhana (*Medium*) dan Tinggi (*High*) menggunakan fungsi keahlian segi tiga yang dibahagi secara sama rata.

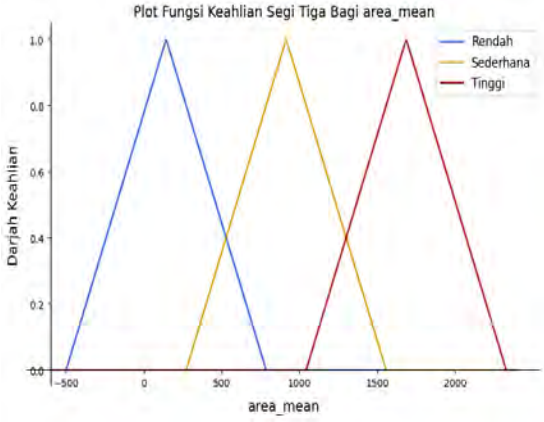
Jadual 4.9 menunjukkan contoh pembentukan set kabur bagi beberapa fitur berterusan dalam kajian ini. Sebagai contoh, fitur *radius_mean* dengan julat [6.98, 24.63] dibahagikan kepada tiga julat kabur berdasarkan fungsi keahlian segi tiga iaitu Rendah [-0.37 6.98 14.33], Sederhana [8.45 15.81 23.16] dan Tinggi [17.28 24.63 31.98]. Begitu juga, *area_mean* dengan julat [143.5, 1686.0] dibahagikan kepada Rendah [-499.21 143.5 786.21], Sederhana [272.04 914.75 1557.46] dan Tinggi [1043.29 1686 2328.71]. Proses ini dilaksanakan bagi setiap fitur berterusan dalam semua set data. Rajah 4.4 memaparkan fungsi keahlian segi tiga yang digunakan untuk menggambarkan julat kabur bagi fitur berkenaan dalam setiap set data.

Jadual 4.9 Set kabur bagi setiap contoh fitur

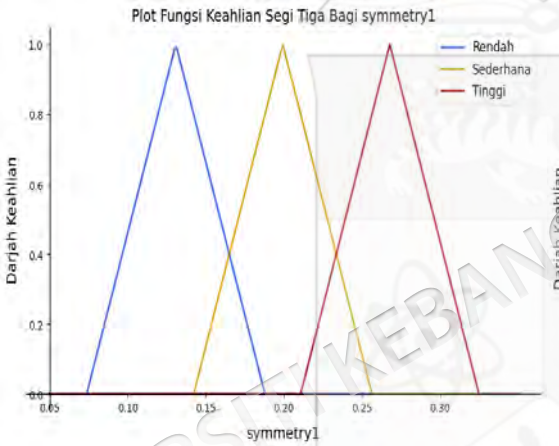
Set Data	Contoh Fitur	Julat Baharu	Set Kabur		
			Rendah	Sederhana	Tinggi
WDBC	*radius_mean*	[6.98, 24.63]	[-0.37 6.98 14.33]	[8.45 15.81 23.16]	[17.28 24.63 31.98]
	area_mean	[143.5, 1686]	[-499.21 143.5 786.21]	[272.04 914.75 1557.46]	[1043.29 1686 2328.71]
WPBC	*symmetry1*	[0.13, 0.27]	[0.07 0.13 0.19]	[0.14 0.20 0.26]	[0.21 0.27 0.32]
	fractal_dimension1	[0.05, 0.08]	[0.04 0.05 0.06]	[0.05 0.07 0.08]	[0.07 0.08 0.10]
BCC	*Glucose*	[60, 152]	[21.67 60 98.33]	[67.67 106 144.33]	[113.67 152 190.33]
	Insulin	[2.43, 36.94]	[-11.95 2.43 16.81]	[5.31 19.69 34.06]	[22.56 36.94 51.32]



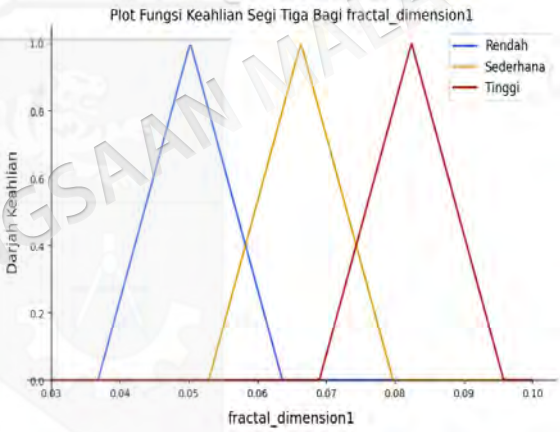
(a)



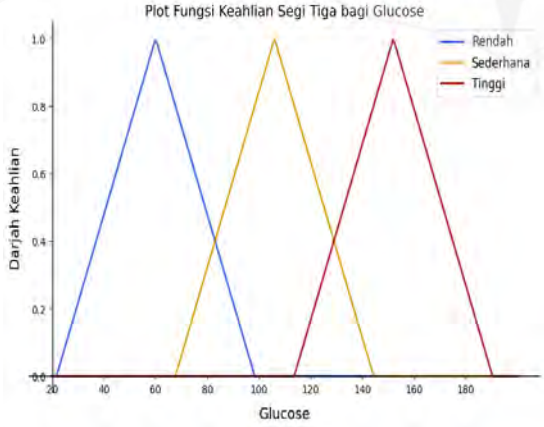
(b)



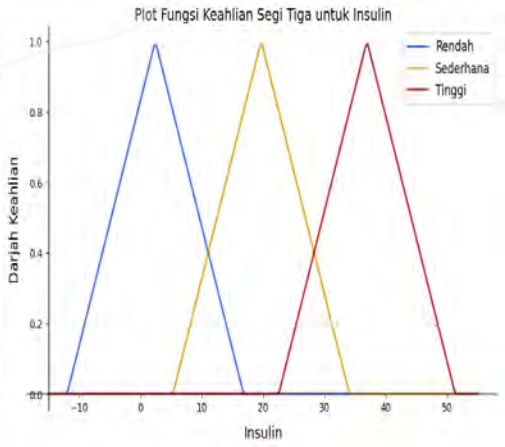
(c)



(d)



(e)



(f)

Rajah 4.4 Plot fungsi keahlian segi tiga bagi fitur (a) *radius_mean* (b) *area_mean* (c) *symmetry1* (d) *fractal_dimension1* (e) *Glucose* (f) *Insulin*

c. Penugasan nilai keahlian kepada setiap sampel

Proses pendiskretan kabur diteruskan dengan menugaskan setiap nilai fitur kepada kategori kabur berdasarkan nilai keahlian tertinggi antara kategori Rendah, Sederhana dan Tinggi. Jadual 4.10 menunjukkan contoh penugasan nilai kabur bagi satu sampel pertama daripada setiap set data. Sebagai contoh, nilai *radius_mean* dalam set data WDBC bernilai 17.99 mempunyai keahlian tertinggi dalam kategori Sederhana (0.7029). Fitur *area_mean* dengan nilai 1001.0 juga dikelaskan sebagai Sederhana (0.8658). Dalam set data WPBC, fitur *symmetry1* dan *fractal_dimension1* masing-masing juga mempunyai nilai keahlian tertinggi dalam kategori Sederhana iaitu 0.7758 dan 0.7755. Sementara itu, nilai sampel pertama bagi fitur *Glucose* dan *Insulin* juga dikelaskan sebagai Rendah kerana keahlian tertingginya masing-masing ialah 0.7391 dan 0.9809. Penugasan ini membolehkan transformasi fitur berterusan kepada fitur kabur kategori bagi tujuan perlombongan peraturan pada fasa seterusnya.

Jadual 4.10 Nilai kabur bagi setiap contoh fitur bagi sampel pertama

Set Data	Contoh Fitur	Nilai Fitur	Nilai Kabur			Penugasan Kategori Kabur
			Rendah	Sederhana	Tinggi	
WDBC	*radius_mean*	17.99	0.0	0.7029	0.0971	Sederhana
	area_mean	1001.0	0.0	0.8658	0.0	Sederhana
WPBC	*symmetry1*	0.1865	0.0242	0.7758	0.0	Sederhana
	fractal_dimension1	0.0633	0.0245	0.7755	0.0	Sederhana
BCC	*Glucose*	70	0.7391	0.0609	0.0	Rendah
	Insulin	2.707	0.9809	0.0	0.0	Rendah

d. Pelarasan sampel terpencil

Sampel yang berada di luar julat biasa atau tidak sesuai dengan sebarang kelas kabur disemak dan diselaraskan ke kelas yang paling hampir. Ini memastikan bahawa semua data dapat digunakan secara konsisten dalam proses perlombongan peraturan seterusnya. Jadual 4.11 menunjukkan bilangan pencilan yang dikesan dan bilangan sampel yang memerlukan pelarasan dalam setiap set data. Sebagai contoh, *area_mean* dalam WDBC mempunyai 8 pencilan dan 2 daripadanya diselaraskan. Bagi WPBC pula, *fractal_dimension1* mencatat 1 pelarasan manakala dalam BCC, *Glucose* dan *Insulin* masing-masing memerlukan 2 pelarasan. Fitur lain seperti *radius_mean* dan *symmetry1* tidak memerlukan pelarasan kerana nilainya masih dalam julat kabur.

Jadual 4.11 Bilangan pencilan dan sampel yang diselaraskan bagi setiap contoh fitur

Set Data	Contoh Fitur	Bilangan Pencilan	Bilangan Pelarasan Sampel
WDBC	*radius_mean*	5	0
	area_mean	8	2
WPBC	*symmetry1*	1	0
	fractal_dimension1	1	1
BCC	*Glucose*	3	2
	Insulin	4	2

Jadual 4.12 pula menunjukkan contoh pelarasan kategori kabur bagi beberapa nilai pencilan yang terletak di luar julat kabur asal dan telah ditugaskan kepada kategori Tinggi.

Jadual 4.12 Kategori kabur selepas pelarasan nilai pencilan

Set Data	Contoh Fitur	Nilai Pencilan	Pelarasan Kategori Kabur
WDBC	*area_mean*	2501	Tinggi
WPBC	*fractal_dimension1*	0.0974	Tinggi
BCC	*Glucose*	201	Tinggi
	Insulin	58.46	Tinggi

Pelarasan ini memastikan kesemua nilai berada dalam skop kategori kabur dan boleh dimasukkan dalam proses pengekstrakan peraturan seterusnya.

4.2.5 Pengekstrakan Subset Fitur Relevan dan Berinteraksi

Seksyen ini membentangkan hasil pengekstrakan fitur menggunakan CAR dan AAR yang merangkumi jumlah peraturan dijana dan bilangan fitur yang memenuhi kriteria relevan, berinteraksi dan tidak bertindih bagi setiap set data.

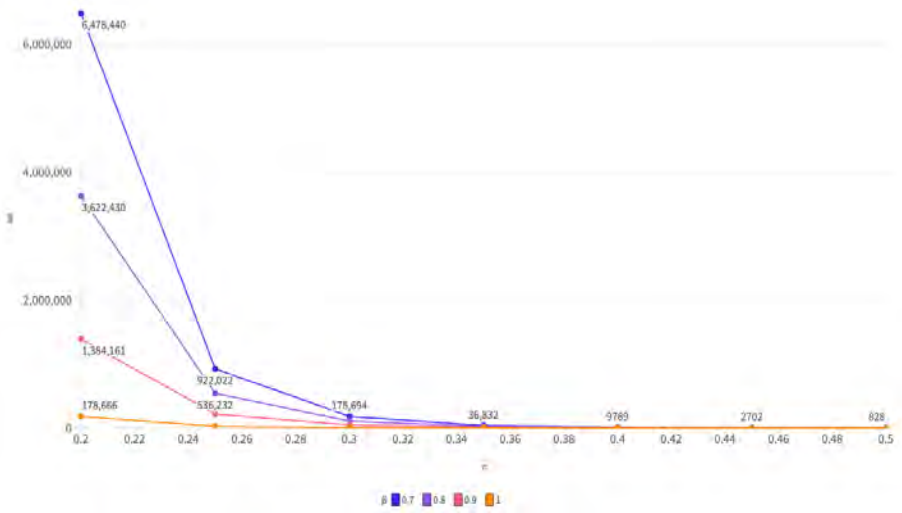
a. Pengekstrakan Peraturan Berdasarkan Kombinasi Parameter α dan β

Pengekstrakan fitur dalam kajian ini dilakukan berdasarkan bilangan peraturan yang dijana oleh kaedah CAR dengan pelbagai kombinasi nilai ambang *minSupp* (α) dan *minConf* (β). Pelbagai kombinasi α dan β telah diuji untuk menilai kesannya terhadap bilangan AR, CAR dan AAR. Jadual 4.13 menunjukkan bilangan peraturan yang dijana berdasarkan kombinasi α dan β bagi set data WDBC dan diterjemahkan kepada Rajah 4.5.

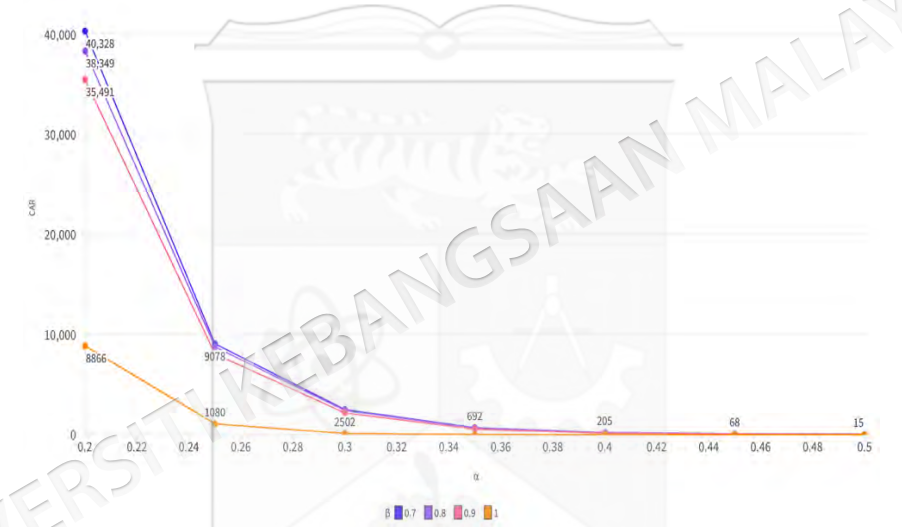
Hasilnya menunjukkan bahawa peningkatan nilai α dan β membawa kepada pengurangan bilangan peraturan, namun memberikan subset fitur yang lebih padat dan bermakna. Sebagai contoh, terdapat lebih 6.4 juta peraturan dijana pada α = 0.20 dan β = 0.70 namun jumlah ini turun kepada 828 apabila α meningkat ke 0.50. Corak yang sama dilihat pada CAR di mana ia berkurang daripada 40,328 kepada hanya 15 peraturan berkaitan kelas "Malignan" dan "Benigna". Bagi AAR pula, sebanyak 570 peraturan dijana pada α = 0.20, β = 0.70. Jumlah ini menurun kepada 156 apabila α dinaikkan dan tinggal hanya 1 peraturan apabila β = 1.00.

Jadual 4.13 Peraturan AR, CAR dan AAR bagi set data WDBC

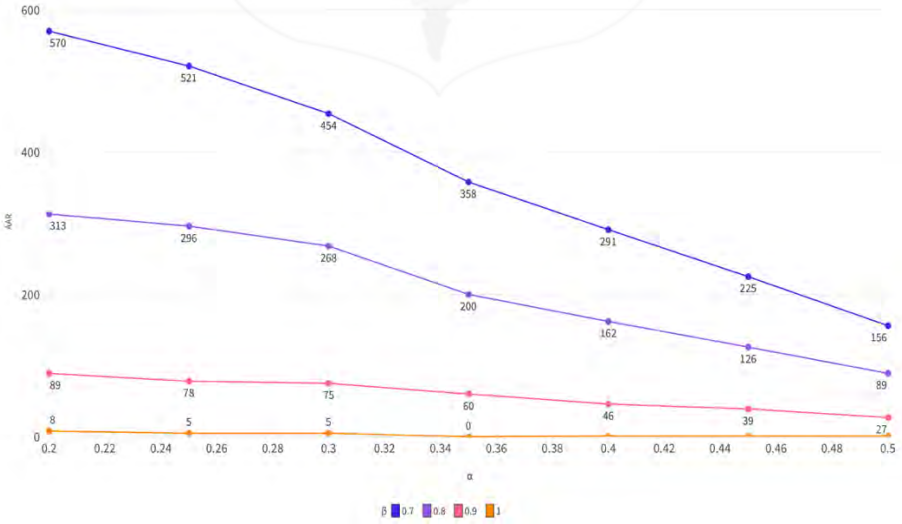
α	β	AR	CAR	AAR
0.20	0.70	6,478,440	40,328	570
	0.80	3,622,430	38,349	313
	0.90	1,384,161	35,491	89
	1.00	178,666	8,866	8
0.25	0.70	922,022	9,078	521
	0.80	536,232	8,772	296
	0.90	214,887	8,071	78
	1.00	24,236	1,080	5
0.30	0.70	178,694	2,502	454
	0.80	106,204	2,417	268
	0.90	44,852	2,171	75
	1.00	4,601	120	5
0.35	0.70	36,832	692	358
	0.80	22,198	660	200
	0.90	9,598	550	60
	1.00	939	16	0
0.40	0.70	9,789	205	291
	0.80	6,042	192	162
	0.90	2,721	147	46
	1.00	239	0	1
0.45	0.70	2,702	68	225
	0.80	1,709	63	126
	0.90	780	44	39
	1.00	64	0	1
0.50	0.70	828	15	156
	0.80	522	13	89
	0.90	241	6	27
	1.00	16	0	1



(a)



(b)



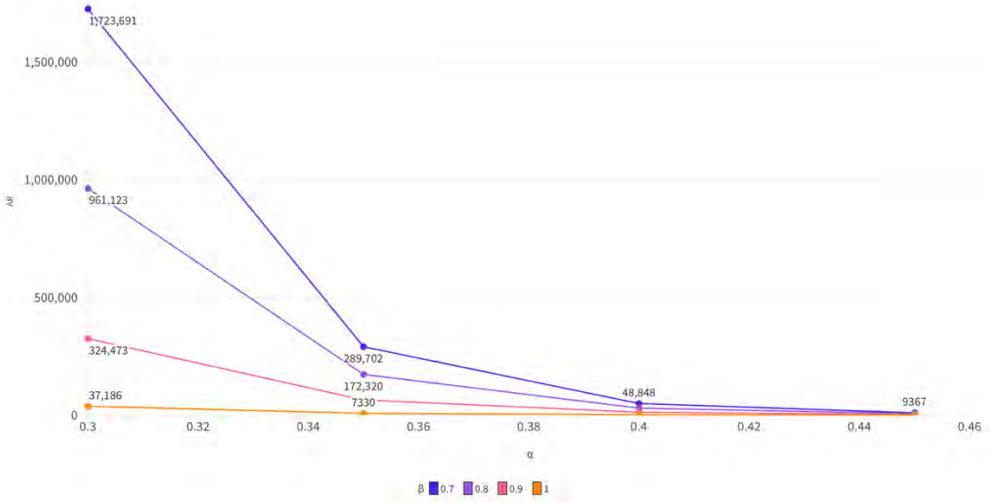
(c)

Rajah 4.5 Bilangan (a) AR (b) CAR (c) AAR berdasarkan α dan β pada WDBC

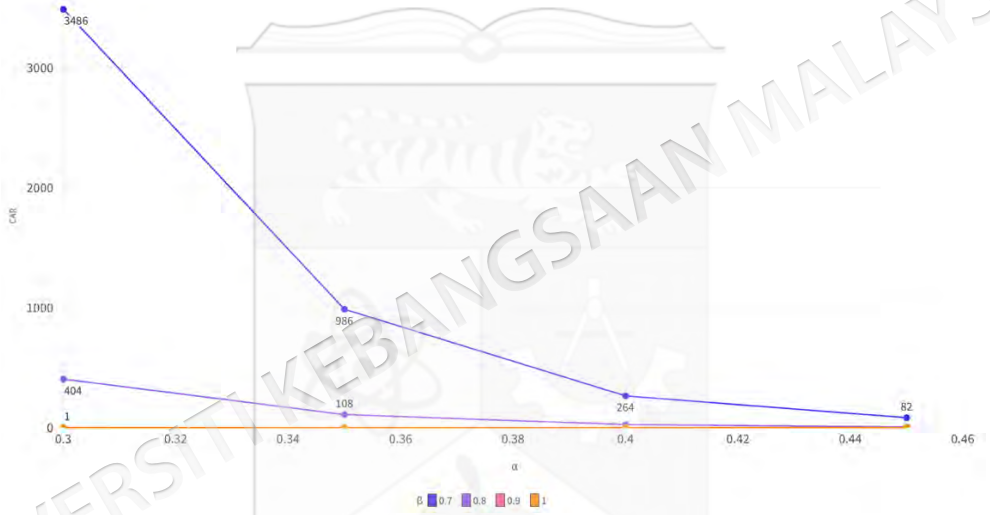
Perkara sama dapat dilihat dalam set data WPBC (rujuk Jadual 4.14 dan Rajah 4.6) dan set data BCC (rujuk Jadual 4.15 dan Rajah 4.7).

Jadual 4.14 Peraturan AR, CAR dan AAR bagi set data WPBC

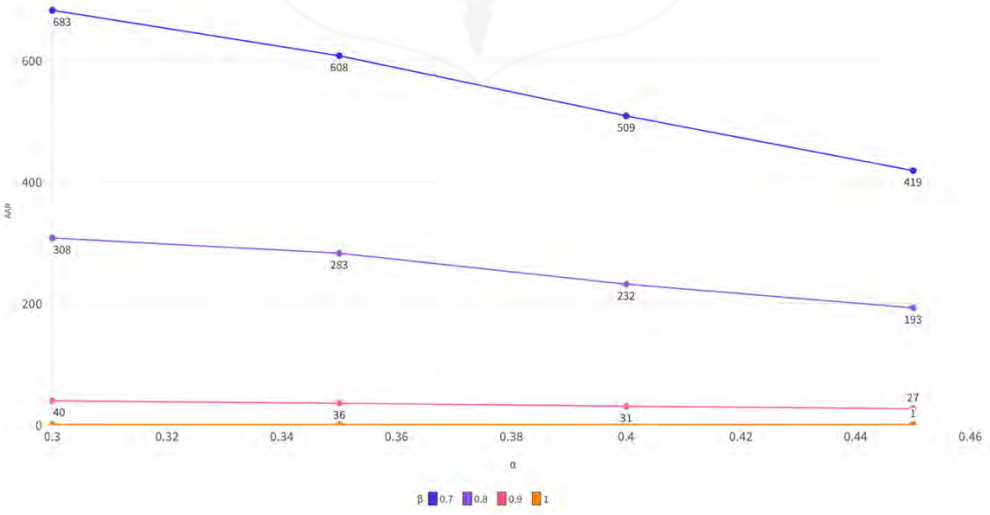
α	β	AR	CAR	AAR
0.30	0.70	1,723,691	3,486	683
	0.80	961,123	404	308
	0.90	324,473	1	40
	1.00	37,186	0	1
0.35	0.70	289,702	986	608
	0.80	172,320	108	283
	0.90	63,908	0	36
	1.00	7,330	0	1
0.40	0.70	48,848	264	509
	0.80	29,150	25	232
	0.90	11,860	0	31
	1.00	1,141	0	1
0.45	0.70	9,367	82	419
	0.80	5,938	5	193
	0.90	2,326	0	27
	1.00	161	0	1



(a)



(b)

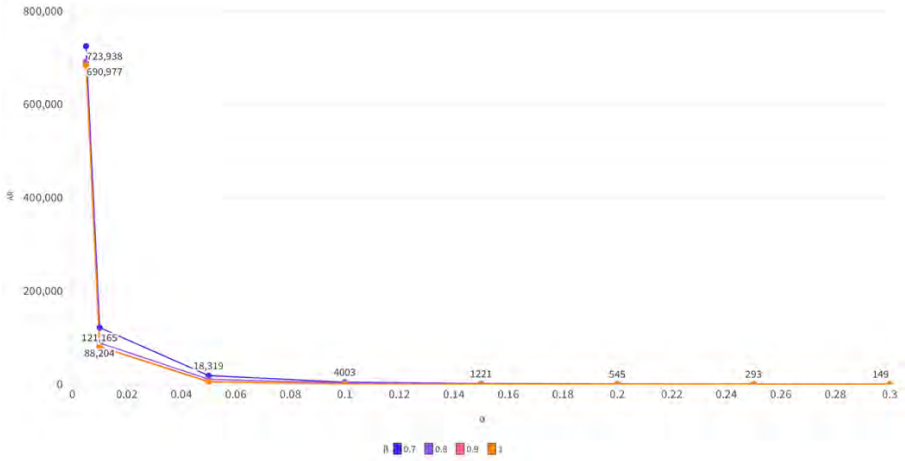


(c)

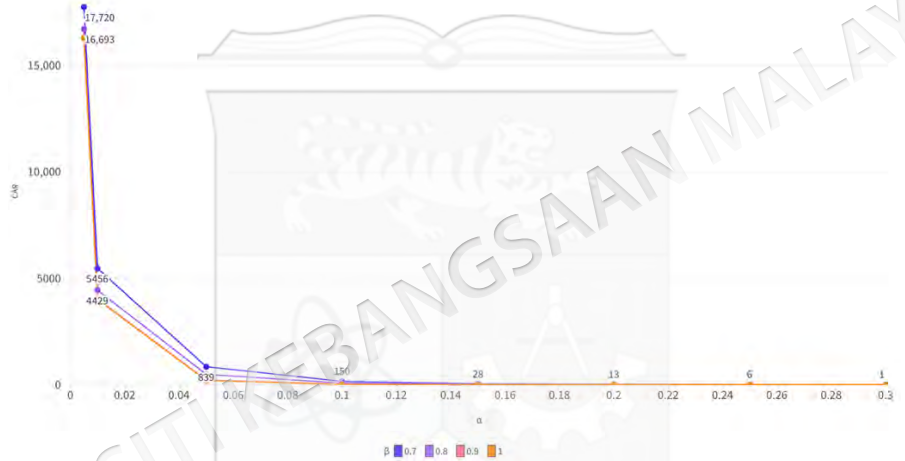
Rajah 4.6 Bilangan (a) AR (b) CAR (c) AAR berdasarkan α dan β pada WPBC

Jadual 4.15 Peraturan AR, CAR dan AAR bagi set data BCC

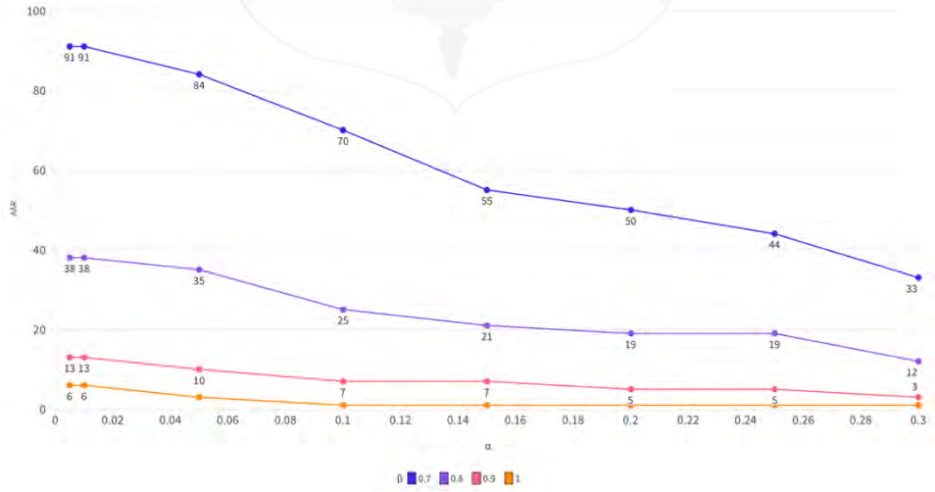
α	β	AR	CAR	AAR
0.005	0.70	723,938	17,720	91
	0.80	690,977	16,693	38
	0.90	683,463	16,284	13
	1.00	682,415	16,260	6
0.01	0.70	121,165	5,456	91
	0.80	88,204	4,429	38
	0.90	80,690	4,020	13
	1.00	79,642	3,996	6
0.05	0.70	18,319	839	84
	0.80	10,368	480	35
	0.90	5,669	214	10
	1.00	4,621	190	3
0.10	0.70	4,003	150	70
	0.80	2,221	69	25
	0.90	1,142	16	7
	1.00	617	7	1
0.15	0.70	1,221	28	55
	0.80	670	8	21
	0.90	351	0	7
	1.00	149	0	1
0.20	0.70	545	13	50
	0.80	287	2	19
	0.90	143	0	5
	1.00	67	0	1
0.25	0.70	293	6	44
	0.80	145	0	19
	0.90	74	0	5
	1.00	36	0	1
0.30	0.70	149	1	33
	0.80	76	0	12
	0.90	40	0	3
	1.00	21	0	1



(a)



(b)



(c)

Rajah 4.7 Bilangan (a) AR (b) CAR (c) AAR berdasarkan α dan β pada BCC

Walaupun penghasilan AR dianalisis bagi pelbagai kombinasi α dan β untuk memahami corak peraturan secara keseluruhan, hanya CAR dan AAR digunakan dalam langkah seterusnya iaitu proses pemangkasan peraturan. Ini kerana hanya CAR mengandungi fitur yang relevan dan berinteraksi dengan kelas sasaran, manakala AAR digunakan untuk mengenalpasti fitur yang tidak bertindih selaras dengan objektif pemilihan fitur dalam kajian ini.

b. Pemangkasan Peraturan

Pemangkasan peraturan dalam kajian ini dijalankan bagi menyaring CAR dan AAR yang telah dijana dengan tujuan mengekalkan hanya peraturan yang berkualiti tinggi dan menyumbang secara langsung kepada pemilihan fitur yang relevan, berinteraksi dan tidak bertindih. Hasil daripada proses ini menunjukkan bahawa bilangan peraturan dapat dikurangkan secara signifikan tanpa menjejaskan maklumat penting yang diperlukan untuk langkah seterusnya. Penapisan dilakukan menggunakan kombinasi ambang metrik seperti *minLift*, *minConv*, *minCF* dan *minLev* yang ditentukan secara berasingan bagi setiap set data berdasarkan taburan nilai peraturan.

Jadual 4.16 menunjukkan hasil pemangkasan CAR (CARpr) dan bilangan subset fitur (RFset) yang diperoleh bagi set data WDBC berdasarkan pelbagai nilai ambang α dan β serta empat ambang tambahan tersebut. Apabila nilai α rendah (0.20 – 0.30), bilangan peraturan dan fitur dalam yang dipilih adalah tinggi. Sebagai conto pada α = 0.20 dan 0.25, sebanyak 1,785 CARpr dikekalkan dan 22 fitur terpilih termasuk seperti *area_mean*, *concavity_mean* dan *perimeter_worst*. Namun, peningkatan nilai α menyebabkan penurunan ketara dalam bilangan peraturan dan fitur di mana pada α= 0.35 hanya tinggal 98 peraturan dengan 12 fitur, α = 0.45, hanya tinggal 2 peraturan dengan 4 fitur. Tiada peraturan atau fitur kekal pada α mencapai 0.50.

Jadual 4.16 Bilangan peraturan yang dipangkas dan senarai subset fitur berdasarkan α dan β bagi set data WDBC

α	β	CARpr	Bilangan RFset	Senarai RFset
0.20	0.70	1,785	22	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'perimeter_mean=Medium', 'texture_mean=Medium', 'radius_mean=Medium']*
	0.80	1,785	22	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'perimeter_mean=Medium', 'texture_mean=Medium', 'radius_mean=Medium']*
	0.90	1,785	22	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'perimeter_mean=Medium', 'texture_mean=Medium', 'radius_mean=Medium']*
	1.00	486	18	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low']*
0.25	0.70	1,785	22	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'perimeter_mean=Medium', 'texture_mean=Medium', 'radius_mean=Medium']*

bersambung…

...sambungan

	0.80	1,785	22	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'perimeter_mean=Medium', 'texture_mean=Medium', 'radius_mean=Medium']*
	0.90	1,785	22	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'perimeter_mean=Medium', 'texture_mean=Medium', 'radius_mean=Medium']*
	1.00	486	18	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'area_worst=Low', 'concave points_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low']*
0.30	0.70	640	19	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'texture_mean=Medium']*
	0.80	640	19	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'texture_mean=Medium']*
	0.90	640	19	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'texture_worst=Medium', 'area_worst=Low', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'area_se=Low', 'texture_mean=Medium']*

bersambung…

...sambungan

	1.00	120	14	*['area_mean=Low', 'area_se=Low', 'area_worst=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'concave points_mean=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'compactness_se=Low', 'radius_se=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium']*
0.35	0.70	98	12	*['area_mean=Low', 'area_se=Low', 'area_worst=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'concave points_mean=Low', 'perimeter_se=Low', 'compactness_worst=Low', 'radius_se=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium']*
	0.80	98	12	*['area_mean=Low', 'area_se=Low', 'area_worst=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'concave points_mean=Low', 'perimeter_se=Low', 'compactness_worst=Low', 'radius_se=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium']*
	0.90	98	12	*['area_mean=Low', 'area_se=Low', 'area_worst=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'concave points_mean=Low', 'perimeter_se=Low', 'compactness_worst=Low', 'radius_se=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium']*
	1.00	16	7	*['concavity_worst=Low', 'concave points_mean=Low', 'perimeter_se=Low', 'area_worst=Low', 'radius_se=Low', 'concavity_mean=Low', 'area_se=Low']*
0.40	0.70	22	7	*['concave points_mean=Low', 'concavity_worst=Low', 'perimeter_se=Low', 'area_worst=Low', 'radius_se=Low', 'concavity_mean=Low', 'area_se=Low']*
	0.80	22	7	*['concave points_mean=Low', 'concavity_worst=Low', 'perimeter_se=Low', 'area_worst=Low', 'radius_se=Low', 'concavity_mean=Low', 'area_se=Low']*
	0.90	22	7	*['concave points_mean=Low', 'concavity_worst=Low', 'perimeter_se=Low', 'area_worst=Low', 'radius_se=Low', 'concavity_mean=Low', 'area_se=Low']*
	1.00	0	0	[]
0.45	0.70	2	4	*['area_se=Low', 'concave points_mean=Low', 'area_worst=Low', 'concavity_mean=Low']*
	0.80	2	4	*['area_se=Low', 'concave points_mean=Low', 'area_worst=Low', 'concavity_mean=Low']*
	0.90	2	4	*['area_se=Low', 'concave points_mean=Low', 'area_worst=Low', 'concavity_mean=Low']*
	1.00	0	0	[]

Jadual 4.17 pula menunjukkan hasil untuk set data WPBC di mana pemangkasan pada $\alpha = 0.30$ mengekalkan sebanyak 59 peraturan dan 20 fitur termasuk *area2, radius3, perimeter3, texture1, concavity3* serta fitur klinikal seperti *tumor_size*, *lymph_node_status* dan *Time*. Apabila β dinaikkan ke 0.90, hanya 1 peraturan dan 3 fitur dikekalkan. Pada $\beta = 1.00$ atau $\alpha = 0.45$, tiada peraturan atau fitur dikekalkan.

Jadual 4.17 Bilangan peraturan yang dipangkas dan senarai subset fitur berdasarkan α dan β bagi set data WPBC

α	β	CARpr	Bilangan RFset	Senarai RFset
0.30	0.70	59	20	*['area2=Low', 'concave_points3=Medium', 'perimeter3=Medium', 'area3=Low', 'compactness1=Medium', 'texture3=Medium', 'tumor_size=Small', 'radius2=Low', 'lymph_node_status=Low', 'perimeter1=Medium', 'symmetry3=Medium', 'compactness2=Medium', 'radius3=Medium', 'compactness3=Medium', 'texture1=Medium', 'radius1=Medium', 'concave_points1=Medium', 'concavity3=Medium', 'smoothness3=Medium', 'Time=Medium']*
	0.80	59	20	*['area2=Low', 'concave_points3=Medium', 'perimeter3=Medium', 'area3=Low', 'compactness1=Medium', 'texture3=Medium', 'tumor_size=Small', 'radius2=Low', 'lymph_node_status=Low', 'perimeter1=Medium', 'symmetry3=Medium', 'compactness2=Medium', 'radius3=Medium', 'compactness3=Medium', 'texture1=Medium', 'radius1=Medium', 'concave_points1=Medium', 'concavity3=Medium', 'smoothness3=Medium', 'Time=Medium']*
	0.90	1	3	*['concavity3=Medium', 'tumor_size=Small', 'lymph_node_status=Low']*
	1.00	0	0	[]
0.35	0.70	10	9	*['area2=Low', 'radius3=Medium', 'perimeter3=Medium', 'texture1=Medium', 'radius1=Medium', 'concavity3=Medium', 'lymph_node_status=Low', 'smoothness3=Medium', 'tumor_size=Small']*
	0.80	10	9	*['area2=Low', 'radius3=Medium', 'perimeter3=Medium', 'texture1=Medium', 'radius1=Medium', 'concavity3=Medium', 'lymph_node_status=Low', 'smoothness3=Medium', 'tumor_size=Small']*
	0.90	0	0	[]
	1.00	0	0	[]

Jadual 4.18 menunjukkan hasil pemangkasan bagi set data BCC. Pada nilai α yang rendah (0.005 – 0.05), sebanyak 25 peraturan dan 11 fitur dikekalkan termasuk *Age, Resistin, HOMA, Insulin, MCP.1, Glucose* dan *Leptin*. Namun, apabila α meningkat ke 0.10, peraturan yang dipangkas menurun kepada 7 dan hanya 7 fitur dikekalkan Bagi $\alpha \geq 0.15$, tiada CARpr dipilih maka tiada fitur dikekalkan.

Jadual 4.18 Bilangan peraturan yang dipangkas dan senarai subset fitur berdasarkan α dan β bagi set data BCC

α	β	CARpr	Bilangan RFset	Senarai RFset
0.005	0.70	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	0.80	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	0.90	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	1.00	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
0.01	0.70	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	0.80	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	0.90	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	1.00	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
0.05	0.70	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	0.80	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	0.90	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
	1.00	25	11	*['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'HOMA=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']*
0.10	0.70	7	7	*['Age=Middle', 'Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'HOMA=Low', 'Leptin=Low']*
	0.80	7	7	*['Age=Middle', 'Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'HOMA=Low', 'Leptin=Low']*
	0.90	7	7	*['Age=Middle', 'Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'HOMA=Low', 'Leptin=Low']*
	1.00	7	7	*['Age=Middle', 'Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'HOMA=Low', 'Leptin=Low']*

c. Penyingkiran Fitur Bertindih

Penyingkiran fitur bertindih merupakan langkah penting untuk memastikan tiada tindihan maklumat yang sama daripada fitur yang berlainan. Jadual 4.19 menunjukkan jumlah RFset sebelum dan selepas proses penyingkiran fitur bertindih bagi set data WDBC.

Jadual 4.19 Bilangan RFset sebelum dan selepas penyingkiran fitur bertindih bagi set data WDBC

α	β	AAR mengandungi RFset	Bilangan RFset Sebelum Penyingkiran	Selepas Penyingkiran	
				Bilangan RFset	Senarai RFset
0.20	0.70	264	22	2	*['concave points_worst=Low', 'perimeter_mean=Medium']*
	0.80	141	22	4	*['texture_worst=Medium', 'concave points_worst=Low', 'radius_worst=Low', 'perimeter_mean=Medium']*
	0.90	45	22	9	*['symmetry_worst=Medium', 'radius_worst=Low', 'texture_worst=Medium', 'concave points_worst=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'perimeter_mean=Medium']*
	1.00	5	18	14	*['concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'concave points_worst=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium']*
0.25	0.70	258	22	3	*['concave points_worst=Low', 'radius_worst=Low', 'perimeter_mean=Medium']*
	0.80	141	22	4	*['texture_worst=Medium', 'concave points_worst=Low', 'radius_worst=Low', 'perimeter_mean=Medium']*
	0.90	45	22	9	*['symmetry_worst=Medium', 'radius_worst=Low', 'texture_worst=Medium', 'concave points_worst=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium', 'perimeter_mean=Medium']*
	1.00	5	18	14	*['concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'concave points_worst=Low', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium']*
0.30	0.70	198	19	3	*['radius_worst=Low', 'texture_worst=Medium', 'compactness_mean=Low']*
	0.80	120	19	3	*['radius_worst=Low', 'texture_worst=Medium', 'compactness_mean=Low']*

bersambung…

...sambungan

	0.90	37	19	9	*['symmetry_worst=Medium', 'concavity_worst=Low', 'radius_worst=Low', 'compactness_worst=Low', 'texture_worst=Medium', 'compactness_se=Low', 'radius_se=Low', 'compactness_mean=Low', 'smoothness_mean=Medium']*
	1.00	2	14	12	*['area_mean=Low', 'concavity_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'perimeter_se=Low', 'perimeter_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'symmetry_mean=Medium', 'concave points_mean=Low', 'compactness_se=Low', 'radius_se=Low']*
0.35	0.70	90	12	2	*['area_mean=Low', 'compactness_worst=Low']*
	0.80	66	12	3	*['area_mean=Low', 'symmetry_worst=Medium', 'compactness_worst=Low']*
	0.90	18	12	7	*['area_mean=Low', 'symmetry_worst=Medium', 'concavity_worst=Low', 'compactness_worst=Low', 'smoothness_worst=Medium', 'concave points_mean=Low', 'radius_se=Low']*
	1.00	1	7	6	*['area_worst=Low', 'concavity_mean=Low', 'concavity_worst=Low', 'concave points_mean=Low', 'perimeter_se=Low', 'radius_se=Low']*
0.40	0.70	38	7	2	*['concavity_worst=Low', 'radius_se=Low']*
	0.80	29	7	2	*['concavity_worst=Low', 'radius_se=Low']*
	0.90	13	7	3	*['concave points_mean=Low', 'concavity_worst=Low', 'radius_se=Low']*
	1.00	0	0	0	[]
0.45	0.70	12	4	1	*['concave points_mean=Low']*
	0.80	9	4	1	*['concave points_mean=Low']*
	0.90	5	4	1	*['concave points_mean=Low']*
	1.00	0	0	0	[]

Jadual 4.19 menunjukkan pada $\alpha = 0.20$ dan $\beta = 0.70$, terdapat 264 peraturan AAR yang mengandungi RFset. Bilangan fitur dalam RFset sebelum penyingkiran iaitu 22 dikurangkan kepada 2 fitur iaitu *concave points_worst* dan *perimeter_mean*. Apabila β meningkat, bilangan RFset yang kekal selepas penyingkiran juga meningkat kerana hanya peraturan AAR yang lebih kuat dikekalkan. Contohnya, pada $\beta = 0.90$, bilangan RFset selepas penyingkiran naik kepada 9 fitur dan 12 fitur pada $\beta = 1.0$. Apabila $\alpha = 0.30$, $\beta = 1.00$ pula, hanya 2 peraturan AAR dihasilkan dan 12 fitur dikekalkan. Namun, apabila α meningkat kepada 0.50, tiada fitur dalam RFset maka tiada juga fitur yang bertindih.

Seterusnya, Jadual 4.20 menunjukkan bilangan RFset sebelum dan selepas penyingkiran fitur bertindih berdasarkan pelbagai nilai α dan β pada set data WPBC.

Jadual 4.20 Bilangan RFset sebelum dan selepas penyingkiran fitur bertindih bagi set data WPBC

α	β	AAR mengandungi RFset	Bilangan RFset Sebelum Penyingkiran	Selepas Penyingkiran	
				Bilangan RFset	Senarai RFset
0.30	0.70	210	20	5	*['area3=Low', 'tumor_size=Small', 'compactness2=Medium', 'radius3=Medium', 'Time=Medium']*
	0.80	105	20	6	*['area3=Low', 'tumor_size=Small', 'compactness2=Medium', 'radius3=Medium', 'concave_points1=Medium', 'Time=Medium']*
	0.90	1	3	2	*['concavity3=Medium', 'tumor_size=Small']*
	1.00	0	0	0	[]
0.35	0.70	42	9	3	*['area2=Low', 'radius3=Medium', 'tumor_size=Small']*
	0.80	30	9	3	*['area2=Low', 'radius3=Medium', 'tumor_size=Small']*
	0.90	0	0	0	[]
	1.00	0	0	0	[]

Jadual 4.20 menunjukkan pada $\alpha = 0.30$ dan $\beta = 0.70$, terdapat 210 peraturan AAR yang mengandungi RFset. Sebanyak 20 fitur dalam RFset sebelum penyingkiran telah dikurangkan kepada 5 fitur sahaja iaitu *area3*, *tumor_size*, *compactness2*, *radius3* dan *Time*. Apabila β meningkat kepada 0.80, bilangan AAR berkurang kepada 105 dan bilangan RFset selepas penyingkiran meningkat kepada 6 fitur. Namun, pada $\beta = 0.90$, hanya 2 fitur kekal iaitu *concavity3* dan *tumor_size* dan tiada fitur dikekalkan apabila β mencapai 1.00. Apabila α meningkat kepada 0.35 dan $\beta = 0.70$, jumlah peraturan AAR berkurang kepada 42 dan RFset sebelum dan selepas penyingkiran mengandungi 9 dan 3 fitur masing-masing. Bilangan fitur kekal tidak berubah apabila β meningkat kepada 0.80 namun apabila nilai α mencapai 0.40, tiada peraturan AAR dijana maka tiada juga fitur dalam RFset. Keadaan ini berlaku kerana tiada CARpr yang wujud pada nilai ambang ini.

Seterusnya, Jadual 4.21 menunjukkan bilangan RFset sebelum dan selepas penyingkiran fitur bertindih berdasarkan pelbagai nilai α dan β pada set data BCC.

Jadual 4.21 Bilangan RFset sebelum dan selepas penyingkiran fitur bertindih bagi set data BCC

α	β	AAR mengandungi RFset	Bilangan RFset Sebelum Penyingkiran	Selepas Penyingkiran	
				Bilangan RFset	Senarai RFset
0.005	0.70	31	11	6	['Age=Old', 'Resistin=Medium', 'Age=Young', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	0.80	13	11	9	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	0.90	4	11	9	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	1.00	1	11	10	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
0.01	0.70	31	11	6	['Age=Old', 'Resistin=Medium', 'Age=Young', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	0.80	13	11	9	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	0.90	4	11	9	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	1.00	1	11	10	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
0.05	0.70	31	11	6	['Age=Old', 'Resistin=Medium', 'Age=Young', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	0.80	13	11	9	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	0.90	4	11	9	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
	1.00	1	11	10	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'Age=Young', 'MCP.1=Medium', 'Glucose=Medium', 'Glucose=Low', 'Age=Middle', 'Leptin=Low']
0.10	0.70	14	7	4	['Age=Old', 'Resistin=Medium', 'Age=Middle', 'Leptin=Low']
	0.80	7	7	5	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Age=Middle', 'Leptin=Low']
	0.90	2	7	6	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'Age=Middle', 'Leptin=Low']
	1.00	1	7	6	['Age=Old', 'Resistin=Medium', 'Resistin=Low', 'Insulin=Low', 'Age=Middle', 'Leptin=Low']

Jadual 4.21 menunjukkan pada α = 0.005 dan β = 0.70, terdapat 31 AAR dengan 11 fitur dalam RF sebelum penyingkiran. Selepas penyingkiran, hanya 6 fitur yang dikekalkan. Apabila β meningkat ke 0.80, bilangan AAR berkurang kepada 13 dan bilangan RFset meningkat kepada 9 fitur. Pada β = 0.10, hanya 1 AAR dan RFset meningkat pada 10 fitur selepas fitur bertindih disingkir. Apabila α meningkat ke 0.10 dan β = 0.70, terdapat 14 AAR dengan 7 fitur dalam RFset sebelum penyingkiran dan hanya 4 fitur yang dikekalkan. Apabila α meningkat ke 0.15 dan lebih tinggi, tiada AAR atau RFset kerana tiada CARpr yang wujud pada nilai ambang ini.

d. Pemilihan Subset Fitur Selepas Penyingkiran Fitur Bertindih

Subset fitur unik ini merujuk kepada fitur tanpa nilai spesifiknya yang dianggap berinteraksi merujuk kepada persamaan ...(2.6). Jadual 4.22 menunjukkan subset fitur unik yang diperoleh selepas proses penyingkiran fitur bertindih bagi set data WDBC. Hasil menunjukkan bahawa semua fitur yang dipilih adalah unik, tanpa sebarang pertindihan dari segi jenis atau nilai fitur. Maka, tiada pengurangan bilangan fitur berlaku selepas proses ini.

Jadual 4.22 Subset fitur unik berdasarkan α dan β bagi set data WDBC

α	β	Bilangan RFset awal	Subset Fitur Unik	
			Bilangan RFset akhir	Senarai RFset
0.20	0.70	2	2	['concave points_worst', 'perimeter_mean']
	0.80	4	4	['texture_worst', 'concave points_worst', 'radius_worst', 'perimeter_mean']
	0.90	9	9	['symmetry_worst', 'radius_worst', 'texture_worst', 'concave points_worst', 'compactness_se', 'radius_se', 'compactness_mean', 'smoothness_mean', 'perimeter_mean']
	1.00	14	14	['concavity_mean', 'symmetry_worst', 'concavity_worst', 'radius_worst', 'perimeter_se', 'perimeter_worst', 'compactness_worst', 'smoothness_worst', 'symmetry_mean', 'concave points_worst', 'compactness_se', 'radius_se', 'compactness_mean', 'smoothness_mean']
0.25	0.70	3	3	['concave points_worst', 'radius_worst', 'perimeter_mean']
	0.80	4	4	['texture_worst', 'concave points_worst', 'radius_worst', 'perimeter_mean']
	0.90	9	9	['symmetry_worst', 'radius_worst', 'texture_worst', 'concave points_worst', 'compactness_se', 'radius_se', 'compactness_mean', 'smoothness_mean', 'perimeter_mean']

bersambung…

...sambungan

	1.00	14	14	*['concavity_mean', 'symmetry_worst', 'concavity_worst', 'radius_worst', 'perimeter_se', 'perimeter_worst', 'compactness_worst', 'smoothness_worst', 'symmetry_mean', 'concave points_worst', 'compactness_se', 'radius_se', 'compactness_mean', 'smoothness_mean']*
0.30	0.70	3	3	*['radius_worst', 'texture_worst', 'compactness_mean']*
	0.80	3	3	*['radius_worst', 'texture_worst', 'compactness_mean']*
	0.90	9	9	*['symmetry_worst', 'concavity_worst', 'radius_worst', 'compactness_worst', 'texture_worst', 'compactness_se', 'radius_se', 'compactness_mean', 'smoothness_mean']*
	1.00	12	12	*['area_mean', 'concavity_mean', 'symmetry_worst', 'concavity_worst', 'perimeter_se', 'perimeter_worst', 'compactness_worst', 'smoothness_worst', 'symmetry_mean', 'concave points_mean', 'compactness_se', 'radius_se']*
0.35	0.70	2	2	*['area_mean', 'compactness_worst']*
	0.80	3	3	*['area_mean', 'symmetry_worst', 'compactness_worst']*
	0.90	7	7	*['area_mean', 'symmetry_worst', 'concavity_worst', 'compactness_worst', 'smoothness_worst', 'concave points_mean', 'radius_se']*
	1.00	6	6	*['area_worst', 'concavity_mean', 'concavity_worst', 'concave points_mean', 'perimeter_se', 'radius_se']*
0.40	0.70	2	2	*['concavity_worst', 'radius_se']*
	0.80	2	2	*['concavity_worst', 'radius_se']*
	0.90	3	3	*['concave points_mean', 'concavity_worst', 'radius_se']*
0.45	0.70	1	1	*['concave points_mean']*
	0.80	1	1	*['concave points_mean']*
	0.90	1	1	*['concave points_mean']*

Seterusnya, Jadual 4.23 mempersembahkan subset fitur unik selepas proses penyingkiran fitur bertindih bagi set data WPBC. Hasil menunjukkan bahawa tiada perubahan dalam bilangan fitur sebelum dan selepas proses tersebut yang menandakan bahawa semua fitur yang dikenal pasti adalah unik dan tidak bertindih.

Jadual 4.23 Subset fitur unik berdasarkan α dan β bagi set data WPBC

α	β	**Bilangan RFset awal**	**Subset Fitur Unik**	
			Bilangan RFset akhir	**Senarai RFset**
0.30	0.70	5	5	*['area3', 'tumor_size', 'compactness2', 'radius3', 'Time']*
	0.80	6	6	*['area3', 'tumor_size', 'compactness2', 'radius3', 'concave_points1', 'Time']*
	0.90	2	2	*['concavity3', 'tumor_size']*
0.35	0.70	3	3	*['area2', 'radius3', 'tumor_size']*
	0.80	3	3	*['area2', 'radius3', 'tumor_size']*

Jadual 4.24 menunjukkan subset fitur unik selepas proses penyingkiran fitur bertindih bagi set data BCC. Berbeza dengan set data sebelumnya, terdapat pengurangan bilangan fitur selepas proses ini disebabkan oleh kewujudan nilai fitur yang serupa. Sebagai contoh, berdasarkan Jadual 4.21, sebanyak 11 fitur dikenal pasti dalam RFset awal pada kombinasi α = 0.005 dan β = 0.70, namun hanya 4 fitur yang kekal selepas penyingkiran kerana terdapat tiga variasi nilai pada fitur *Age* iaitu "*Young*", "*Middle*" dan "*Old*". Begitu juga, apabila β meningkat kepada 0.80, jumlah fitur dalam RFset berkurang daripada 9 menjadi 5 akibat tiga variasi nilai pada *Age* dan 2 variasi nilai "*Low*" dan "*Medium*" pada kedua-dua fitur *Resistin* dan *Glucose* . Proses ini diteruskan untuk semua kombinasi α dan β bagi mendapatkan subset fitur unik.

Jadual 4.24 Subset fitur unik berdasarkan α dan β bagi set data BCC

α	β	Bilangan RF_{set} awal	Subset Fitur Unik Tanpa Nilai Fitur	
			Bilangan RF_{set}	Senarai RF_{set}
0.005	0.70	6	4	*['Age', 'Glucose', 'Leptin', 'Resistin']*
	0.80	9	5	*['Age', 'Glucose', 'Leptin', 'MCP.1', 'Resistin']*
	0.90	9	5	*['Age', 'Glucose', 'Leptin', 'MCP.1', 'Resistin']*
	1.00	10	6	*['Age', 'Glucose', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']*
0.01	0.70	6	4	*['Age', 'Resistin', 'Glucose', 'Leptin']*
	0.80	9	5	*['Age', 'Resistin', 'MCP.1', 'Glucose', 'Leptin']*
	0.90	9	5	*['Age', 'Resistin', 'MCP.1', 'Glucose', 'Leptin']*
	1.00	10	6	*['Age', 'Resistin', 'Insulin', 'MCP.1', 'Glucose', 'Leptin']*
0.05	0.70	6	4	*['Age', 'Resistin', 'Glucose', 'Leptin']*
	0.80	9	5	*['Age', 'Resistin', 'MCP.1', 'Glucose', 'Leptin']*
	0.90	9	5	*['Age', 'Resistin', 'MCP.1', 'Glucose', 'Leptin']*
	1.00	10	6	*['Age', 'Resistin', 'Insulin', 'MCP.1', 'Glucose', 'Leptin']*
0.10	0.70	4	3	*['Age', 'Resistin', 'Leptin']*
	0.80	5	3	*['Age', 'Resistin', 'Leptin']*
	0.90	6	4	*['Age', 'Resistin', 'Insulin', 'Leptin']*
	1.00	6	4	*['Age', 'Resistin', 'Insulin', 'Leptin']*

e. Penilaian Prestasi Subset Fitur Terbaik Menggunakan Model Pengelasan

Pemilihan subset fitur unik selepas penyingkiran fitur bertindih dinilai mengunakan lima model pengelasan iaitu DT, RF, NB, LR dan SVM. Tujuan utama adalah untuk mengenalpasti kombinasi nilai α dan β yang menghasilkan subset fitur terbaik berdasarkan skor AUC.

Jadual 4.25 menunjukkan prestasi pengelasan untuk set data WDBC. Subset terbaik dikenal pasti pada kombinasi α = 0.25, β = 0.80 yang menghasilkan 4 fitur utama iaitu *concave points_worst*, *perimeter_mean*, *radius_worst* dan *texture_worst* dengan purata skor AUC tertinggi 0.9857. Model NB, LR dan SVM mencatatkan skor AUC melebihi 0.99, manakala model RF juga menunjukkan prestasi cemerlang melebihi 0.98.

Jadual 4.25 Skor AUC bagi setiap model pengelasan bagi set data WDBC

α	β	Bilangan Fitur	Purata Skor AUC ± SD					
			DT	RF	NB	LR	SVM	Purata
0.20	0.70	2	0.9638 ± 0.0087	0.9765 ± 0.0073	0.9808 ± 0.0056	0.9802 ± 0.0059	0.9806 ± 0.0061	0.9764
0.25	0.70	3	0.9614 ± 0.0116	0.9787 ± 0.0065	0.9833 ± 0.0062	0.9854 ± 0.0054	0.9850 ± 0.0051	0.9788
	0.80	**4**	**0.9683 ± 0.0115**	**0.9853 ± 0.0057**	**0.9911 ± 0.0049**	**0.9919 ± 0.0049**	**0.9917 ± 0.0049**	**0.9857**
	0.90	9	0.9592 ± 0.0141	0.9875 ± 0.0067	0.9868 ± 0.0046	0.9943 ± 0.0036	0.9913 ± 0.0035	0.9838
	1.00	14	0.9649 ± 0.0124	0.9868 ± 0.0057	0.9752 ± 0.0066	0.9899 ± 0.0041	0.9914 ± 0.0057	0.9816
0.30	0.80	3	0.9564 ± 0.0162	0.9810 ± 0.0075	0.9853 ± 0.0046	0.9887 ± 0.0050	0.9890 ± 0.0052	0.9801
	0.90	9	0.9572 ± 0.0188	0.9900 ± 0.0051	0.9741 ± 0.0079	0.9936 ± 0.0036	0.9930 ± 0.0031	0.9816
	1.00	12	0.9601 ± 0.0153	0.9876 ± 0.0055	0.9755 ± 0.0063	0.9895 ± 0.0046	0.9898 ± 0.0045	0.9805
0.35	0.70	2	0.9513 ± 0.0160	0.9631 ± 0.0093	0.9657 ± 0.0087	0.9691 ± 0.0090	0.9679 ± 0.0082	0.9634
	0.80	3	0.9544 ± 0.0148	0.9673 ± 0.0104	0.9653 ± 0.0075	0.9726 ± 0.0088	0.9713 ± 0.0076	0.9662
	0.90	7	0.9584 ± 0.0158	0.9838 ± 0.0083	0.9760 ± 0.0059	0.9852 ± 0.0059	0.9830 ± 0.0081	0.9773
	1.00	6	0.9624 ± 0.0128	0.9865 ± 0.0049	0.9781 ± 0.0074	0.9889 ± 0.0035	0.9884 ± 0.0038	0.9809
0.40	0.80	2	0.9396 ± 0.0100	0.9579 ± 0.0085	0.9597 ± 0.0066	0.9623 ± 0.0068	0.9653 ± 0.0062	0.9570
	0.90	3	0.9547 ± 0.0142	0.9707 ± 0.0089	0.9682 ± 0.0076	0.9680 ± 0.0077	0.9712 ± 0.0074	0.9666
0.45	0.90	1	0.9458 ± 0.0136	0.9561 ± 0.0111	0.9626 ± 0.0082	0.9626 ± 0.0082	0.9627 ± 0.0082	0.9580

Untuk set data WPBC, Jadual 4.26 memperlihatkan bahawa α = 0.30 dan β = 0.70 menghasilkan subset terbaik yang terdiri daripada 5 fitur (*area3*, *tumor_size*, *compactness2*, *radius3* dan *Time*) dengan purata skor AUC sebanyak 0.7015. Model

LR menunjukkan prestasi tertinggi (AUC 0.7383 ± 0.0544), diikuti rapat oleh SVM (AUC 0.7352 ± 0.0740).

Jadual 4.26 Skor AUC bagi setiap model pengelasan bagi set data WPBC

α	β	Bilangan Fitur	Purata Skor AUC ± SD					
			DT	RF	NB	LR	SVM	Purata
0.30	**0.70**	**5**	**0.6865 ± 0.0582**	**0.6685 ± 0.0789**	**0.6789 ± 0.0839**	**0.7383 ± 0.0544**	**0.7352 ± 0.0740**	**0.7015**
	0.80	6	0.6412 ± 0.1194	0.6606 ± 0.0882	0.6620 ± 0.0928	0.7405 ± 0.0529	0.7142 ± 0.0553	0.6837
	0.90	2	0.5707 ± 0.0744	0.5829 ± 0.0644	0.6021 ± 0.0417	0.6104 ± 0.0681	0.5525 ± 0.1261	0.5837
0.35	0.80	3	0.5240 ± 0.0862	0.5521 ± 0.0771	0.6080 ± 0.0462	0.6540 ± 0.0683	0.5564 ± 0.1634	0.5789

Seterusnya, untuk set data BCC, Jadual 4.27 bahawa subset fitur terbaik diperoleh pada α = 0.05 dan β = 0.70, terdiri daripada 4 fitur iaitu *Age*, *Resistin*, *Glucose* dan *Leptin*. Subset ini menghasilkan purata skor AUC tertinggi 0.8044 dengan model SVM mencapai skor AUC tertinggi (0.8439 ± 0.0717), diikuti oleh model RF (0.8412 ± 0.0602.

Jadual 4.27 Skor AUC bagi setiap model pengelasan bagi set data BCC

α	β	Bilangan Fitur	Purata Skor AUC ± SD					
			DT	RF	NB	LR	SVM	Purata
0.05	**0.70**	**4**	**0.7676 ± 0.0712**	**0.8412 ± 0.0602**	**0.7773 ± 0.0861**	**0.7922 ± 0.0893**	**0.8439 ± 0.0717**	**0.8044**
	0.90	5	0.7364 ± 0.0782	0.8201 ± 0.0732	0.7658 ± 0.0611	0.7872 ± 0.0785	0.8169 ± 0.0610	0.7852
	1.00	6	0.6798 ± 0.0846	0.8233 ± 0.0606	0.7941 ± 0.0434	0.7947 ± 0.0723	0.8100 ± 0.0575	0.7804
0.10	0.80	3	0.6281 ± 0.0840	0.7007 ± 0.0885	0.6546 ± 0.1204	0.6481 ± 0.0638	0.6251 ± 0.0744	0.6513
	1.00	4	0.6766 ± 0.0854	0.7449 ± 0.0603	0.7317 ± 0.1032	0.7128 ± 0.0526	0.6075 ± 0.1162	0.6947

4.3 HASIL MODEL PENGELASAN

Seksyen ini membincangkan keputusan pengelasan menggunakan pengukuran seperti skor AUC, ACC, SEN, SPE dan F1. Hiperparameter bagi lima model pengelasan ini telah ditala menggunakan kaedah *Grid Search* dan menggunakan belahan data dengan *random_state*=42 dan pengesahan silang sebanyak 5 kali. Konfigurasi hiperparameter daripada penalaan ini digunakan dalam 10 percubaan penilaian model.

Jadual 4.28 menunjukkan hasil prestasi menggunakan subset fitur terbaik yang telah dipilih bagi set data WDBC.

Jadual 4.28 Nilai skor metrik penilaian setiap model pengelasan bagi set data WDBC

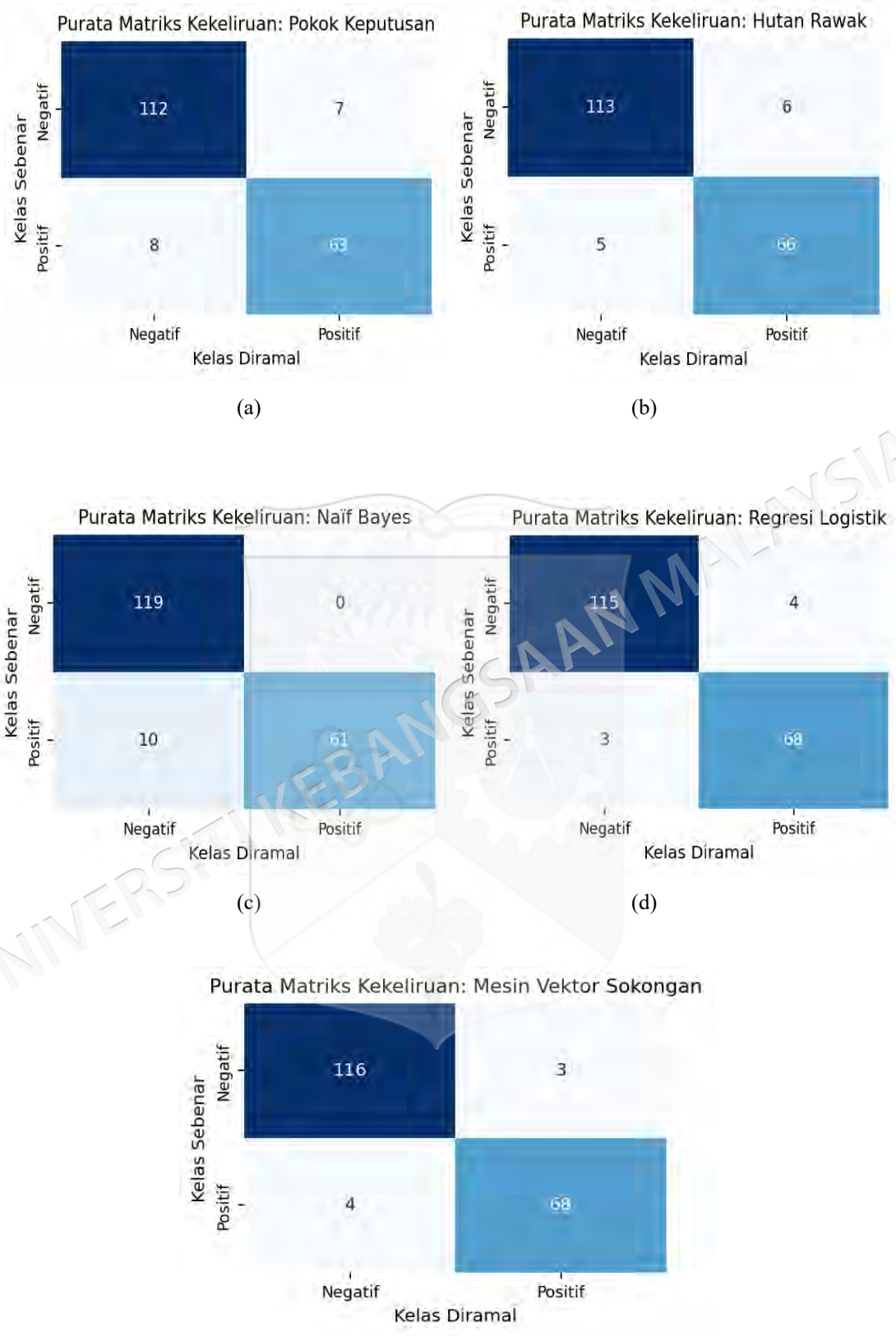
Metrik	Model Pengelasan (Skor ± SD)				
	DT	**RF**	**NB**	**LR**	**SVM**
AUC	0.9683 ± 0.0115	0.9853 ± 0.0057	0.9911 ± 0.0049	**0.9919 ± 0.0049**	0.9917 ± 0.0049
ACC	0.9242 ± 0.0188	0.9432 ± 0.0195	0.9479 ± 0.0132	0.9637 ± 0.0128	**0.9647 ± 0.0114**
SEN	0.8901 ± 0.0350	0.9296 ± 0.0188	0.8648 ± 0.0383	**0.9535 ± 0.0188**	0.9507 ± 0.0179
SPE	0.9445 ± 0.0337	0.9513 ± 0.0250	**0.9975 ± 0.0041**	0.9697 ± 0.0149	0.9731 ± 0.0103
F1	0.8980 ± 0.0238	0.9247 ± 0.0251	0.9250 ± 0.0207	0.9516 ± 0.0168	**0.9527 ± 0.0153**

Berdasarkan Jadual 4.28, model LR menunjukkan prestasi tertinggi dengan skor AUC sebanyak 0.9919 ± 0.0049, diikuti rapat oleh SVM (0.9917 ± 0.0049) dan NB (0.9911 ± 0.0049). Model RF juga menunjukkan prestasi yang sangat baik dengan skor AUC sebanyak 0.9853 ± 0.0057. Model DT pula menunjukkan skor AUC yang lebih rendah berbanding model lain dengan nilai 0.9683 ± 0.0115.

Bagi penilaian skor ACC, model LR dan SVM mencatatkan nilai tertinggi masing-masing 96.37% dan 96.47% diikuti oleh RF (94.32%). Model NB dan DT menunjukkan ACC yang lebih rendah masing-masing 94.79% dan 92.42%. Bagi penilaian skor SEN pula, model LR mencatatkan nilai tertinggi (0.9535 ± 0.0188), diikuti oleh SVM (0.9507 ± 0.0179) dan RF (0.9296 ± 0.0188). Kepekaan yang tinggi ini menunjukkan bahawa model ini sangat berkesan dalam mengenal pasti kes malignan.

Selain itu, bagi penilaian skor SPE, model NB mencatatkan prestasi terbaik (0.9975 ± 0.0041) menandakan keupayaan model dalam mengurangkan kesilapan pengelasan bagi kes benigna. Model SVM (0.9731 ± 0.0103) dan LR (0.9697 ± 0.0149) juga menunjukkan skor SPE yang tinggi membuktikan keupayaan mereka dalam membezakan tumor benigna dengan betul. Akhir sekali, bagi penilaian skor F1, model SVM dan LR mencatatkan prestasi terbaik dengan skor masing-masing 0.9527 ± 0.0153 dan 0.9516 ± 0.0168. Model RF juga mencatatkan prestasi yang tinggi dengan skor 0.9247 ± 0.0251 membuktikan model ini mencapai keseimbangan antara malignan dan benigna serta mengurangkan kesilapan pengelasan. Secara keseluruhan, model LR dan SVM adalah model terbaik untuk pengelasan tumor dalam set data WDBC berdasarkan subset fitur yang dipilih.

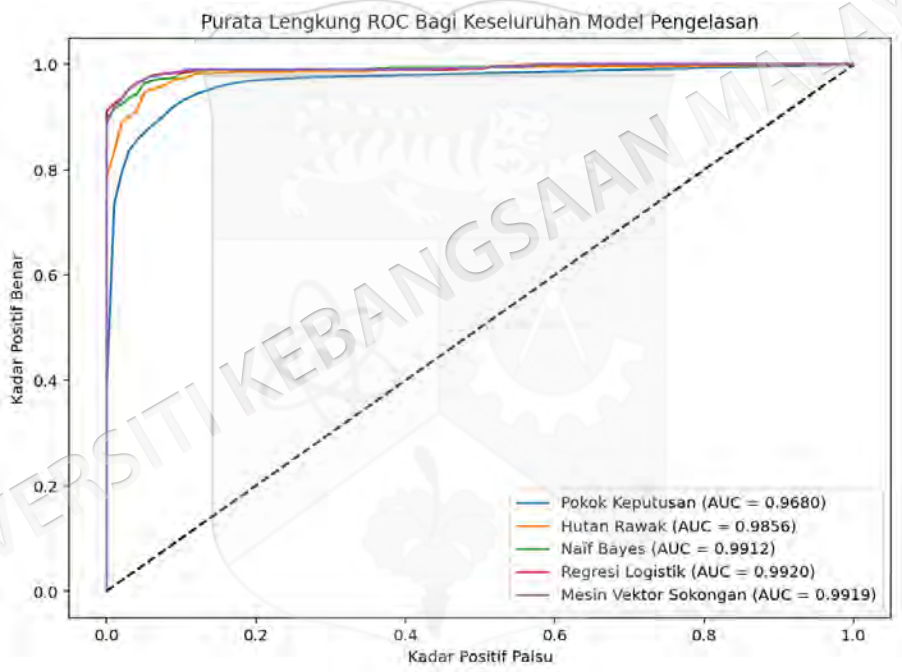
Rajah 4.8 menunjukkan purata matriks kekeliruan bagi setiap model pengelasan bagi set data WDBC.



Rajah 4.8 Purata matriks kekeliruan bagi model pengelasan (a) DT (b) RF (c) NB (d) LR (e) SVM bagi set data WDBC

Berdasarkan Rajah 4.8, model NB mencatatkan TN tertinggi dengan 119 kes namun mencatatkan FN tertinggi dengan 10 kes. Model LR dan SVM pula menunjukkan TP tertinggi dengan 68 kes. Tambahan pula, kedua-dua model juga menunjukkan FP dan FN yang rendah. Model RF pula mencatatkan prestasi yang seimbang dengan 113 TN dan 66 TP.

Rajah 4.9 menunjukkan purata lengkung ROC untuk setiap model pengelasan bagi set data WDBC. Lengkung model LR dan SVM jelas lebih mendekati sudut kiri atas berbanding NB, RF dan DT. Ini membuktikan model LR dan SVM merupakan model terbaik dalam membezakan kelas malignan dan benigna.



Rajah 4.9 Purata lengkung ROC bagi setiap model pengelasan bagi set data WDBC

Seterusnya, Jadual 4.29 menunjukkan hasil prestasi menggunakan subset fitur terbaik yang telah dipilih bagi set data WPBC.

Jadual 4.29 Nilai skor metrik penilaian setiap model pengelasan bagi set data WPBC

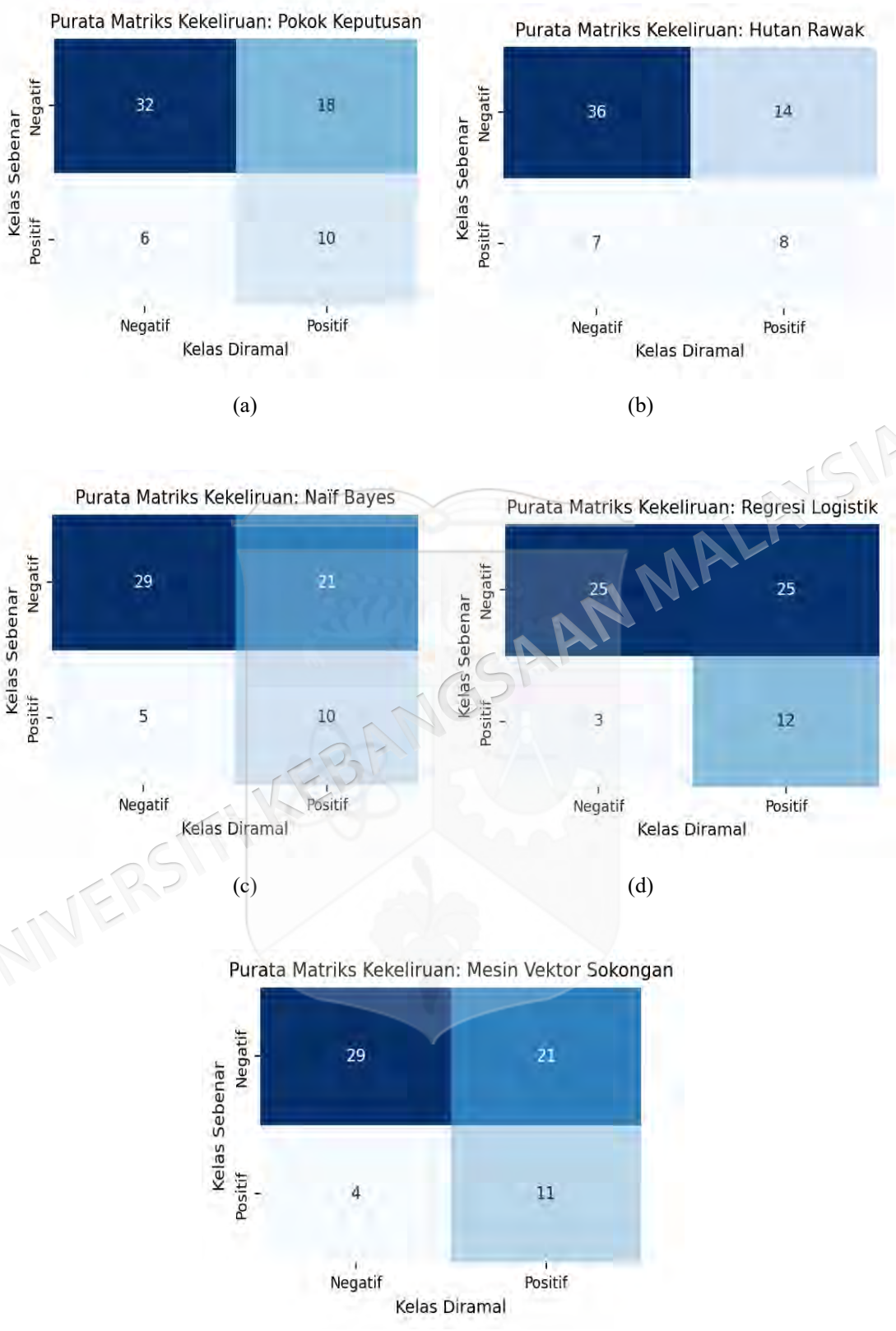
Metrik	Model Pengelasan (Skor ± SD)				
	DT	**RF**	**NB**	**LR**	**SVM**
AUC	0.6865 ± 0.0581	0.6685 ± 0.0789	0.6789 ± 0.0839	**0.7383 ± 0.0544**	0.7243 ± 0.0811
ACC	0.6431 ± 0.0422	**0.6723 ± 0.0696**	0.5985 ± 0.0727	0.5769 ± 0.0504	0.6138 ± 0.0545
SEN	0.6333 ± 0.1379	0.5067 ± 0.1184	0.6867 ± 0.0773	**0.8200 ± 0.0945**	0.7200 ± 0.0820
SPE	0.6460 ± 0.0611	**0.7220 ± 0.0780**	0.5720 ± 0.0790	0.5040 ± 0.0580	0.5820 ± 0.0553
F1	0.4470 ± 0.0686	0.4180 ± 0.1001	0.4446 ± 0.0703	**0.4725 ± 0.0500**	0.4638 ± 0.0588

Berdasarkan Jadual 4.29, model LR menunjukkan skor AUC tertinggi iaitu 0.7383 ± 0.0544) diikuti oleh SVM (0.7243 ± 0.0811). Model DT dan NB mempunyai skor AUC yang hampir sama sekitar 0.68, manakala RF menunjukkan nilai terendah (0.6685 ± 0.0789).

Bagi penilaian skor ACC, model RF mencatatkan nilai tertinggi (0.6723 ± 0.0696), diikuti oleh DT (0.6431 ± 0.0422) 0.6138 ± 0.0545). Model SVM dan NB masing-masing menunjukkan ACC sebanyak 0.6138 ± 0.0545 dan 0.5985 ± 0.0727, manakala LR mencatatkan ACC terendah dengan 0.5769 ± 0.0504. Bagi penilaian skor SEN pula, model LR menunjukkan skor tertinggi (0.8200 ± 0.0945) diikuti oleh SVM (0.7200 ± 0.0820). Model NB mencatatkan nilai SEN yang lebih rendah (0.6867 ± 0.0773), manakala model RF mencatatkan nilai SEN yang paling rendah (0.5067 ± 0.1184).

Di samping itu, bagi penilaian skor SPE, model RF mencatatkan nilai tertinggi (0.7220 ± 0.0780) diikuti oleh DT (0.6460 ± 0.0611). Model SVM dan NB menunjukkan SPE yang sederhana sekitar 0.58, manakala LR mempunyai nilai SPE terendah (0.5040 ± 0.0580). Akhir sekali, bagi skor F1, model LR mencatatkan nilai tertinggi (0.4725 ± 0.0500) diikuti oleh SVM (0.4638 ± 0.0588). Model DT dan NB mencatatkan skor F1 terendah sekitar 0.44 dan 0.45 masing-masing. Secara keseluruhan, model LR dan SVM adalah model terbaik untuk pengelasan pengulangan kanser dalam set data WPBC.

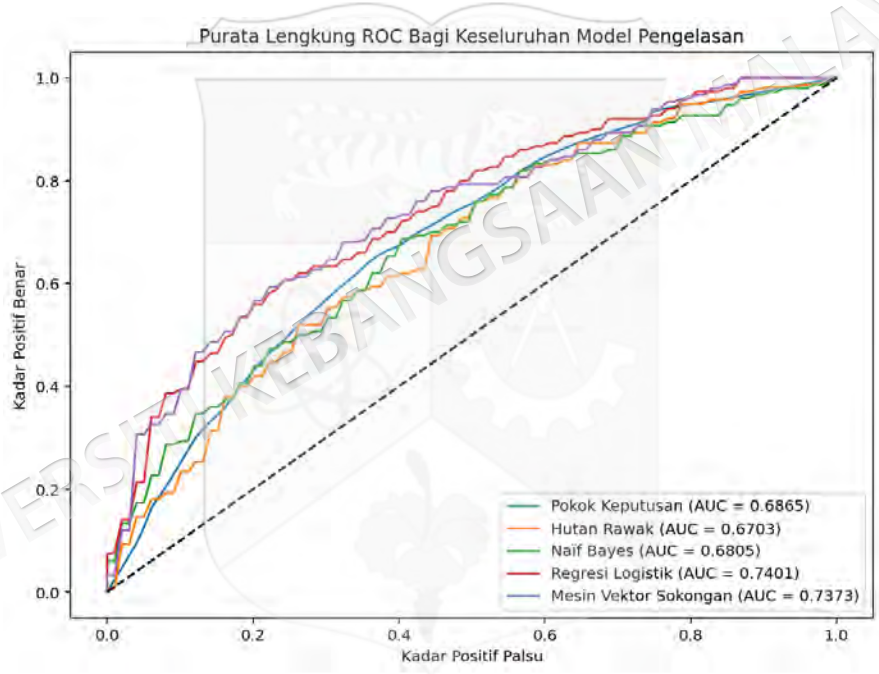
Rajah 4.10 menunjukkan purata matriks kekeliruan bagi setiap model pengelasan bagi set data WPBC.



Rajah 4.10 Purata matriks kekeliruan bagi model pengelasan (a) DT (b) RF (c) NB (d) LR (e) SVM bagi set data WPBC

Berdasarkan Rajah 4.10, model RF mencatatkan TN tertinggi dengan 36 kes dan terendah bagi FP sebanyak 14 kes. Model LR pula menunjukkan TP tertinggi dengan 12 kes dan terendah pada FN sebanyak 3 kes namun tertinggi pada FP sebanyak 25 kes. Model DT pula mencatatkan prestasi yang seimbang dengan 32 TN dan 10 TP.

Rajah 4.11 menunjukkan purata lengkung ROC bagi keseluruhan model pengelasan bagi set data WPBC. Lengkung ROC bagi model LR dan SVM jelas lebih mendekati sudut kiri atas berbanding DT, NB dan RF. Ini membuktikan model LR dan SVM merupakan model terbaik dalam membezakan kelas kanser berulang dan tidak berulang.



Rajah 4.11 Purata lengkung ROC bagi keseluruhan model pengelasan bagi set data WPBC

Seterusnya, Jadual 4.30 menunjukkan hasil prestasi menggunakan subset fitur terbaik yang telah dipilih bagi set data BCC.

Jadual 4.30 Nilai skor metrik penilaian bagi setiap model pengelasan bagi set data BCC

Metrik	Model Pengelasan (Skor ± SD)				
	DT	**RF**	**NB**	**LR**	**SVM**
AUC	0.7676 ± 0.0712	0.8412 ± 0.0602	0.7773 ± 0.0861	0.7922 ± 0.0893	**0.8439 ± 0.0717**
ACC	0.7256 ± 0.0663	**0.7718 ± 0.049**	0.6103 ± 0.0878	0.7103 ± 0.0746	0.7667 ± 0.0924
SEN	0.7227 ± 0.1202	**0.8000 ± 0.0963**	0.3909 ± 0.1425	0.6636 ± 0.0987	0.7273 ± 0.1389
SPE	0.7294 ± 0.1446	0.7353 ± 0.0635	**0.8941 ± 0.0723**	0.7706 ± 0.1158	0.8176 ± 0.0515
F1	0.7450 ± 0.0758	**0.7959 ± 0.0533**	0.5196 ± 0.1427	0.7191 ± 0.0799	0.7731 ± 0.1011

Berdasarkan Jadual 4.30, model SVM mencatatkan skor AUC tertinggi (0.8439 ± 0.0717) diikuti oleh RF (0.8412 ± 0.0602). Model LR menunjukkan skor AUC yang agak baik (0.7922 ± 0.0893), manakala NB mencatatkan nilai AUC yang lebih rendah (0.7773 ± 0.0861) dan DT mempunyai skor AUC terendah (0.7676 ± 0.0712).

Bagi penilaian skor ACC, model RF mencatatkan nilai tertinggi (0.7718 ± 0.049), diikuti oleh SVM (0.7667 ± 0.0924) dan DT (0.7256 ± 0.0663). Model LR menunjukkan ACC sebanyak 0.7103 ± 0.0746 manakala model NB mencatatkan ACC paling rendah (0.6103 ± 0.0878). Bagi penilaian skor SEN, model RF menunjukkan nilai tertinggi (0.8000 ± 0.0963), diikuti oleh SVM (0.7273 ± 0.1389). DT mencatatkan kepekaan yang agak baik (0.7227 ± 0.1202), manakala LR menunjukkan nilai yang lebih rendah (0.6636 ± 0.0987). Model NB mencatatkan nilai SEN terendah (0.3909 ± 0.1425).

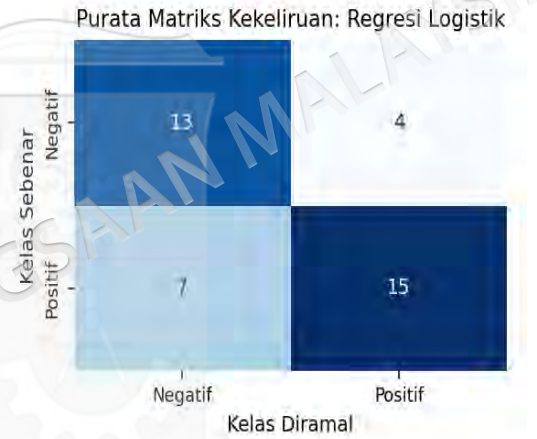
Bagi penilaian skor SPE, model NB mencatatkan nilai tertinggi (0.8941 ± 0.0723), diikuti oleh SVM (0.8176 ± 0.0515) dan RF (0.7353 ± 0.0635). Model LR dan DT menunjukkan SPE yang baik masing-masing 0.7706 ± 0.1158 dan 0.7294 ± 0.1446. Akhir sekali, bagi skor F1, RF mencatatkan nilai tertinggi (0.7959 ± 0.0533), diikuti oleh SVM (0.7731 ± 0.1011) dan DT (0.7450 ± 0.0758). Model LR mencatatkan skor F1 sebanyak 0.7191 ± 0.0799, manakala NB mempunyai skor F1 terendah (0.5196 ± 0.1427). Secara keseluruhan, SVM dan RF adalah model terbaik untuk pengelasan individu kepada pesakit atau kawalan sihat dalam set data BCC.

Rajah 4.12 menunjukkan purata matriks kekeliruan bagi setiap model pengelasan bagi set data BCC.

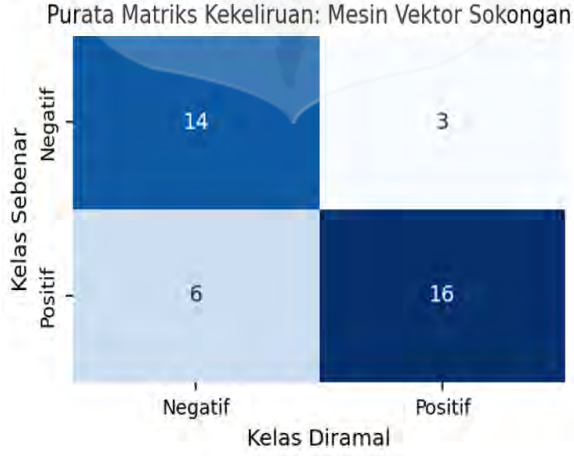
(a)

(b)

(c)



(d)

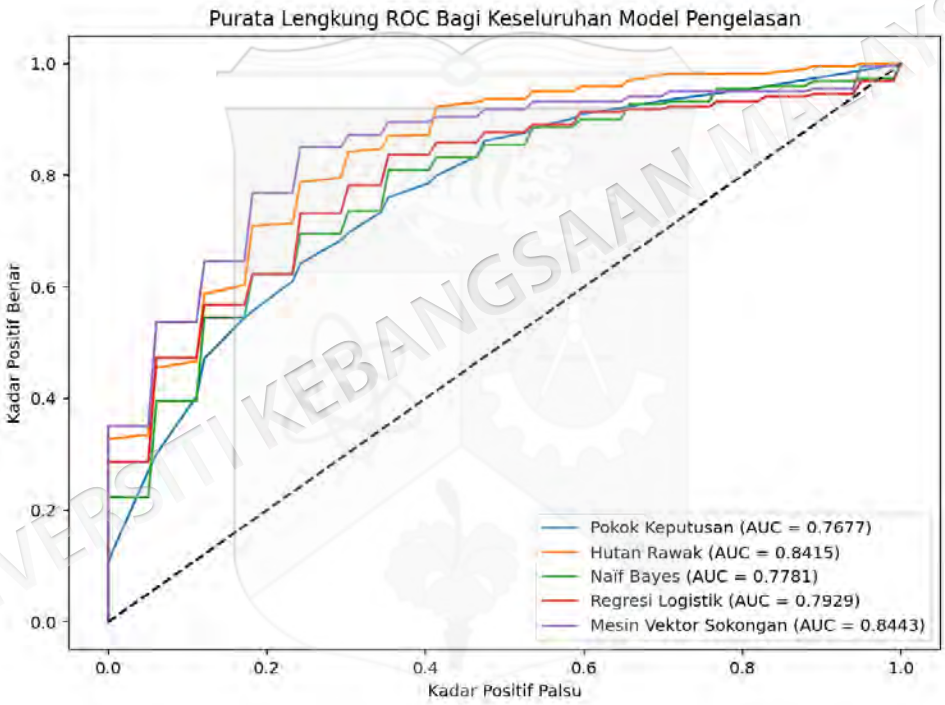


(e)

Rajah 4.12 Purata matriks kekeliruan bagi model pengelasan (a) DT (b) RF (c) NB (d) LR (e) SVM bagi set data BCC

Berdasarkan Rajah 4.12, model RF menunjukkan TP tertinggi dengan 18 kes dan terendah pada FN sebanyak 4 kes. Model SVM pula mencatatkan prestasi yang seimbang dengan 14 TN dan 16 TP. Sebaliknya, model NB mencatatkan agak tinggi FN iaitu 13 kes.

Rajah 4.13 menunjukkan purata lengkung ROC bagi keseluruhan model pengelasan bagi set data BCC. Lengkung ROC bagi model SVM dan RF jelas lebih mendekati sudut kiri atas berbanding LR, NB dan DT. Ini membuktikan model SVM dan RF merupakan model terbaik dalam membezakan kelas pesakit dan kawalan sihat.



Rajah 4.13 Purata lengkung ROC bagi keseluruhan model pengelasan bagi set data BCC

4.4 PERBANDINGAN BILANGAN SUBSET FITUR DIPILIH

Seksyen ini membandingkan bilangan subset fitur yang dipilih oleh kaedah FD-CARI dan kaedah pemilihan fitur piawai bagi setiap set data. Jadual 4.31 menyenaraikan subset fitur terpilih oleh setiap kaedah pemilihan fitur bagi set data WDBC.

Jadual 4.31 Subset fitur terpilih oleh kaedah pemilihan fitur bagi set data WDBC

Kaedah Pemilihan Fitur	Bilangan Fitur	Senarai Fitur	Peratusan Subset Fitur Terpilih
CFS	11	*texture_mean, concavity_mean, concave points_mean, area_se, symmetry_se, radius_worst, perimeter_worst, area_worst, smoothness_worst, concavity_worst, concave points_worst*	36.67%
FCBF	7	*perimeter_worst, concave points_worst, radius_se, texture_mean, smoothness_worst, symmetry_worst, symmetry_se*	23.33%
Consistency	8	*texture_mean, radius_se, perimeter_se, radius_worst, texture_worst, concavity_worst, concave points_worst, symmetry_worst*	27%
Relief-F	4	*radius_worst, concave points_worst, perimeter_worst, texture_worst*	13.33%
mRMR	4	*perimeter_worst, fractal_dimension_se, texture_mean, symmetry_worst*	13.33%
FD-CARI	4	*concave points_worst, perimeter_mean, radius_worst, texture_worst*	13.33%

Berdasarkan Jadual 4.31, kaedah CFS memilih 11 fitur (36.67%), diikuti dengan Consistency dengan 8 fitur (27%) dan FCBF dengan 7 fitur (23.33%). Kaedah FD-CARI pula memilih 4 fitur (13.33%). Oleh disebabkan kaedah Relief-F dan mRMR adalah kaedah berdasarkan penarafan fitur, maka jumlah fitur yang dipilih diselaraskan sama kaedah dengan FD-CARI iaitu 4 fitur bagi memastikan perbandingan yang adil.

Seterusnya, subset fitur terpilih bagi setiap kaedah pemilihan fitur bagi set data WPBC disenaraikan pada Jadual 4.32.

Jadual 4.32 Subset fitur terpilih oleh kaedah pemilihan fitur bagi set data WPBC

Kaedah Pemilihan Fitur	Bilangan Fitur	Senarai Fitur	Peratusan Subset Fitur Terpilih
CFS	2	*Time, lymph_node_status*	6.25%
FCBF	2	*Time, lymph_node_status*	6.25%
Consistency	2	*Time, lymph_node_status*	6.25%
Relief-F	5	*Time, texture3, lymph_node_status, compactness1, symmetry1*	15.63%
mRMR	5	*Time, concavity3, smoothness2, area3, lymph_node_status*	15.63%
FD-CARI	5	*Time, area3, tumor_size, compactness2, radius3*	15.63%

Berdasarkan Jadual 4.32, kaedah CFS, FCBF dan Consistency hanya memilih 2 fitur (6.25%) iaitu *Time* dan *lymph_node_status*. Sebaliknya, kaedah Relief-F, mRMR, dan FD-CARI memilih lebih banyak fitur iaitu 5 fitur (15.63%).

Seterusnya Jadual 4.33 menyenaraikan subset fitur terpilih oleh setiap kaedah pemilihan fitur bagi set data BCC.

Jadual 4.33 Subset fitur terpilih oleh kaedah pemilihan fitur bagi set data BCC

Kaedah Pemilihan Fitur	**Bilangan Fitur**	**Senarai Fitur**	**Peratusan Subset Fitur Terpilih**
CFS	3	*Age, Glucose, HOMA*	33.33%
FCBF	2	*Glucose, HOMA*	22.22%
Consistency	3	*Age, Glucose, HOMA*	33.33%
Relief-F	4	*Age, Glucose, Resistin, Insulin*	44.44%
mRMR	4	*Glucose, Adiponectin, Resistin, MCP.1*	44.44%
FD-CARI	4	*Age, Glucose, Leptin, Resistin*	44.44%

Berdasarkan Jadual 4.33, kaedah CFS dan Consistency memilih 3 fitur (33.33%) iaitu *Age*, *Glucose* dan *HOMA*. Kaedah FCBF pula hanya memilih 2 fitur (22.22%) iaitu *Glucose* dan *HOMA*. Sebaliknya, kaedah FD-CARI, Relief-F dan mRMR masing-masing menambah 1 lagi fitur menjadikannya 4 fitur (44.44%).

4.5 PERBANDINGAN PRESTASI MODEL PEMILIHAN FITUR

Seksyen ini membandingkan prestasi model pengelasan berdasarkan subset fitur yang dipilih oleh kaedah FD-CARI dan kaedah pemilihan fitur piawai. Keseluruhan 10 percubaan model kaedah FD-CARI untuk setiap set data dirujuk pada Lampiran A.

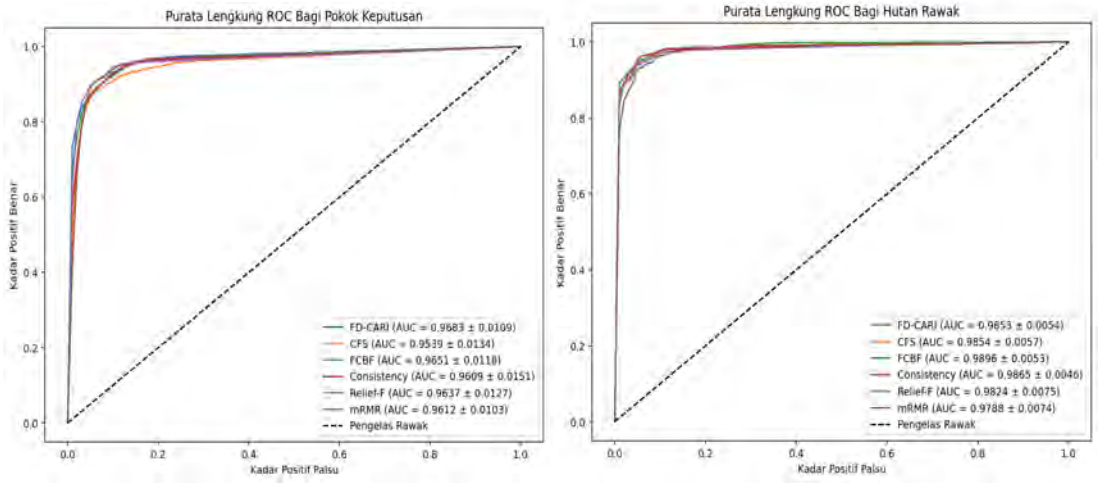
4.5.1 Set Data WDBC

Subseksyen ini membincangkan prestasi penilaian metrik AUC, ACC, SEN, SPE dan F1 untuk setiap kaedah pemilihan fitur bagi set data WDBC. Jadual 4.34 mencatatkan skor AUC bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.34 Nilai skor AUC bagi setiap model pengelasan

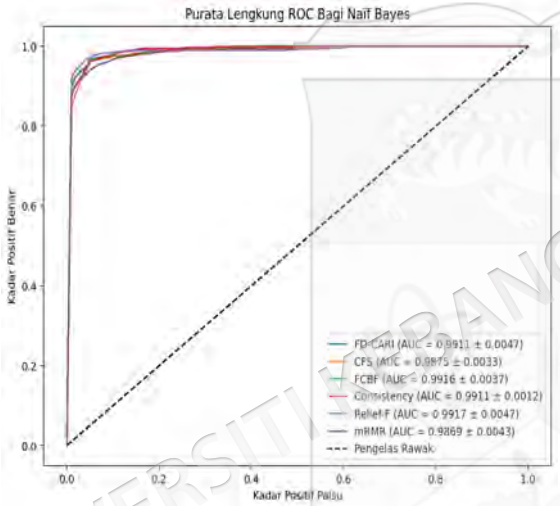
Kaedah Pemilihan Fitur	Nilai Skor AUC ± SD					Purata AUC
	DT	RF	NB	LR	SVM	
CFS	0.9539 ± 0.0141	0.9854 ± 0.0060	0.9875 ± 0.0035	0.9920 ± 0.0035	0.9878 ± 0.0054	0.9813
FCBF	0.9651 ± 0.0124	**0.9896 ± 0.0056**	0.9916 ± 0.0039	0.9924 ± 0.0029	0.9942 ± 0.0035	**0.9866**
Consistency	0.9609 ± 0.0159	0.9865 ± 0.0048	0.9911 ± 0.0012	**0.9944 ± 0.0031**	**0.9946 ± 0.0027**	0.9855
Relief-F	0.9637 ± 0.0134	0.9824 ± 0.0079	**0.9917 ± 0.0049**	0.9916 ± 0.0049	0.9918 ± 0.0047	0.9843
mRMR	0.9612 ± 0.0108	0.9788 ± 0.0077	0.9869 ± 0.0046	0.9856 ± 0.0060	0.9857 ± 0.0056	0.9796
FD-CARI	**0.9683 ± 0.0115**	0.9853 ± 0.0057	0.9911 ± 0.0049	0.9919 ± 0.0049	0.9917 ± 0.0049	0.9857

Berdasarkan Jadual 4.34, kaedah FCBF mencatatkan purata skor AUC tertinggi iaitu 0.9866, diikuti oleh FD-CARI (0.9857) dan Consistency (0.9855). Kaedah FCBF menunjukkan prestasi terbaik pada model RF dengan nilai skor AUC sebanyak 0.9896 ± 0.0056 manakala kaedah FD-CARI mencatatkan skor AUC tertinggi bagi DT iaitu 0.9683 ± 0.0115. Kaedah Consistency pula menunjukkan prestasi cemerlang pada model LR dan SVM dengan skor AUC tertinggi masing-masing 0.9944 ± 0.0031 dan 0.9946 ± 0.0027. Kaedah Relief-F dan CFS pula masing-masing mencatatkan purata skor AUC sebanyak 0.9843 dan 0.9813 menunjukkan prestasi yang agak baik namun sedikit lebih rendah. Sebaliknya, kaedah mRMR mencatatkan purata skor AUC terendah iaitu 0.9796. Rajah 4.14 memaparkan purata lengkungan ROC bagi setiap model pengelasan.

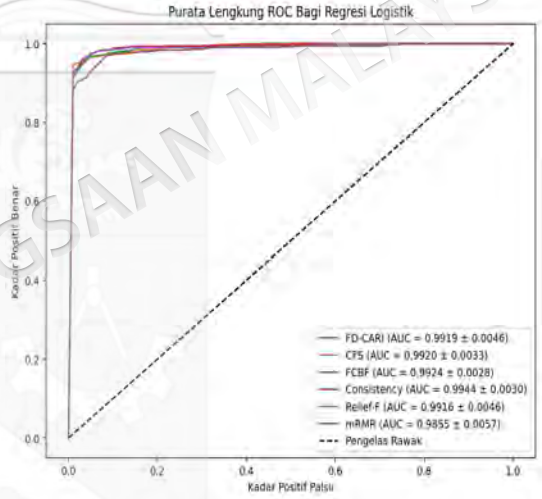


(a)

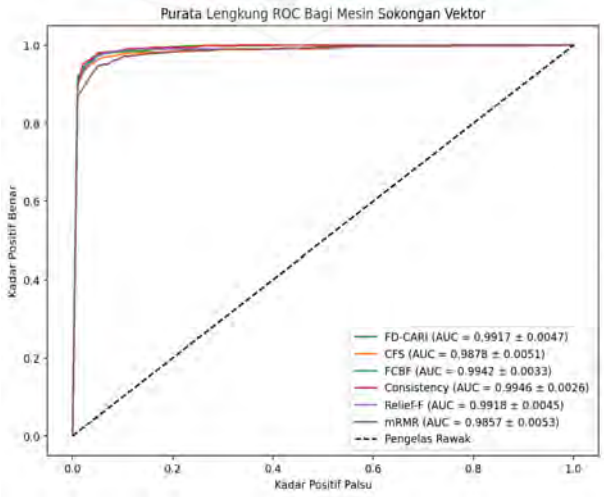
(b)



(c)



(d)



(e)

Rajah 4.14 Purata lengkungan ROC bagi model (a) DT (b) RF (c) NB (d) LR (e) SVM

Rajah 4.14 menunjukkan lengkung ROC bagi model LR dan SVM menggunakan kaedah Consistency dan FCBF jelas lebih mendekati sudut kiri yang mencerminkan prestasi yang lebih cemerlang berbanding model DT, RF dan NB.

Seterusnya, skor ACC bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur dicatatkan pada Jadual 4.35.

Jadual 4.35 Nilai skor ACC bagi setiap model pengelasan

Kaedah Pemilihan Fitur	**Nilai Skor ACC ± SD**					**Purata ACC**
	DT	**RF**	**NB**	**LR**	**SVM**	
CFS	0.9142 ± 0.0179	0.9447 ± 0.0134	0.9389 ± 0.0125	**0.9684 ± 0.0134**	0.9632 ± 0.0093	0.9459
FCBF	0.9189 ± 0.0167	**0.9505 ± 0.0170**	0.9426 ± 0.0130	0.9579 ± 0.0134	0.9605 ± 0.0117	0.9461
Consistency	0.9200 ± 0.0172	0.9495 ± 0.0117	0.9326 ± 0.0135	0.9637 ± 0.0123	**0.9663 ± 0.0100**	0.9464
Relief-F	**0.9326 ± 0.0162**	0.9463 ± 0.0170	**0.9526 ± 0.0111**	0.9668 ± 0.0111	0.9658 ± 0.0122	**0.9528**
mRMR	0.9295 ± 0.0109	0.9395 ± 0.0153	0.9521 ± 0.0123	0.9405 ± 0.0177	0.9516 ± 0.0131	0.9426
FD-CARI	0.9242 ± 0.0188	0.9432 ± 0.0195	0.9479 ± 0.0132	0.9637 ± 0.0128	0.9647 ± 0.0114	0.9487

Berdasarkan Jadual 4.35, kaedah Relief-F mencatatkan purata skor ACC tertinggi iaitu 0.9528, diikuti oleh FD-CARI (0.9487) dan Consistency (0.9464). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model DT dan RF dengan nilai skor ACC masing-masing sebanyak 0.9326 ± 0.0162 dan 0.9526 ± 0.0111 manakala Consistency menunjukkan prestasi cemerlang pada model SVM dengan skor ACC tertinggi sebanyak 0.9663 ± 0.0100. Kaedah FCBF dan CFS pula masing-masing mencatatkan purata skor ACC sebanyak 0.9461 dan 0.9459 menunjukkan prestasi yang agak baik namun sedikit lebih rendah berbanding tiga kaedah utama tersebut. Sebaliknya, kaedah mRMR mencatatkan purata skor ACC terendah iaitu 0.9426.

Seterusnya, Jadual 4.36 mencatatkan skor SEN bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.36 Nilai skor SEN bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor SEN ± SD					Purata SEN
	DT	RF	NB	LR	SVM	
CFS	0.8803 ± 0.0377	0.9254 ± 0.0319	0.8761 ± 0.0331	0.9493 ± 0.0119	0.9380 ± 0.0178	0.9138
FCBF	0.8676 ± 0.0502	0.9239 ± 0.0232	0.8592 ± 0.0409	0.9366 ± 0.0223	0.9296 ± 0.0230	0.9034
Consistency	0.8803 ± 0.0389	0.9254 ± 0.0230	0.8282 ± 0.0397	0.9493 ± 0.0291	**0.9535 ± 0.0282**	0.9073
Relief-F	**0.9056 ± 0.0455**	0.9239 ± 0.0327	0.8761 ± 0.0324	0.9521 ± 0.0201	0.9507 ± 0.0191	**0.9217**
mRMR	0.8845 ± 0.0344	0.9155 ± 0.0376	**0.9000 ± 0.0366**	0.9211 ± 0.0405	0.9268 ± 0.0368	0.9096
FD-CARI	0.8901 ± 0.0350	**0.9296 ± 0.0188**	0.8648 ± 0.0383	**0.9535 ± 0.0188**	0.9507 ± 0.0179	0.9177

Berdasarkan Jadual 4.36, kaedah Relief-F mencatatkan purata skor SEN tertinggi iaitu 0.9217, diikuti oleh FD-CARI (0.9177) dan CFS (0.9138). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model DT dengan nilai skor SEN sebanyak 0.9056 ± 0.0455 manakala kaedah FD-CARI menunjukkan prestasi cemerlang pada model RF dan LR dengan skor SEN tertinggi masing-masing sebanyak 0.9296 ± 0.0188 dan 0.9535 ± 0.0188. Kaedah mRMR dan Consistency pula masing-masing mencatatkan purata skor SEN sebanyak 0.9096 dan 0.9073 menunjukkan prestasi yang agak baik namun sedikit lebih rendah berbanding tiga kaedah utama tersebut. Sebaliknya, kaedah FCBF mencatatkan purata skor SEN terendah iaitu 0.9034.

Seterusnya, skor SPE bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur dicatatkan pada Jadual 4.37.

Jadual 4.37 Nilai skor SPE bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor SPE ± SD					Purata SPE
	DT	RF	NB	LR	SVM	
CFS	0.9345 ± 0.0396	0.9563 ± 0.0220	0.9765 ± 0.0142	**0.9798 ± 0.0210**	0.9782 ± 0.0120	0.9650
FCBF	0.9496 ± 0.0324	**0.9664 ± 0.0213**	0.9924 ± 0.0108	0.9706 ± 0.0178	**0.9790 ± 0.0155**	**0.9716**
Consistency	0.9437 ± 0.0217	0.9639 ± 0.0186	0.9950 ± 0.0071	0.9723 ± 0.0119	0.9739 ± 0.0101	0.9697
Relief-F	0.9487 ± 0.0196	0.9597 ± 0.0189	**0.9983 ± 0.0035**	0.9756 ± 0.0108	0.9748 ± 0.0137	0.9714
mRMR	**0.9563 ± 0.0152**	0.9538 ± 0.0071	0.9832 ± 0.0105	0.9521 ± 0.0168	0.9664 ± 0.0125	0.9624
FD-CARI	0.9445 ± 0.0337	0.9513 ± 0.0250	0.9975 ± 0.0041	0.9697 ± 0.0149	0.9731 ± 0.0103	0.9672

Berdasarkan Jadual 4.37, kaedah FCBF mencatatkan purata skor SPE tertinggi iaitu 0.9716, diikuti oleh Relief-F (0.9714) dan Consistency (0.9697). Kaedah FCBF menunjukkan prestasi terbaik berbanding kaedah lain pada model RF dan SVM dengan nilai skor SPE masing-masing sebanyak 0.9664 ± 0.0213 dan 0.9790 ± 0.0155 manakala kaedah Relief-F menunjukkan prestasi cemerlang pada model NB dengan skor SPE tertinggi sebanyak 0.9983 ± 0.0035. Kaedah FD-CARI dan CFS pula masing-masing mencatatkan purata skor SPE sebanyak 0.9672 dan 0.9650 menunjukkan prestasi yang agak baik namun sedikit lebih rendah berbanding tiga kaedah utama tersebut. Sebaliknya, kaedah mRMR mencatatkan purata skor SPE terendah iaitu 0.9624.

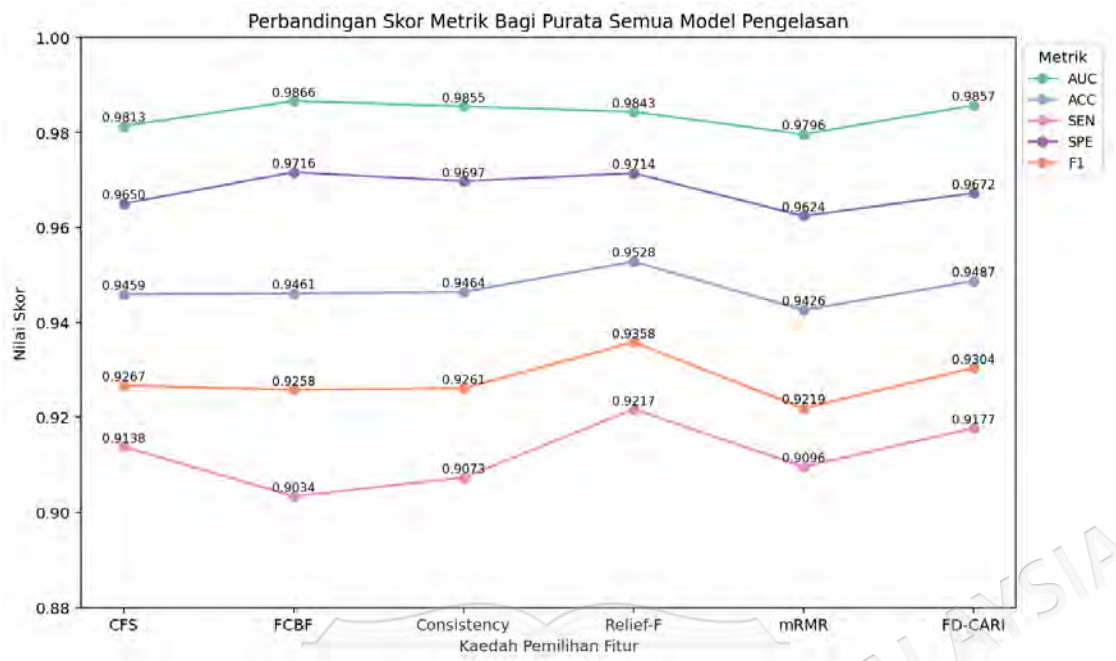
Seterusnya, Jadual 4.38 mencatatkan skor F1 bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.38 Nilai skor F1 bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor F1 ± SD					Purata F1
	DT	RF	NB	LR	SVM	
CFS	0.8851 ± 0.0209	0.9260 ± 0.0177	0.9145 ± 0.0181	**0.9576 ± 0.0169**	0.9501 ± 0.0125	0.9267
FCBF	0.8887 ± 0.0231	**0.9333 ± 0.0225**	0.9176 ± 0.0202	0.9433 ± 0.0176	0.9462 ± 0.0159	0.9258
Consistency	0.8914 ± 0.0239	0.9320 ± 0.0150	0.9014 ± 0.0220	0.9512 ± 0.0169	**0.9547 ± 0.0139**	0.9261
Relief-F	**0.9092 ± 0.0231**	0.9278 ± 0.0227	0.9323 ± 0.0171	0.9555 ± 0.0150	0.9541 ± 0.0163	**0.9358**
mRMR	0.9034 ± 0.0161	0.9184 ± 0.0221	**0.9332 ± 0.0182**	0.9202 ± 0.0245	0.9344 ± 0.0188	0.9219
FD-CARI	0.8980 ± 0.0238	0.9247 ± 0.0251	0.9250 ± 0.0207	0.9516 ± 0.0168	0.9527 ± 0.0153	0.9304

Berdasarkan Jadual 4.38, kaedah Relief-F mencatatkan purata skor F1 tertinggi iaitu 0.9358, diikuti oleh FD-CARI (0.9304) dan CFS (0.9267). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model DT dengan nilai skor F1 sebanyak 0.9092 ± 0.0231 manakala kaedah CFS menunjukkan prestasi cemerlang pada model LR dengan skor F1 tertinggi sebanyak 0.9576 ± 0.0169. Kaedah Consistency dan FCBF pula masing-masing mencatatkan purata skor F1 sebanyak 0.9261 dan 0.9258 menunjukkan prestasi yang agak baik namun sedikit lebih rendah berbanding tiga kaedah utama tersebut. Sebaliknya, kaedah mRMR mencatatkan purata skor F1 terendah iaitu 0.9219.

Akhir sekali, bagi mengenal pasti kaedah pemilihan yang mencatatkan prestasi paling konsisten dan seimbang berdasarkan semua metrik penilaian, Rajah 4.15 membandingkan keseluruhan purata skor metrik lima model pengelasan bagi set data WDBC.



Rajah 4.15 Perbandingan purata skor metrik penilaian bagi keseluruhan model pengelasan

Berdasarkan Rajah 4.15, kaedah Relief-F dan FD-CARI menunjukkan prestasi yang paling konsisten dan seimbang dalam semua metrik utama. Kaedah Relief-F mencatatkan prestasi tertinggi pada ACC (0.9528), SEN (0.9217) dan F1 (0.9358) manakala FD-CARI mencatatkan prestasi kedua terbaik pada AUC (0.9857), ACC (0.9847), SEN (0.9177) dan F1 (0.9672). Walaupun kaedah FCBF mencatatkan skor AUC (0.9866) dan SPE (0.9176) tertinggi, namun skor SEN (0.9034) dan F1 (0.9258) yang rendah menunjukkan kaedah tersebut kurang cekap dalam mengenal pasti kes malignan dengan betul. Secara keseluruhan, kaedah Relief-F dan FD-CARI merupakan pilihan utama untuk membangunkan model pengelasan tumor pada set data WDBC.

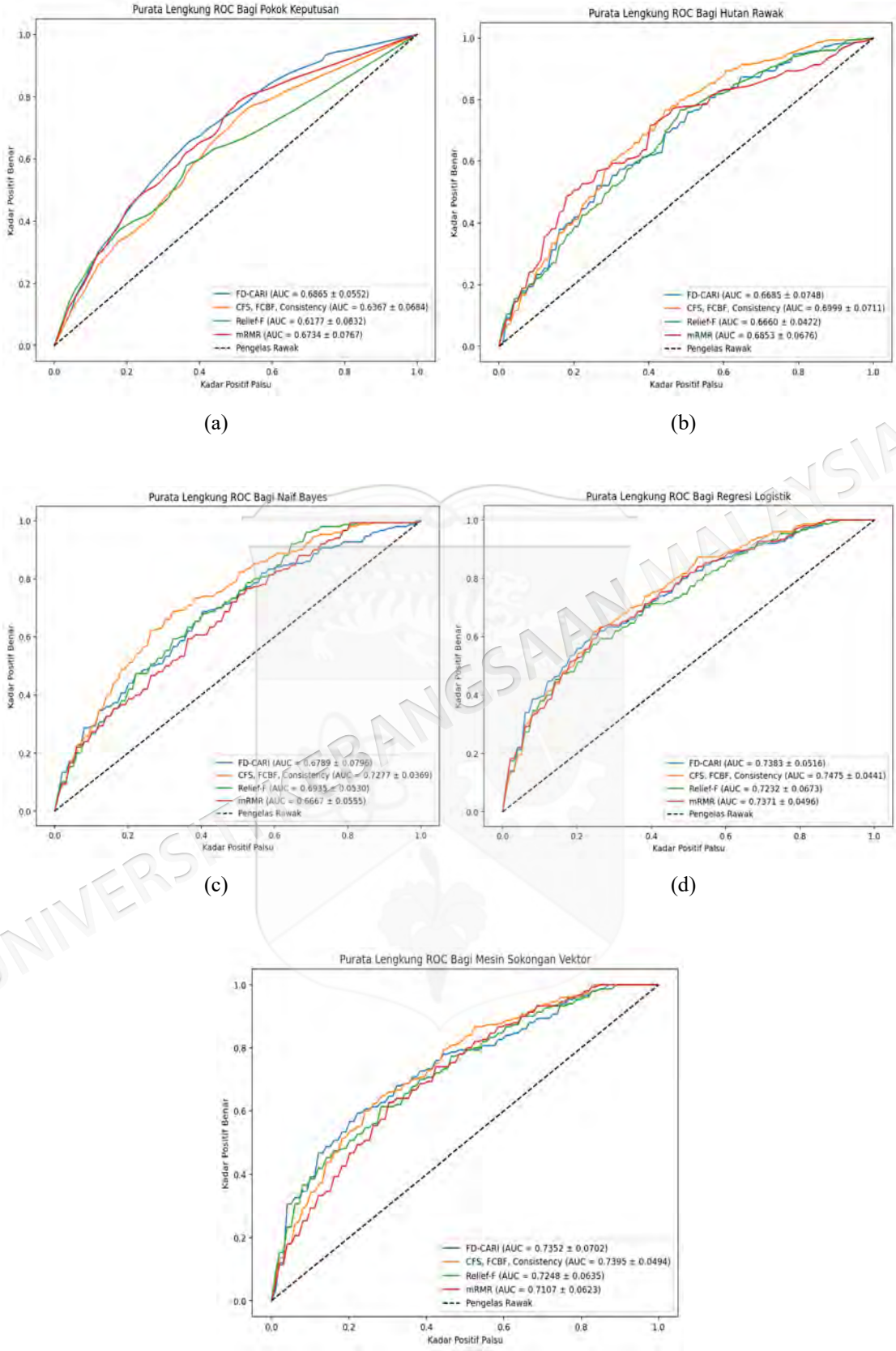
4.5.2 Set Data WPBC

Subseksyen ini membincangkan prestasi penilaian metrik AUC, ACC, SEN, SPE dan F1 untuk setiap kaedah pemilihan fitur bagi set data WPBC. Oleh disebabkan kaedah CFS, FCBF dan Consistency telah memilih subset fitur yang sama, maka prestasi bagi kaedah-kaedah tersebut adalah serupa. Jadual 4.39 mencatatkan skor AUC bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.39 Nilai skor AUC bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor AUC ± SD					Purata AUC
	DT	RF	NB	LR	SVM	
CFS	0.6367 ± 0.0721	**0.6999 ± 0.0749**	**0.7277 ± 0.0389**	**0.7475 ± 0.0465**	**0.7395 ± 0.0520**	**0.7103**
FCBF	0.6367 ± 0.0721	**0.6999 ± 0.0749**	**0.7277 ± 0.0389**	**0.7475 ± 0.0465**	**0.7395 ± 0.0520**	**0.7103**
Consistency	0.6367 ± 0.0721	**0.6999 ± 0.0749**	**0.7277 ± 0.0389**	**0.7475 ± 0.0465**	**0.7395 ± 0.0520**	**0.7103**
Relief-F	0.6177 ± 0.0877	0.6660 ± 0.0445	0.6935 ± 0.0559	0.7232 ± 0.0709	0.7248 ± 0.0669	0.6850
mRMR	0.6734 ± 0.0808	0.6853 ± 0.0712	0.6667 ± 0.0585	0.7371 ± 0.0523	0.7107 ± 0.0657	0.6946
FD-CARI	**0.6865 ± 0.0581**	0.6685 ± 0.0789	0.6789 ± 0.0839	0.7383 ± 0.0544	0.7352 ± 0.0740	0.7015

Berdasarkan Jadual 4.39, kaedah CFS, FCBF dan Consistency mencatatkan purata skor AUC tertinggi iaitu 0.7103, diikuti oleh FD-CARI (0.7015) dan mRMR (0.6946). Kaedah CFS, FCBF dan Consistency menunjukkan prestasi terbaik berbanding kaedah lain pada model RF, NB, LR dan SVM dengan nilai skor AUC masing-masing sebanyak 0.6999 ± 0.0749, 0.7277 ± 0.0389, 0.7475 ± 0.0465 dan 0.7395 ± 0.0520 manakala kaedah FD-CARI mencatatkan skor AUC tertinggi bagi DT iaitu 0.6865 ± 0.0581. Sebaliknya, kaedah Relief-F mencatatkan purata skor AUC terendah iaitu 0.6850. Rajah 4.23 memaparkan purata lengkungan ROC bagi setiap model pengelasan.



Rajah 4.16 Purata lengkungan ROC bagi model (a) DT (b) RF (c) NB (d) LR (e) SVM

Rajah 4.16 menunjukkan lengkung ROC bagi model LR dan SVM menggunakan kaedah CFS, FCBF dan Consistency serta FD-CARI jelas lebih mendekati sudut kiri yang mencerminkan prestasi yang lebih cemerlang berbanding model DT, RF dan NB.

Seterusnya, skor ACC bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur dicatatkan pada Jadual 4.40.

Jadual 4.40 Nilai skor ACC bagi setiap model pengelasan

Kaedah Pemilihan Fitur	**Nilai Skor ACC ± SD**					**Purata ACC**
	DT	**RF**	**NB**	**LR**	**SVM**	
CFS	0.6585 ± 0.0725	0.6831 ± 0.0514	0.5600 ± 0.0378	0.6308 ± 0.0459	0.6123 ± 0.0422	0.6289
FCBF	0.6585 ± 0.0725	0.6831 ± 0.0514	0.5600 ± 0.0378	0.6308 ± 0.0459	0.6123 ± 0.0422	0.6289
Consistency	0.6585 ± 0.0725	0.6831 ± 0.0514	0.5600 ± 0.0378	0.6308 ± 0.0459	0.6123 ± 0.0422	0.6289
Relief-F	0.6492 ± 0.0683	0.6954 ± 0.0331	0.5585 ± 0.0503	**0.6400 ± 0.0392**	0.6077 ± 0.0399	**0.6302**
mRMR	**0.6846 ± 0.0831**	**0.7108 ± 0.0588**	0.5708 ± 0.0505	0.5723 ± 0.0532	0.6046 ± 0.0523	0.6286
FD-CARI	0.6431 ± 0.0422	0.6723 ± 0.0696	**0.5985 ± 0.0727**	0.5769 ± 0.0504	**0.6138 ± 0.0535**	0.6209

Berdasarkan Jadual 4.40, kaedah Relief-F mencatatkan purata skor ACC tertinggi iaitu 0.6302, diikuti oleh kaedah CFS, FCBF dan Consistency (0.6289) serta mRMR (0.6286). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model LR dengan nilai skor ACC sebanyak 0.6400 ± 0.0392 manakala mRMR menunjukkan prestasi cemerlang pada model DT dan RF dengan skor ACC tertinggi masing-masing sebanyak 0.6846 ± 0.0831 dan 0.7108 ± 0.0588. Sebaliknya, kaedah FD-CARI mencatatkan purata skor ACC terendah iaitu 0.6209.

Seterusnya, Jadual 4.41 mencatatkan skor SEN bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.41 Nilai skor SEN bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor SEN ± SD					Purata SEN
	DT	RF	NB	LR	SVM	
CFS	0.4200 ± 0.1259	0.4867 ± 0.1476	**0.8067 ± 0.0858**	0.7400 ± 0.0584	**0.7733 ± 0.0783**	0.6453
FCBF	0.4200 ± 0.1259	0.4867 ± 0.1476	**0.8067 ± 0.0858**	0.7400 ± 0.0584	**0.7733 ± 0.0783**	0.6453
Consistency	0.4200 ± 0.1259	0.4867 ± 0.1476	**0.8067 ± 0.0858**	0.7400 ± 0.0584	**0.7733 ± 0.0783**	0.6453
Relief-F	0.4533 ± 0.1249	0.3800 ± 0.0996	0.7000 ± 0.1144	0.6933 ± 0.0717	0.7133 ± 0.0706	0.5880
mRMR	0.4800 ± 0.1209	0.4867 ± 0.1259	0.6733 ± 0.1016	**0.8200 ± 0.0834**	0.7600 ± 0.0843	0.6440
FD-CARI	**0.6333 ± 0.1379**	**0.5067 ± 0.1184**	0.6867 ± 0.0773	**0.8200 ± 0.0945**	0.7400 ± 0.0858	**0.6773**

Berdasarkan Jadual 4.41, kaedah FD-CARI mencatatkan purata skor SEN tertinggi iaitu 0.6773, diikuti oleh CFS, FCBF dan Consistency (0.6453) dan mRMR (0.6440). Kaedah FD-CARI menunjukkan prestasi terbaik berbanding kaedah lain pada model DT, RF dan LR dengan nilai skor SEN masing-masing sebanyak 0.6333 ± 0.1379, 0.5067 ± 0.1184 dan 0.8200 ± 0.0945 manakala kaedah FCBF, CFS dan Consistency menunjukkan prestasi cemerlang pada model NB dan SVM dengan skor SEN tertinggi masing-masing sebanyak 0.8067 ± 0.0858 dan 0.7733 ± 0.0783. Sebaliknya, kaedah Relief-F mencatatkan purata skor SEN terendah iaitu 0.5880.

Seterusnya, skor SPE bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur dicatatkan pada Jadual 4.42.

Jadual 4.42 Nilai skor SPE bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor SPE ± SD					Purata SPE
	DT	RF	NB	LR	SVM	
CFS	0.7300 ± 0.0823	0.7420 ± 0.0739	0.4860 ± 0.0534	0.5980 ± 0.0529	0.5640 ± 0.0515	0.6240
FCBF	0.7300 ± 0.0823	0.7420 ± 0.0739	0.4860 ± 0.0534	0.5980 ± 0.0529	0.5640 ± 0.0515	0.6240
Consistency	0.7300 ± 0.0823	0.7420 ± 0.0739	0.4860 ± 0.0534	0.5980 ± 0.0529	0.5640 ± 0.0515	0.6240
Relief-F	0.7080 ± 0.0732	**0.7900 ± 0.0380**	0.5160 ± 0.0685	**0.6240 ± 0.0420**	**0.5760 ± 0.0440**	**0.6428**
mRMR	**0.7460 ± 0.0900**	0.7780 ± 0.0649	0.5400 ± 0.0573	0.4980 ± 0.0649	0.5580 ± 0.0592	0.6240
FD-CARI	0.6460 ± 0.0611	0.7220 ± 0.0780	**0.5720 ± 0.0790**	0.5040 ± 0.0580	**0.5760 ± 0.0610**	0.6040

Berdasarkan Jadual 4.42, kaedah Relief-F mencatatkan purata skor SPE tertinggi iaitu 0.6428, diikuti oleh kaedah CFS, FCBF dan Consistency (0.6240) serta mRMR (0.6240). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model RF, LR dan SVM dengan nilai skor SPE masing-masing sebanyak 0.7900 ± 0.0380, 0.6240 ± 0.0420 dan 0.5760 ± 0.0440 manakala kaedah mRMR menunjukkan prestasi cemerlang pada model DT dengan skor SPE tertinggi sebanyak 0.7460 ± 0.0900. Sebaliknya, kaedah FD-CARI mencatatkan purata skor SPE terendah iaitu 0.6040.

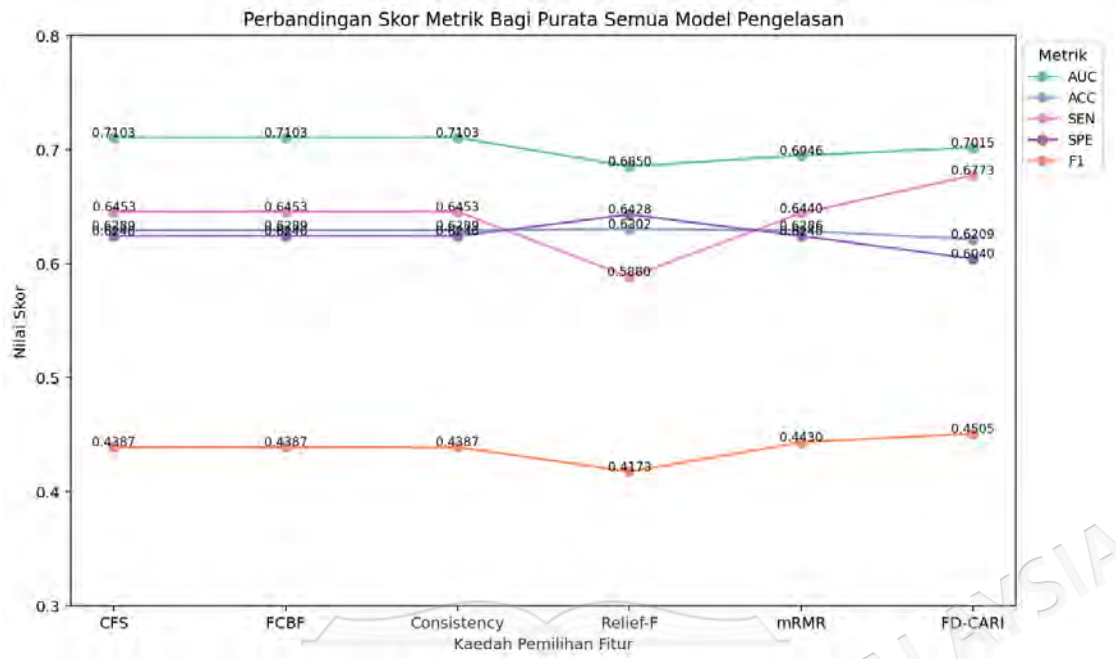
Seterusnya, Jadual 4.43 mencatatkan skor F1 bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.43 Nilai skor F1 bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor F1 ± SD					Purata F1
	DT	RF	NB	LR	SVM	
CFS	0.3635 ± 0.1057	0.4103 ± 0.0918	**0.4581 ± 0.0361**	**0.4818 ± 0.0464**	**0.4798 ± 0.0444**	0.4387
FCBF	0.3635 ± 0.1057	0.4103 ± 0.0918	**0.4581 ± 0.0361**	**0.4818 ± 0.0464**	**0.4798 ± 0.0444**	0.4387
Consistency	0.3635 ± 0.1057	0.4103 ± 0.0918	**0.4581 ± 0.0361**	**0.4818 ± 0.0464**	**0.4798 ± 0.0444**	0.4387
Relief-F	0.3744 ± 0.1039	0.3626 ± 0.0784	0.4218 ± 0.0505	0.4709 ± 0.0479	0.4567 ± 0.0449	0.4173
mRMR	0.4174 ± 0.1156	0.4363 ± 0.0983	0.4199 ± 0.0566	0.4703 ± 0.0474	0.4711 ± 0.0561	0.4430
FD-CARI	**0.4470 ± 0.0686**	**0.4180 ± 0.1001**	0.4446 ± 0.0703	0.4725 ± 0.0500	0.4704 ± 0.0553	**0.4505**

Berdasarkan Jadual 4.43, kaedah FD-CARI mencatatkan purata skor F1 tertinggi iaitu 0.4505, diikuti oleh mRMR (0.4430) serta CFS, FCBF dan Consistency (0.4387). Kaedah FD-CARI menunjukkan prestasi terbaik berbanding kaedah lain pada model DT dan RF dengan nilai skor F1 masing-masing sebanyak 0.4470 ± 0.0686 dan 0.4180 ± 0.1001 manakala kaedah CFS, FCBF dan Consistency menunjukkan prestasi cemerlang pada model NB, LR dan SVM dengan skor F1 tertinggi masing-masing sebanyak 0.4581 ± 0.0361, 0.4818 ± 0.0464 dan 0.4798 ± 0.0444. Sebaliknya, kaedah Relief-F mencatatkan purata skor F1 terendah iaitu 0.4173.

Akhir sekali, Rajah 4.17 membandingkan keseluruhan purata skor metrik lima model pengelasan bagi set data WPBC bagi mengenal pasti kaedah pemilihan fitur yang mencatatkan prestasi paling konsisten dan seimbang.



Rajah 4.17 Perbandingan purata skor metrik penilaian bagi keseluruhan model pengelasan

Berdasarkan Rajah 4.17, kaedah CFS, FCBF dan Consistency serta FD-CARI menunjukkan prestasi yang paling konsisten dan seimbang dalam semua metrik utama. Kaedah CFS, FCBF dan Consistency mencatatkan prestasi tertinggi pada AUC (0.7103) manakala FD-CARI mencatatkan prestasi tertinggi pada SEN (0.6773) dan F1 (0.4505) serta prestasi kedua terbaik pada AUC (0.7015). Walaupun kaedah Relief-F mencatatkan skor ACC (0.6302) dan SPE (0.6428) tertinggi, namun skor SEN (0.5880) dan F1 (0.4173) yang rendah menunjukkan kaedah tersebut kurang cekap dalam mengenal pasti kes kanser berulang dengan betul. Secara keseluruhan, kaedah CFS, FCBF dan Consistency serta FD-CARI merupakan pilihan utama untuk membangunkan model pengelasan pengulangan kanser pada set data WPBC.

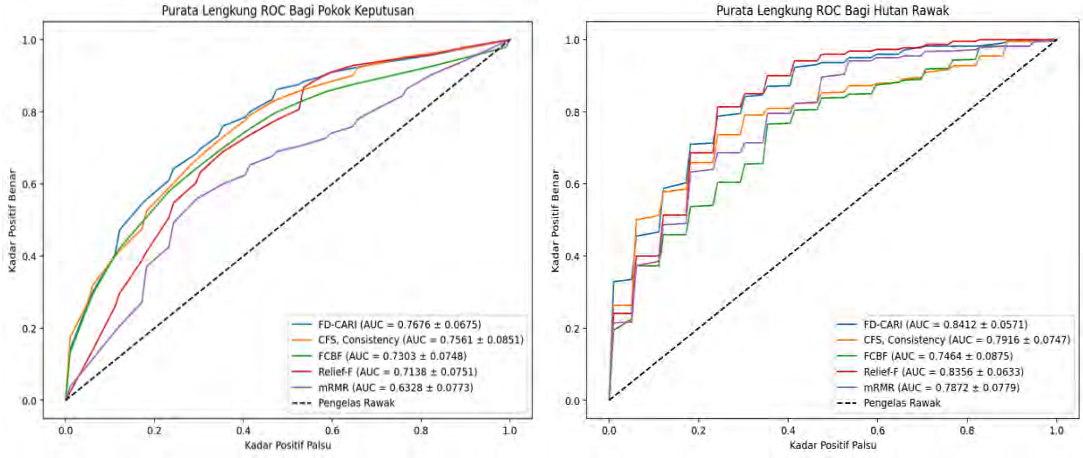
4.5.3 Set Data BCC

Subseksyen ini membincangkan prestasi penilaian metrik AUC, ACC, SEN, SPE dan F1 untuk setiap kaedah pemilihan fitur bagi set data BCC. Oleh disebabkan kaedah CFS dan Consistency telah memilih subset fitur yang sama, maka prestasi bagi kaedah-kaedah tersebut adalah serupa. Jadual 4.44 mencatatkan skor AUC bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.44 Nilai skor AUC bagi setiap model pengelasan

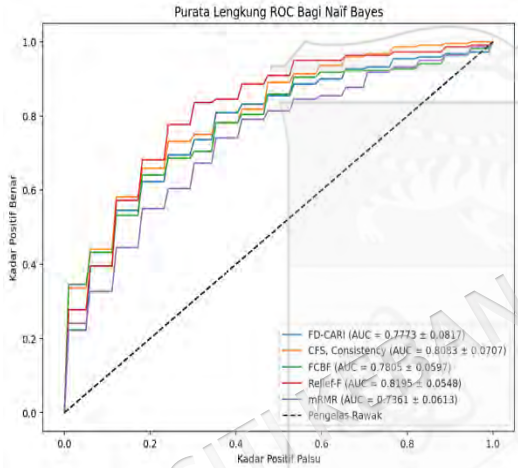
Kaedah Pemilihan Fitur	Nilai Skor AUC ± SD					Purata AUC
	DT	RF	NB	LR	SVM	
CFS	0.7561 ± 0.0897	0.7916 ± 0.0787	0.8083 ± 0.0745	0.7647 ± 0.0765	0.7870 ± 0.1566	0.7816
FCBF	0.7303 ± 0.0788	0.7464 ± 0.0923	0.7805 ± 0.0630	0.7647 ± 0.0765	0.6675 ± 0.1607	0.7379
Consistency	0.7561 ± 0.0897	0.7916 ± 0.0787	0.8083 ± 0.0745	0.7647 ± 0.0765	0.7870 ± 0.1566	0.7816
Relief-F	0.7138 ± 0.0792	0.8356 ± 0.0667	**0.8195 ± 0.0578**	**0.8003 ± 0.0765**	0.8414 ± 0.0562	0.8021
mRMR	0.6328 ± 0.0815	0.7872 ± 0.0821	0.7361 ± 0.0647	0.7840 ± 0.0816	0.7858 ± 0.0811	0.7452
FD-CARI	**0.7676 ± 0.0712**	**0.8412 ± 0.0602**	0.7773 ± 0.0861	0.7922 ± 0.0893	**0.8439 ± 0.0717**	**0.8044**

Berdasarkan Jadual 4.44, kaedah FD-CARI mencatatkan purata skor AUC tertinggi iaitu 0.8044, diikuti oleh Relief-F (0.8021) serta CFS dan Consistency (0.7816). Kaedah FD-CARI menunjukkan prestasi terbaik berbanding kaedah lain pada model DT, RF dan SVM dengan nilai skor AUC masing-masing sebanyak 0.7676 ± 0.0712, 0.8412 ± 0.0602 dan 0.8439 ± 0.0717 manakala kaedah Relief-F mencatatkan skor AUC tertinggi bagi NB dan LR iaitu 0.8195 ± 0.0578 dan 0.8003 ± 0.0765 masing-masing. Sebaliknya, kaedah mRMR dan FCBF mencatatkan purata skor AUC agak rendah masing-masing sebanyak 0.7452 dan 0.7379. Rajah 4.18 memaparkan purata lengkungan ROC bagi setiap model pengelasan.

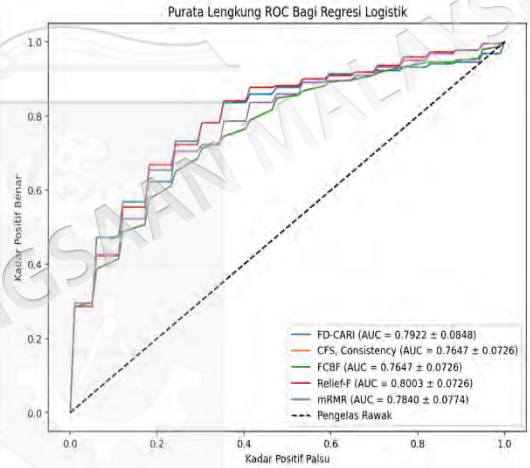


(a)

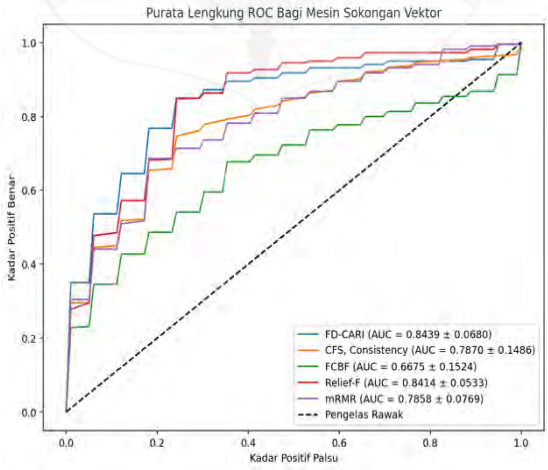
(b)



(c)



(d)



(e)

Rajah 4.18 Purata lengkungan ROC bagi model (a) DT (b) RF (c) NB (d) LR (e) SVM

Rajah 4.18 menunjukkan lengkung ROC bagi model RF dan SVM menggunakan kaedah FD-CARI dan Relief-F jelas lebih mendekati sudut kiri di mana ia mencerminkan prestasi yang lebih cemerlang berbanding model DT, NB dan LR.

Seterusnya, skor ACC bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur dicatatkan pada Jadual 4.45.

Jadual 4.45 Nilai skor ACC bagi setiap model pengelasan

Kaedah Pemilihan Fitur	**Nilai Skor ACC ± SD**					**Purata ACC**
	DT	**RF**	**NB**	**LR**	**SVM**	
CFS	0.7103 ± 0.0847	0.7231 ± 0.0827	0.5923 ± 0.0720	0.6744 ± 0.0784	0.7231 ± 0.0415	0.6846
FCBF	0.6821 ± 0.0675	0.7000 ± 0.0580	0.5718 ± 0.0617	0.6744 ± 0.0784	0.6000 ± 0.0595	0.6456
Consistency	0.7103 ± 0.0847	0.7231 ± 0.0827	0.5923 ± 0.0720	0.6744 ± 0.0784	0.7231 ± 0.0415	0.6846
Relief-F	0.7026 ± 0.0727	**0.7872 ± 0.0514**	**0.6256 ± 0.0747**	**0.7128 ± 0.0713**	**0.7923 ± 0.0372**	**0.7241**
mRMR	0.5795 ± 0.0717	0.6872 ± 0.0614	0.5897 ± 0.0554	0.7103 ± 0.0640	0.7077 ± 0.0582	0.6549
FD-CARI	**0.7256 ± 0.0663**	0.7718 ± 0.0490	0.6103 ± 0.0878	0.7103 ± 0.0746	0.7667 ± 0.0924	0.7169

Berdasarkan Jadual 4.45, kaedah Relief-F mencatatkan purata skor ACC tertinggi iaitu 0.7241, diikuti oleh kaedah FD-CARI (0.7169) serta CFS dan Consistency (0.6846). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model RF, NB, LR dan SVM dengan nilai skor ACC masing-masing sebanyak 0.7872 ± 0.0514, 0.6256 ± 0.0747, 0.7128 ± 0.0713 dan 0.7923 ± 0.0372 manakala FD-CARI menunjukkan prestasi cemerlang pada model DT dengan skor ACC tertinggi sebanyak 0. 0.7256 ± 0.0663. Sebaliknya, kaedah mRMR dan FCBF mencatatkan purata skor ACC agak rendah masing masing sebanyak 0.6549 dan 0.6456.

Seterusnya, Jadual 4.47 mencatatkan skor SEN bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.46 Nilai skor SEN bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor SEN ± SD					Purata SEN
	DT	RF	NB	LR	SVM	
CFS	**0.7682 ± 0.1163**	0.7364 ± 0.1045	0.3045 ± 0.1135	0.5409 ± 0.1202	**0.8818 ± 0.0614**	0.6464
FCBF	0.6773 ± 0.1182	0.7364 ± 0.0929	0.2591 ± 0.1006	0.5409 ± 0.1202	0.3727 ± 0.1962	0.5173
Consistency	**0.7682 ± 0.1163**	0.7364 ± 0.1045	0.3045 ± 0.1135	0.5409 ± 0.1202	**0.8818 ± 0.0614**	0.6464
Relief-F	0.7545 ± 0.1635	**0.8182 ± 0.0742**	**0.3955 ± 0.1094**	0.6273 ± 0.1170	0.8636 ± 0.0678	**0.6918**
mRMR	0.5500 ± 0.1015	0.6909 ± 0.1045	0.3682 ± 0.1102	0.6455 ± 0.0878	0.6136 ± 0.0780	0.5736
FD-CARI	0.7227 ± 0.1202	0.8000 ± 0.0963	0.3909 ± 0.1425	**0.6636 ± 0.0987**	0.7273 ± 0.1389	0.6609

Berdasarkan Jadual 4.46, kaedah Relief-F mencatatkan purata skor SEN tertinggi iaitu 0.6918, diikuti oleh FD-CARI (0.6609) serta CFS dan Consistency (0.6464). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model RF dan NB dengan nilai skor SEN masing-masing sebanyak 0.8182 ± 0.0742 dan 0.3955 ± 0.1094 manakala kaedah FD-CARI menunjukkan prestasi cemerlang pada model LR dengan skor SEN tertinggi sebanyak 0.6636 ± 0.0987. Kaedah CFS dan Consistency juga menunjukkan prestasi terbaik pada model SVM sebanyak 0.8818 ± 0.0614. Sebaliknya, kaedah mRMR dan FCBF mencatatkan purata skor SEN agak rendah masing masing sebanyak 0.5736 dan 0.5173.

Seterusnya, skor SPE bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur dicatatkan pada Jadual 4.47.

Jadual 4.47 Nilai skor SPE bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor SPE ± SD					Purata SPE
	DT	RF	NB	LR	SVM	
CFS	0.6353 ± 0.1588	0.7059 ± 0.1271	0.9647 ± 0.0411	**0.8471 ± 0.0568**	0.5176 ± 0.0953	0.7341
FCBF	0.6882 ± 0.1272	0.6529 ± 0.1054	**0.9765 ± 0.0411**	**0.8471 ± 0.0568**	**0.8941 ± 0.1917**	**0.8118**
Consistency	0.6353 ± 0.1588	0.7059 ± 0.1271	0.9647 ± 0.0411	**0.8471 ± 0.0568**	0.5176 ± 0.0953	0.7341
Relief-F	0.6353 ± 0.1462	**0.7471 ± 0.0922**	0.9235 ± 0.0623	0.8235 ± 0.0679	0.7000 ± 0.0757	0.7659
mRMR	0.6176 ± 0.1865	0.6824 ± 0.1081	0.8765 ± 0.0434	0.7941 ± 0.0971	0.8294 ± 0.0704	0.7600
FD-CARI	**0.7294 ± 0.1446**	0.7353 ± 0.0635	0.8941 ± 0.0723	0.7706 ± 0.1158	0.8176 ± 0.0515	0.7894

Berdasarkan Jadual 4.47, kaedah FCBF mencatatkan purata skor SPE tertinggi iaitu 0.8118, diikuti oleh kaedah FD-CARI (0.7894) dan Relief-F (0.7659). Kaedah FCBF menunjukkan prestasi terbaik berbanding kaedah lain pada model NB, LR dan SVM dengan nilai skor SPE masing-masing sebanyak 0.9765 ± 0.0411, 0.8471 ± 0.0568 dan 0.8941 ± 0.1917 manakala kaedah FD-CARI menunjukkan prestasi cemerlang pada model DT dengan skor SPE tertinggi sebanyak 0.7294 ± 0.1446. Kaedah Relief-F juga menunjukkan prestasi terbaik pada model RF sebanyak 0.7471 ± 0.0922. Sebaliknya, kaedah mRMR serta CFS dan Consistency mencatatkan purata skor SPE agak rendah masing masing sebanyak 0.7600 dan 0.7341.

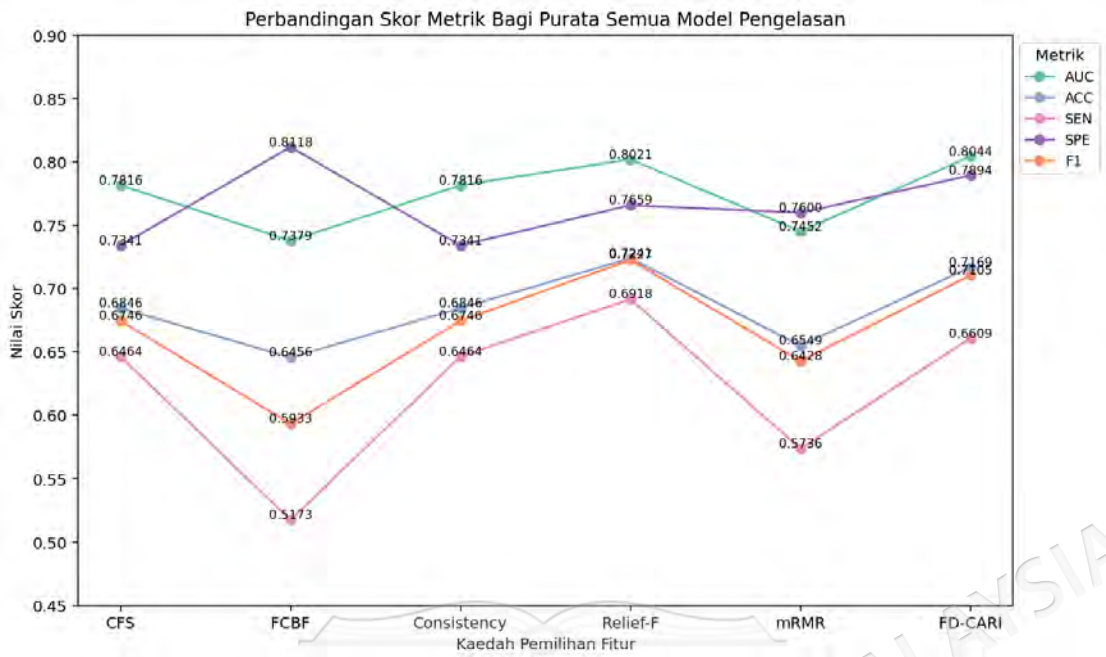
Seterusnya, Jadual 4.48 mencatatkan skor F1 bagi setiap model pengelasan untuk setiap kaedah pemilihan fitur.

Jadual 4.48 Nilai skor F1 bagi setiap model pengelasan

Kaedah Pemilihan Fitur	Nilai Skor F1 ± SD					Purata F1
	DT	RF	NB	LR	SVM	
CFS	**0.7479 ± 0.0761**	0.7484 ± 0.0787	0.4484 ± 0.1326	0.6462 ± 0.1065	0.7820 ± 0.0319	0.6746
FCBF	0.7025 ± 0.0738	0.7331 ± 0.0559	0.3973 ± 0.1251	0.6462 ± 0.1065	0.4875 ± 0.1475	0.5933
Consistency	**0.7479 ± 0.0761**	0.7484 ± 0.0787	0.4484 ± 0.1326	0.6462 ± 0.1065	0.7820 ± 0.0319	0.6746
Relief-F	0.7342 ± 0.0878	**0.8119 ± 0.0468**	**0.5372 ± 0.1194**	0.7067 ± 0.0866	**0.8236 ± 0.0340**	**0.7227**
mRMR	0.5936 ± 0.0678	0.7110 ± 0.0672	0.4944 ± 0.1126	0.7136 ± 0.0694	0.7015 ± 0.0670	0.6428
FD-CARI	0.7450 ± 0.0758	0.7959 ± 0.0533	0.5196 ± 0.1427	**0.7191 ± 0.0799**	0.7731 ± 0.1011	0.7105

Berdasarkan Jadual 4.48, kaedah Relief-F mencatatkan purata skor F1 tertinggi iaitu 0.7227, diikuti oleh FD-CARI (0.7105) serta CFS dan Consistency (0.6746). Kaedah Relief-F menunjukkan prestasi terbaik berbanding kaedah lain pada model RF, NB dan LR dengan nilai skor F1 masing-masing sebanyak 0.8119 ± 0.0468, 0.5372 ± 0.1194 dan 0.8236 ± 0.0340 manakala kaedah FD-CARI menunjukkan prestasi cemerlang pada model LR dengan skor F1 tertinggi sebanyak 0.7191 ± 0.0799. Kaedah CFS dan Consistency juga menunjukkan prestasi terbaik pada model DT sebanyak 0.7479 ± 0.0761. Sebaliknya, kaedah mRMR dan FCBF mencatatkan purata skor SPE agak rendah masing masing sebanyak 0.6428 dan 0.5933.

Akhir sekali, bagi mengenal pasti kaedah pemilihan fitur yang mencatatkan prestasi paling konsisten dan seimbang berdasarkan semua metrik penilaian, Rajah 4.19 membandingkan keseluruhan purata skor metrik lima model pengelasan bagi set data BCC.



Rajah 4.19 Perbandingan purata skor metrik penilaian bagi keseluruhan model pengelasan

Berdasarkan Rajah 4.19, kaedah Relief-F dan FD-CARI menunjukkan prestasi yang paling konsisten dan seimbang dalam semua metrik utama. Kaedah Relief-F mencatatkan prestasi tertinggi pada ACC (0.8021), SEN (0.6198) dan F1 (0.7227) manakala FD-CARI mencatatkan prestasi tertinggi pada AUC (0.8044) dan serta prestasi kedua terbaik pada ACC (0.7169), SEN (0.6609), SPE (0.7894) dan F1 (0.7105). Walaupun kaedah FCBF mencatatkan skor SPE (0.8118) tertinggi, namun skor ACC (0.6456), SEN (0.5173) dan F1 (0.5933) yang rendah menunjukkan kaedah tersebut kurang cekap dalam mengenal pasti pesakit kanser dengan betul. Secara keseluruhan, kaedah Relief-F dan FD-CARI merupakan pilihan utama untuk membangunkan model pengelasan pesakit kanser pada set data BCC.

4.6 UJIAN SIGNIFIKAN

Analisis statistik dijalankan untuk menilai sama ada terdapat perbezaan signifikan antara prestasi kaedah FD-CARI dan kaedah pemilihan fitur terbaik bagi setiap model pengelasan. Hipotesis nul (H_0) menyatakan bahawa tiada perbezaan signifikan antara kaedah FD-CARI dan kaedah FS terbaik, manakala hipotesis alternatif (H_1) menyatakan bahawa terdapat perbezaan signifikan antara kedua-duanya. Ujian *Wilcoxon Signed-Rank* atau *t*-berpasangan digunakan bergantung pada jenis taburan

data dan keputusan ditunjukkan melalui nilai *p*. Jika nilai *p* kurang daripada 0.05, ia menunjukkan perbezaan yang signifikan. Jika nilai *p* sama atau lebih daripada 0.05, maka tiada perubahan signifikan dan prestasi FD-CARI adalah setanding dengan kaedah terbaik.

Skor AUC digunakan sebagai metrik perbandingan ini kerana ia memberikan gambaran menyeluruh terhadap kebolehan model dalam membezakan antara kelas positif dan negatif. Lajur "Signifikan" dalam jadual menunjukkan sama ada perbezaan itu signifikan (+) atau tidak (-).

4.6.1 Set Data WDBC

Jadual 4.49 mencatatkan skor AUC kaedah FD-CARI dan kaedah terbaik bagi setiap model pengelasan bagi set data WDBC.

Jadual 4.49 Skor AUC bagi setiap model pengelasan

Model	Kaedah FD-CARI Nilai AUC ± SD	Kaedah Terbaik	Nilai AUC ± SD
DT	0.9683 ± 0.0115	FD-CARI	0.9683 ± 0.0115
RF	0.9853 ± 0.0057	FCBF	0.9896 ± 0.0056
NB	0.9911 ± 0.0049	Relief-F	0.9917 ± 0.0049
LR	0.9919 ± 0.0049	Consistency	0.9944 ± 0.0031
SVM	0.9917 ± 0.0049	Consistency	0.9946 ± 0.0027

Berdasarkan Jadual 4.49, kaedah Consistency mencatatkan prestasi terbaik bagi model LR dan SVM manakala FCBF terbaik pada model RF. Model NB pula didahului skor AUC oleh kaedah Relief-F dan kaedah FD-CARI hanya terbaik pada model DT. Justeru itu, prestasi kaedah FD-CARI diuji pada keempat-empat model iaitu RF, NB, LR dan SVM.

Jadual 4.50 menunjukkan keputusan ujian *t*-berpasangan berdasarkan nilai *p* dan penjelasan prestasi bagi kaedah yang diuji.

Jadual 4.50 Hasil ujian signifikan pada kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan Prestasi
RF	0.0213	(+)	FD-CARI < FCBF
NB	0.1377	(-)	FD-CARI = Relief-F
LR	0.0270	(+)	FD-CARI < Consistency
SVM	0.0269	(+)	FD-CARI < Consistency

Jadual 4.50 menunjukkan model RF mencatatkan perbezaan signifikan dengan nilai *p* sebanyak 0.0213 di mana ia menunjukkan bahawa kaedah FD-CARI mempunyai prestasi yang lebih rendah berbanding FCBF. Model NB mencatatkan nilai *p* sebanyak 0.1377 menunjukkan tiada perbezaan signifikan yang menunjukkan prestasi kaedah FD-CARI setanding dengan Relief-F. Bagi model LR dan SVM, nilai *p* masing-masing 0.0270 dan 0.0269 mencatatkan perbezaan signifikan menunjukkan prestasi kaedah FD-CARI lebih rendah daripada kaedah Consistency pada kedua-dua model.

Hasil signifikan menunjukkan bahawa FD-CARI tidak dapat menandingi kaedah terbaik mungkin disebabkan oleh kaedah tersebut memilih lebih banyak fitur berbanding FD-CARI. Oleh itu, analisis tambahan dilakukan dengan membandingkan FD-CARI dengan selain kaedah terbaik. Jadual 4.51 mencatatkan keputusan ujian *t*-berpasangan berdasarkan nilai *p* dan penjelasan prestasi bagi kaedah yang diuji.

Jadual 4.51 Hasil ujian signifikan pada selain kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan
RF	0.9775	(-)	FD-CARI = CFS
	0.4427	(-)	FD-CARI = Consistency
	0.0717	(-)	FD-CARI = Relief-F
	0.0015	(+)	FD-CARI > mRMR
LR	0.9838	(-)	FD-CARI = CFS
	0.7672	(-)	FD-CARI = FCBF
	0.0435	(+)	FD-CARI > Relief-F
	0.0000	(+)	FD-CARI > mRMR
SVM	0.0503	(-)	FD-CARI = CFS
	0.1735	(-)	FD-CARI = FCBF
	0.7337	(-)	FD-CARI = Relief-F
	0.0075	(+)	FD-CARI = mRMR

Berdasarkan Jadual 4.51, kaedah FD-CARI mempunyai prestasi yang lebih baik berbanding mRMR dalam semua model RF, LR dan SVM. Kaedah FD-CARI juga lebih baik daripada Relief-F dalam model LR. Namun, prestasi kaedah FD-CARI setanding dengan CFS, Consistency dan Relief-F dalam model RF serta dengan CFS dan FCBF dalam model LR. Bagi model SVM pula, kaedah FD-CARI setanding dengan ketiga-tiga kaedah CFS, FCBF dan Relief-F. Secara keseluruhan, FD-CARI menunjukkan prestasi yang kompetitif dan lebih baik daripada mRMR dalam semua model serta lebih baik daripada kaedah Relief-F dalam model LR.

4.6.2 Set Data WPBC

Jadual 4.52 mencatatkan skor AUC kaedah FD-CARI dan kaedah terbaik bagi setiap model pengelasan pada set data WPBC.

Jadual 4.52 Skor AUC bagi setiap model pengelasan

Model	Kaedah FD-CARI Nilai AUC ± SD	Kaedah Terbaik	Nilai AUC ± SD
DT	0.6865 ± 0.0581	FD-CARI	0.6865 ± 0.0581
RF	0.6685 ± 0.0789	CFS, FCBF, Consistency	0.6999 ± 0.0749
NB	0.6789 ± 0.0839	CFS, FCBF, Consistency	0.7277 ± 0.0389
LR	0.7383 ± 0.0544	CFS, FCBF, Consistency	0.7475 ± 0.0465
SVM	0.7352 ± 0.0740	CFS, FCBF, Consistency	0.7395 ± 0.0520

Berdasarkan Jadual 4.52, kaedah CFS, FCBF dan Consistency mencatatkan prestasi terbaik bagi model RF, NB, LR dan SVM manakala FD-CARI terbaik pada model DT. Justeru itu, prestasi kaedah FD-CARI diuji pada keempat-empat model iaitu RF, NB, LR dan SVM.

Jadual 4.53 menunjukkan keputusan ujian *t*-berpasangan berdasarkan nilai *p* dan penjelasan prestasi bagi kaedah yang diuji.

Jadual 4.53 Hasil ujian signifikan pada kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan Prestasi
RF	0.2677	(-)	FD-CARI = CFS, FCBF, Consistency
NB	0.0877	(-)	FD-CARI = CFS, FCBF, Consistency
LR	0.3159	(-)	FD-CARI = CFS, FCBF, Consistency
SVM	0.7607	(-)	FD-CARI =CFS, FCBF, Consistency

Jadual 4.53 menunjukkan model RF, NB, LR dan SVM mencatatkan tiada perbezaan signifikan dengan nilai *p* masing-masing sebanyak 0.2677, 0.0877, 0.3159 dan 0.7607 di mana ia menunjukkan bahawa kaedah FD-CARI mempunyai prestasi setanding dengan kaedah CFS, FCBF dan Consistency.

4.6.3 Set Data BCC

Jadual 4.54 mencatatkan skor AUC kaedah FD-CARI dan kaedah terbaik bagi setiap model pengelasan pada set data BCC.

Jadual 4.54 Skor AUC bagi setiap model pengelasan

Model Pengelasan	Nilai AUC ± SD (Model FD-CARI)	Kaedah Pemilihan Fitur Terbaik	Nilai AUC ± SD
DT	0.7676 ± 0.0712	FD-CARI	0.7676 ± 0.0712
RF	0.8412 ± 0.0602	FD-CARI	0.8412 ± 0.0602
NB	0.7773 ± 0.0861	Relief-F	0.8195 ± 0.0578
LR	0.7922 ± 0.0893	Relief-F	0.8003 ± 0.0765
SVM	0.8439 ± 0.0717	FD-CARI	0.8439 ± 0.0717

Berdasarkan Jadual 4.54, kaedah FD-CARI mencatatkan prestasi terbaik bagi model DT, RF dan SVM manakala Relief-F terbaik pada model NB dan LR. Justeru itu, prestasi kaedah FD-CARI diuji pada kedua-dua model iaitu NB dan LR.

Jadual 4.55 menunjukkan keputusan ujian *t*-berpasangan berdasarkan nilai *p* dan penjelasan prestasi bagi kaedah yang diuji.

Jadual 4.55 Hasil ujian signifikan pada kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan
NB	0.0687	(-)	FD-CARI = Relief-F
LR	0.7390	(-)	FD-CARI = Relief-F

Jadual 4.55 menunjukkan model NB dan LR mencatatkan tiada perbezaan signifikan dengan nilai *p* masing-masing sebanyak 0.0687 dan 0.7390 di mana ia menunjukkan bahawa kaedah FD-CARI mempunyai prestasi setanding dengan kaedah Relief-F.

4.7 PERBINCANGAN

Bab ini membentangkan keseluruhan penemuan bagi Objektif 1 iaitu mengenal pasti subset fitur yang relevan, berinteraksi dan tidak bertindih menggunakan kaedah FD-CARI berasaskan pendiskretan kabur dan perlombongan peraturan kesatuan kelas. Melalui pendekatan berfasa, kaedah ini bermula dengan pra-pemprosesan data yang merangkumi pembersihan, penyeimbangan dan penormalan data. Seterusnya, fitur-fitur ditapis melalui tiga kaedah penarafan dan digabungkan untuk membentuk calon subset fitur awal yang relevan. Proses pemangkasan seterusnya menghasilkan subset fitur yang lebih padat dengan menyingkirkan fitur bertindih. Hasil akhir ialah subset fitur yang lebih kecil tetapi mempunyai maklumat interaksi yang penting.

Prestasi setiap subset fitur ini dinilai melalui pengujian dengan lima model pembelajaran mesin (DT, RF, NB, LR, SVM). Bagi set data WDBC, subset terbaik ($\alpha = 0.25$, $\beta = 0.80$) mengandungi hanya 4 fitur namun mencatat purata skor AUC setinggi 0.9857, dengan model LR dan SVM mencapai skor melebihi 0.99. Dalam WPBC, subset fitur terbaik terdiri daripada 5 fitur dan menghasilkan purata skor AUC yang tinggi iaitu 0.7015. Manakala dalam BCC, 4 fitur menghasilkan purata skor AUC tertinggi 0.8044.

Hasil ini membuktikan bahawa kaedah FD-CARI mampu menghasilkan subset fitur yang bukan sahaja lebih padat, malah berkeupayaan mengekalkan prestasi pengelasan yang tinggi.

4.8 KESIMPULAN

Secara keseluruhan, hasil daripada seksyen 4.1 hingga 4.5 menunjukkan bahawa kaedah FD-CARI telah mencapai matlamat utama bagi objektif kajian, iaitu menyaring dan memilih subset fitur yang bermakna dan tidak bertindih sambil mengekalkan atau meningkatkan prestasi pengelasan pelbagai model pembelajaran mesin dalam tiga set data yang berbeza.

BAB V

PENGESAHAN FITUR BERINTERAKSI BERDASARKAN NILAI SHAP DAN PENGUNDIAN MAJORITI BERWAJARAN

5.1 PENGENALAN

Bab ini membentangkan hasil keputusan bagi Objektif 2 yang bertujuan untuk mengesahkan bahawa subset fitur yang dipilih oleh kaedah FD-CARI bukan sahaja relevan tidak bertindih, tetapi turut mempunyai hubungan interaksi yang signifikan. Bagi mencapai objektif ini, kaedah pengesahan berdasarkan nilai interaksi SHAP digunakan untuk mengenal pasti interaksi dua arah yang bermakna antara fitur. Selain itu, teknik pengundian majoriti berwajaran turut digunakan untuk mengesahkan kestabilan keputusan yang dihasilkan model. Seterusnya, fitur yang dikenal pasti sebagai penting melalui kaedah FD-CARI sebelum penyingkiran pertindihan dibandingkan dengan pasangan interaksi SHAP untuk memastikan sama ada pemilihan tersebut melibatkan fitur yang benar-benar berinteraksi. Ini penting bagi memastikan subset yang dipilih bukan sahaja padat dan tidak bertindih, malah menyumbang secara berinteraksi terhadap prestasi pengelasan.

5.2 HASIL ANALISIS INTERAKSI FITUR BERDASARKAN NILAI SHAP

5.2.1 Penilaian Awal Prestasi Model Pengelasan Sebelum Penjanaan Nilai SHAP

Bagi membolehkan analisis interaksi fitur yang menyeluruh dijalankan menggunakan model SHAP, kajian ini terlebih dahulu menilai prestasi tiga model pembelajaran mesin berasaskan pokok iaitu RF, *Gradient Boosting* (GB) dan *XGBoost* (XGB). Ketiga-tiga model ini dipilih kerana sifatnya yang serasi dengan algoritma SHAP serta keupayaannya menjana skor kebolehpercayaan interpretasi melalui nilai kepentingan dan interaksi fitur. Penilaian skor AUC dan lengkung ROC dilakukan untuk

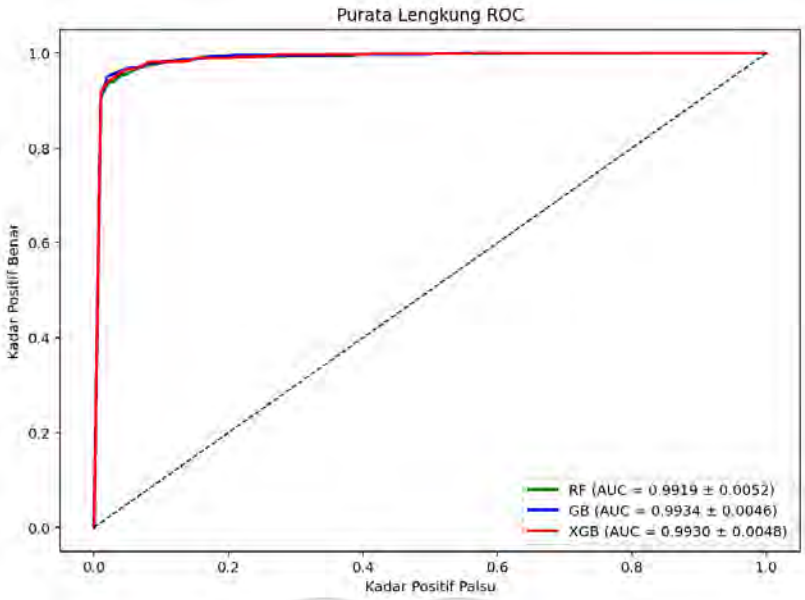
memastikan bahawa model-model ini mencapai prestasi pengelasan yang stabil dan konsisten.

Jadual 5.1 menunjukkan purata skor AUC bersama sisihan piawai (SD) bagi setiap model pengelasan untuk setiap set data. Untuk set data WDBC, model GB mencatatkan skor AUC tertinggi iaitu 0.9934 ± 0.0048, diikuti rapat oleh XGB (0.9930 ± 0.0051) dan RF (0.9919 ± 0.0055). Untuk set data WPBC pula, model XGB mencatatkan AUC tertinggi iaitu 0.6611 ± 0.0713, diikuti oleh GB (0.6550 ± 0.0711) dan RF (0.6412 ± 0.0692). Manakala, untuk set data BCC, model XGB mencatatkan AUC tertinggi iaitu 0.8149 ± 0.0745, diikuti oleh GB (0.8099 ± 0.0584) dan RF (0.8085 ± 0.0628). Walaupun perbezaan prestasi antara model adalah kecil atau hampir setara, variasi prestasi ini tetap penting kerana ia mempengaruhi penetapan pemberat dalam proses pengundian majoriti berwajaran yang digunakan dalam penarafan fitur berdasarkan nilai SHAP bagi mengenal pasti fitur kurang relevan.

Jadual 5.1 Skor AUC bagi setiap model pengelasan bagi set data WDBC

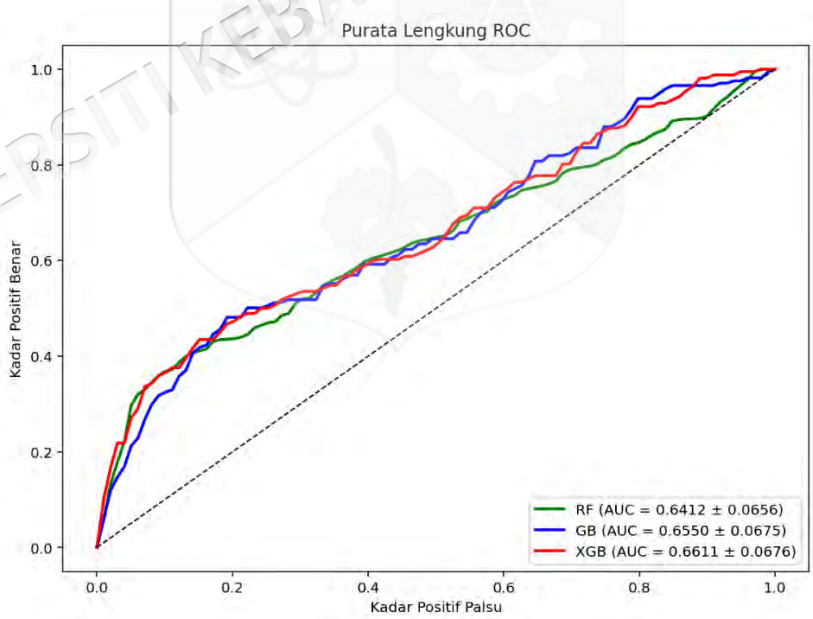
	WDBC	WPBC	BCC
Model Pengelasan	**Purata skor AUC ± SD (10 percubaan)**		
RF	0.9919 ± 0.0055	0.6412 ± 0.0692	0.8085 ± 0.0628
GB	0.9934 ± 0.0048	0.6550 ± 0.0711	0.8099 ± 0.0584
XGB	0.9930 ± 0.0051	0.6611 ± 0.0713	0.8149 ± 0.0745

Rajah 5.1 menunjukkan purata lengkung ROC bagi setiap model pengelasan untuk set data WDBC di mana model GB menunjukkan prestasi terbaik apabila lengkung ROC paling hampir ke sudut kiri atas yang menandakan keupayaan pengelasan yang lebih cemerlang berbanding model RF dan XGB.



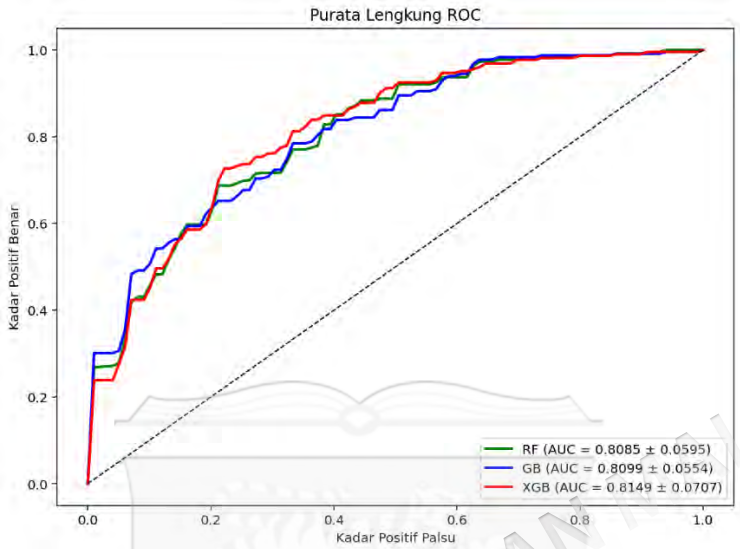
Rajah 5.1 Purata lengkung ROC bagi set data WDBC

Bagi set data WPBC, Rajah 5.2 menunjukkan bahawa model XGB pula mencatatkan lengkung ROC yang paling hampir ke sudut kiri atas yang mencerminkan prestasi pengelasan yang lebih baik berbanding model lain.



Rajah 5.2 Purata lengkung ROC bagi set data WPBC

Rajah 5.3 turut mengukuhkan dapatan tersebut bagi set data BCC apabila model XGB sekali lagi menunjukkan lengkung ROC yang paling hampir ke sudut kiri atas yang menegaskan keunggulannya dari segi prestasi pengelasan berbanding GB dan RF.



Rajah 5.3 Purata lengkung ROC bagi set data BCC

5.2.2 Penjanaan Nilai Kepentingan Fitur Menggunakan SHAP

Selepas melatih ketiga-tiga model pengelasan, nilai kepentingan setiap fitur dinilai menggunakan model SHAP. Model ini digunakan kerana ia menyediakan justifikasi yang boleh dijelaskan bagi sumbangan setiap fitur terhadap ramalan model serta dapat menilai interaksi antara fitur yang selari dengan objektif kajian untuk mengenal pasti fitur yang benar-benar relevan, berinteraksi dan tidak bertindih.

Dalam konteks ini, nilai kepentingan SHAP dikira untuk setiap fitur merentas ketiga-tiga model. Nilai purata SHAP bagi setiap fitur kemudian digunakan sebagai asas dalam proses pengundian majoriti berwajaran. Pemberat diberikan berdasarkan prestasi purata AUC setiap model seperti yang ditunjukkan dalam Jadual 5.1. Ini bertujuan bagi memastikan model yang lebih tepat memberikan sumbangan yang lebih besar terhadap proses penaratan fitur. Langkah ini juga membolehkan pemilihan fitur yang konsisten berpengaruh merentas model, tidak berat sebelah kepada mana-mana model tunggal dan lebih telus dari segi interpretasi hasil pemilihan fitur. Nilai analisis

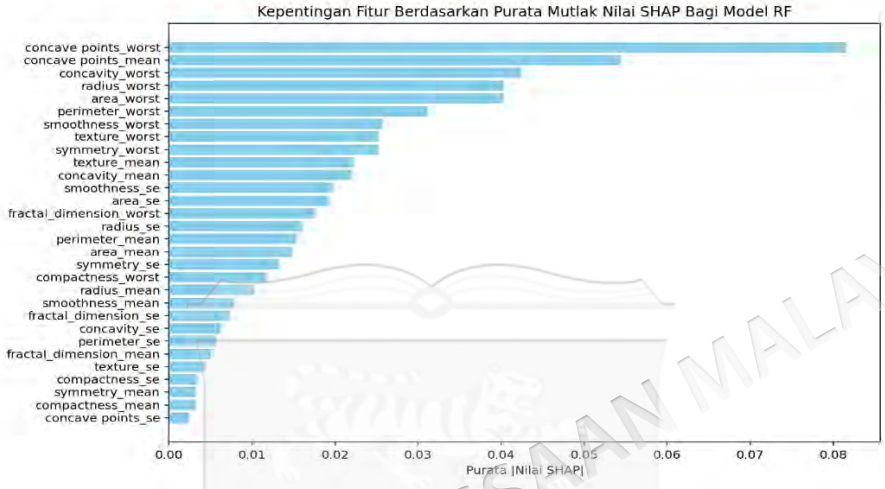
interaksi juga digunakan bagi menilai sama ada fitur-fitur yang dipilih benar-benar menunjukkan hubungan interaktif yang signifikan dalam konteks pengelasan tumor.

Berdasarkan Jadual 5.2, terdapat beberapa fitur dalam set data WDBC menunjukkan nilai SHAP tertinggi pada ketiga-tiga model termasuk *concave_points_worst*, *texture_worst*, *concavity_worst* dan *concave_points_mean*.

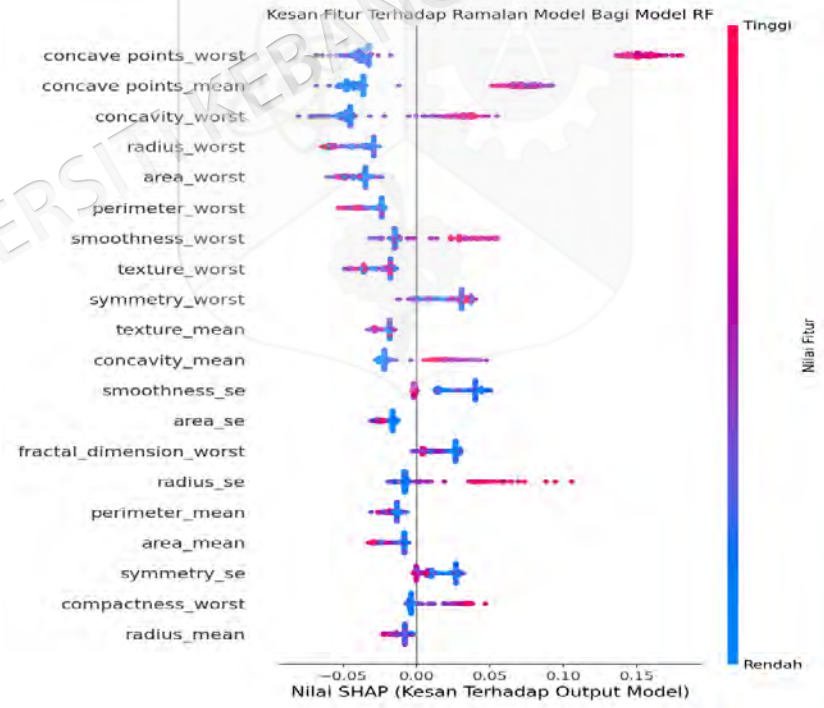
Jadual 5.2 Purata nilai SHAP bagi setiap model pengelasan bagi set data WDBC

Fitur	**Model Pengelasan**		
	RF	**GB**	**XGB**
radius_mean	0.0103	0.3605	0.0392
texture_mean	0.0222	1.6795	0.3303
perimeter_mean	0.0154	0.3047	0.0000
area_mean	0.0148	0.1201	0.0163
smoothness_mean	0.0079	0.2886	0.1404
compactness_mean	0.0032	0.1658	0.3497
concavity_mean	0.0221	0.7364	0.2220
concave points_mean	**0.0544**	**0.9032**	**0.7857**
symmetry_mean	0.0032	0.3281	0.0587
fractal_dimension_mean	0.0050	0.3384	0.1880
radius_se	0.0160	0.5615	0.1610
texture_se	0.0044	2.0803	0.0286
perimeter_se	0.0057	0.3486	0.0424
area_se	0.0192	0.8515	0.4564
smoothness_se	0.0197	1.2024	0.0728
compactness_se	0.0034	0.2070	0.5652
concavity_se	0.0063	0.5009	0.0000
concave points_se	0.0024	0.2610	0.4975
symmetry_se	0.0133	0.6959	0.4598
fractal_dimension_se	0.0072	0.8430	0.0288
radius_worst	0.0404	1.4391	0.2481
texture_worst	**0.0252**	**1.6548**	**1.9461**
perimeter_worst	0.0312	0.6925	0.5453
area_worst	0.0403	0.6501	1.3568
smoothness_worst	0.0258	1.6848	0.6183
compactness_worst	0.0118	0.6519	0.0546
concavity_worst	**0.0425**	**1.5117**	**0.7622**
concave points_worst	**0.0816**	**2.6828**	**0.7727**
symmetry_worst	0.0252	0.8491	0.6780
fractal_dimension_worst	0.0176	0.1313	0.2363

Rajah 5.4(a) menunjukkan fitur *concave points_worst*, *concave points_mean* dan *concavity_worst* paling penting dalam menentukan keputusan model RF. Rajah 5.4(b) juga menunjukkan bahawa bagi nilai tinggi bagi fitur-fitur ini cenderung kepada pengelasan sebagai tumor malignan. Sebaliknya, fitur *compactness_mean* dan *concave point_se* tidak menunjukkan sumbangan tinggi pada keputusan model RF.



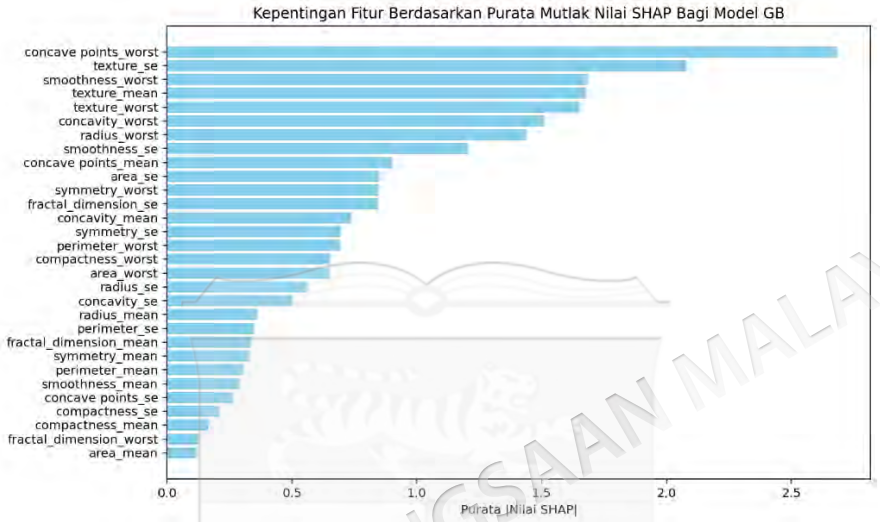
(a)



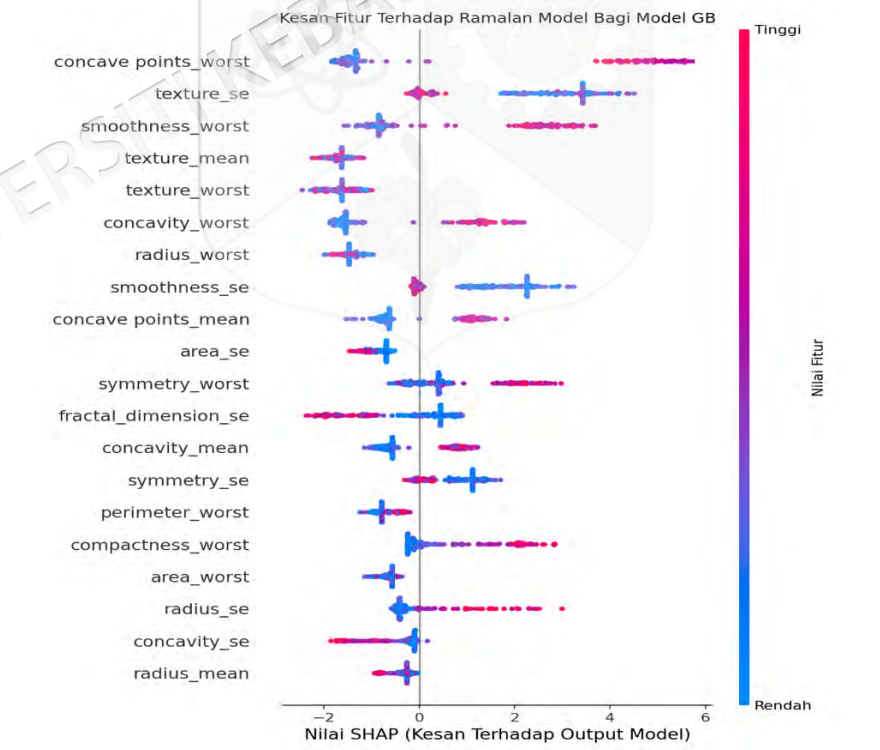
(b)

Rajah 5.4 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model RF

Rajah 5.5(a) menunjukkan fitur *concave points_worst*, *texture_se* dan *smoothness_worst* paling penting dalam menentukan keputusan model GB. Rajah 5.5(b) juga menunjukkan bahawa nilai tinggi bagi fitur *concave points_worst* dan *smoothness_worst* serta nilai rendah fitur *texture_se* cenderung kepada pengelasan tumor malignan. Sebaliknya, fitur *fractal_dimension_worst* dan *area_mean* tidak menunjukkan sumbangan tinggi pada keputusan model GB.



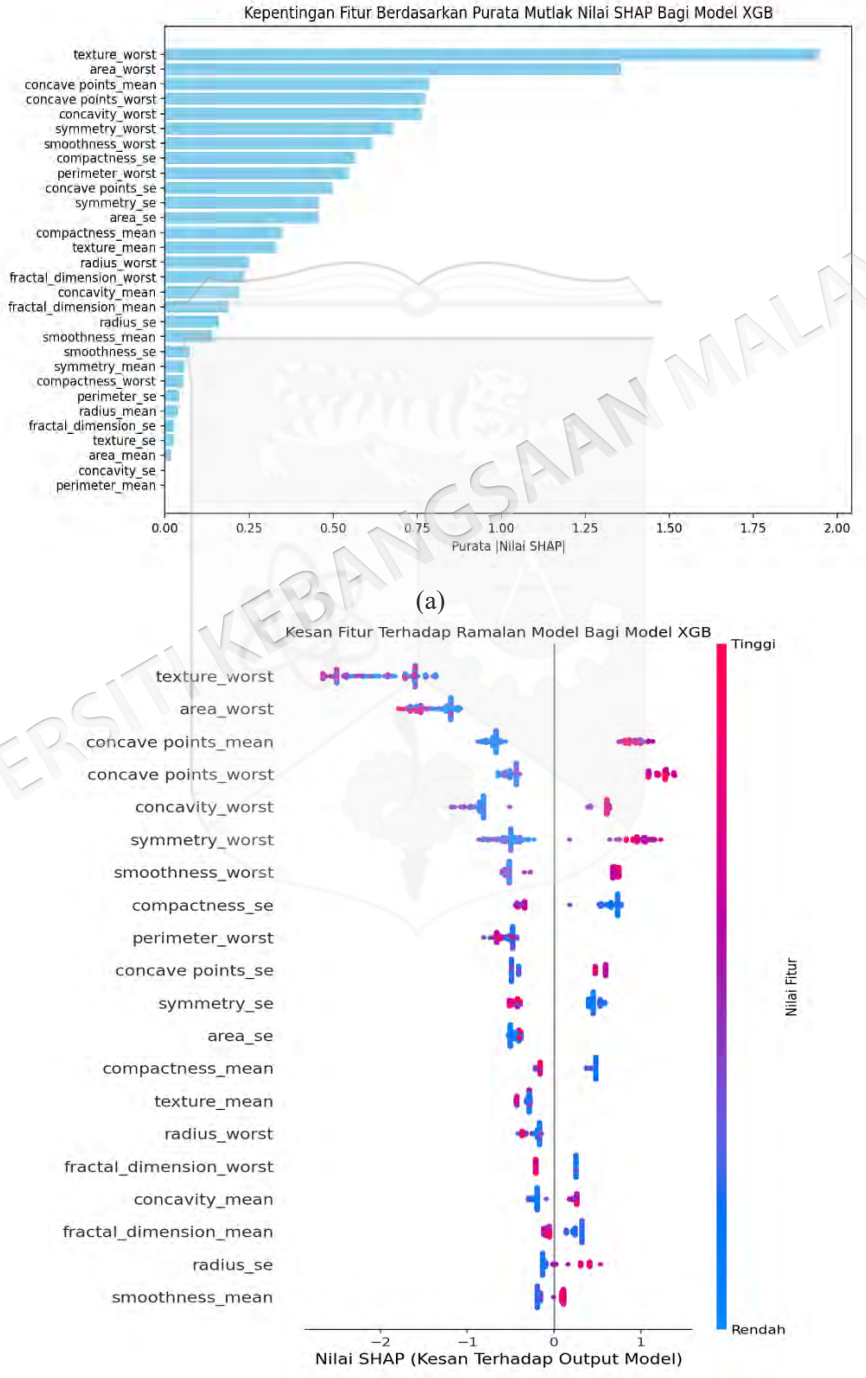
(a)



(b)

Rajah 5.5 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model GB

Rajah 5.6(a) menunjukkan fitur *texture_worst*, *area_worst* dan *concave points_mean* paling penting dalam menentukan keputusan model XGB. Rajah 5.6(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan tumor benigna. Sebaliknya, fitur *concavity_se* dan *perimeter_mean* tidak menunjukkan sumbangan tinggi pada keputusan model XGB.



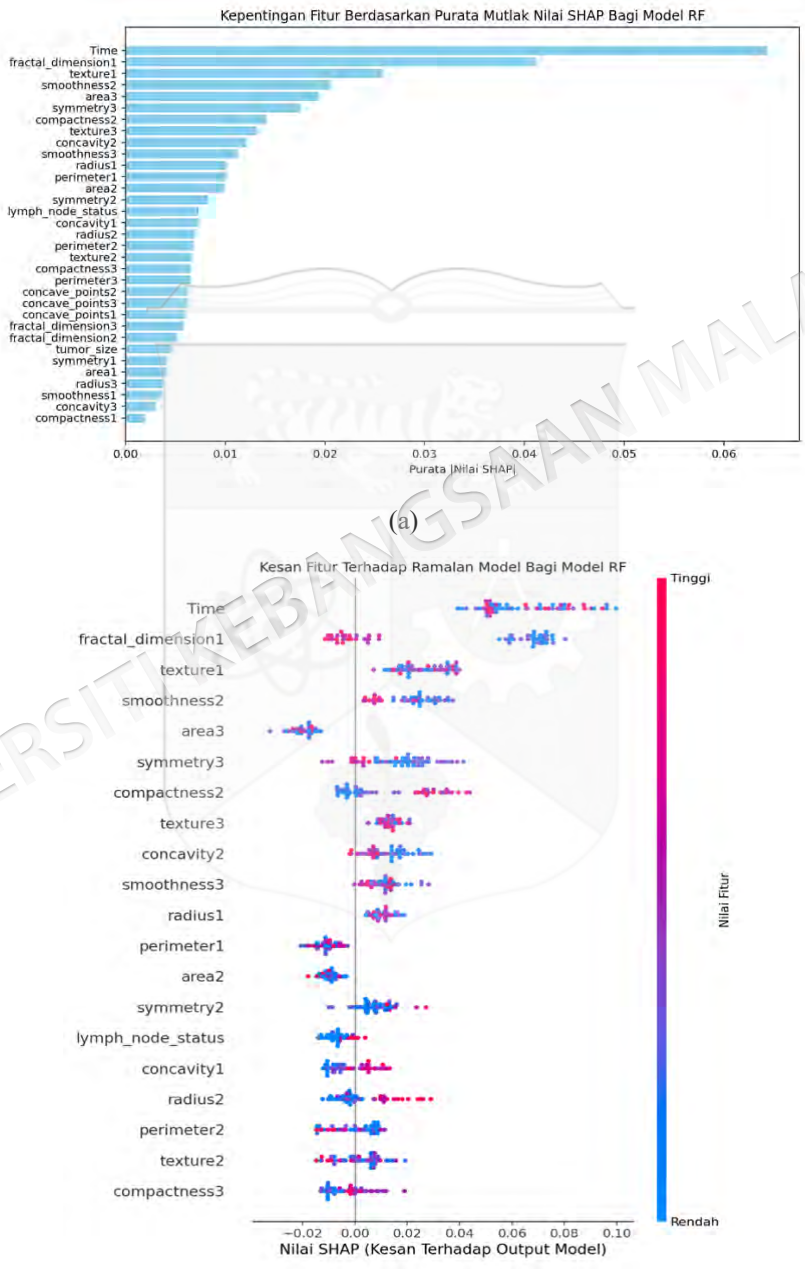
Rajah 5.6 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model XGB

Seterusnya, bagi set data WPBC, terdapat beberapa fitur menunjukkan nilai SHAP tertinggi pada ketiga-tiga model termasuk *Time*, *tumor_size* dan *lymph_node_status* berdasarkan Jadual 5.3.

Jadual 5.3 Purata nilai SHAP bagi setiap model pengelasan bagi set data WPBC

Fitur	**Model Pengelasan**		
	RF	**GB**	**XGB**
Time	**0.0644**	**3.4797**	**0.8898**
radius1	0.0102	0.3863	0.1421
texture1	0.0258	0.0813	0.2561
perimeter1	0.0101	0.0794	0.2062
area1	0.0042	0.4015	0.1597
smoothness1	0.0036	0.3560	0.3903
compactness1	0.0020	0.3978	0.0894
concavity1	0.0073	0.5839	0.0762
concave_points1	0.0060	0.3204	0.0691
symmetry1	0.0042	0.3299	0.0561
fractal_dimension1	0.0411	1.9777	0.1419
radius2	0.0070	0.5357	0.1647
texture2	0.0066	0.2740	0.0955
perimeter2	0.0068	0.3089	0.0871
area2	0.0099	0.2370	0.1637
smoothness2	0.0205	0.7708	0.5226
compactness2	0.0141	0.8667	0.2480
concavity2	0.0121	0.4344	0.1055
concave_points2	0.0063	0.5584	0.2879
symmetry2	0.0083	0.4756	0.1794
fractal_dimension2	0.0052	0.5219	0.4708
radius3	0.0038	0.6049	0.4071
texture3	0.0132	0.4606	0.2233
perimeter3	0.0065	0.3000	0.1340
area3	0.0194	0.3282	0.4327
smoothness3	0.0112	0.6021	0.4161
compactness3	0.0066	0.2888	0.1550
concavity3	0.0030	0.3435	0.4445
concave_points3	0.0063	0.3273	0.1570
symmetry3	0.0175	1.2007	0.4823
fractal_dimension3	0.0058	0.3950	0.2632
tumor_size	**0.0047**	**0.4173**	**0.2746**
lymph_node_status	**0.0073**	**0.2934**	**0.4545**

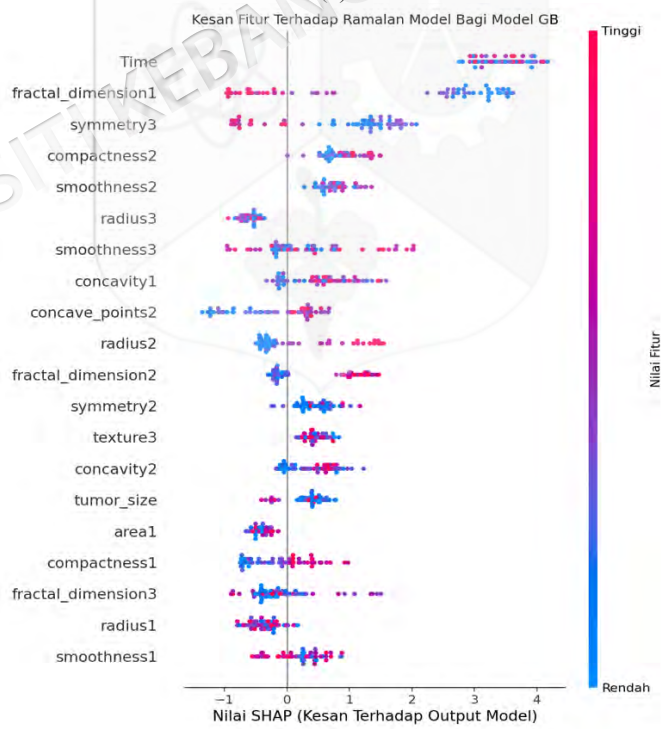
Rajah 5.7(a) menunjukkan fitur *Time*, *fractal_dimension1* dan *texture1* paling penting dalam menentukan keputusan model RF. Rajah 5.7(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan kanser berulang. Sebaliknya, fitur *concavity3* dan *compactness1* tidak menunjukkan sumbangan tinggi pada keputusan model RF.



Rajah 5.7 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model RF

Rajah 5.8(a) menunjukkan fitur *Time*, *fractal_dimension1* dan *symmetry3* paling penting dalam menentukan keputusan model GB. Rajah 5.8(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan kanser berulang Sebaliknya, fitur *texture1* dan *perimeter1* tidak menunjukkan sumbangan tinggi pada keputusan model GB.

(a)

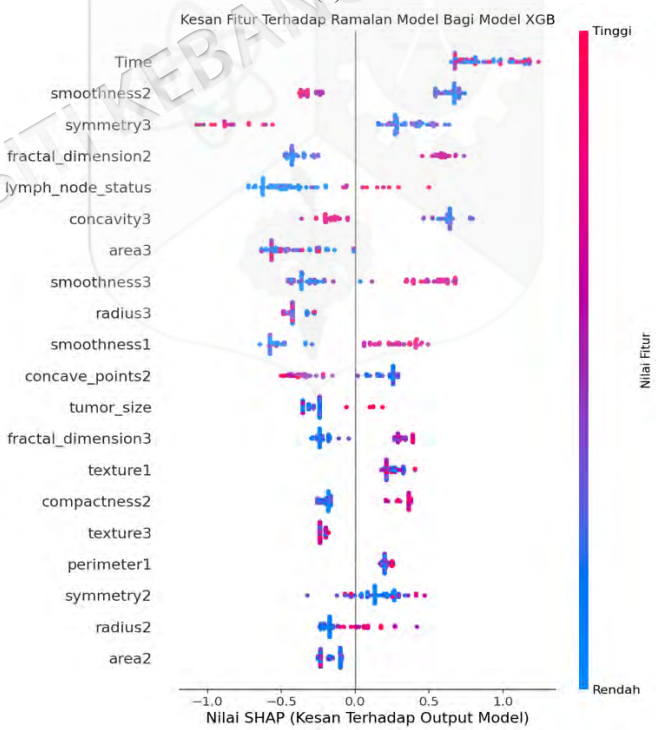


(b)

Rajah 5.8 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model GB

Rajah 5.9(a) menunjukkan fitur *Time*, *smoothness2* dan *symmetry3* paling penting dalam menentukan keputusan model XGB. Rajah 5.9(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan kanser berulang. Sebaliknya, fitur *concave_points1* dan *symmetry1* tidak menunjukkan sumbangan tinggi pada keputusan model XGB.

(a)



(b)

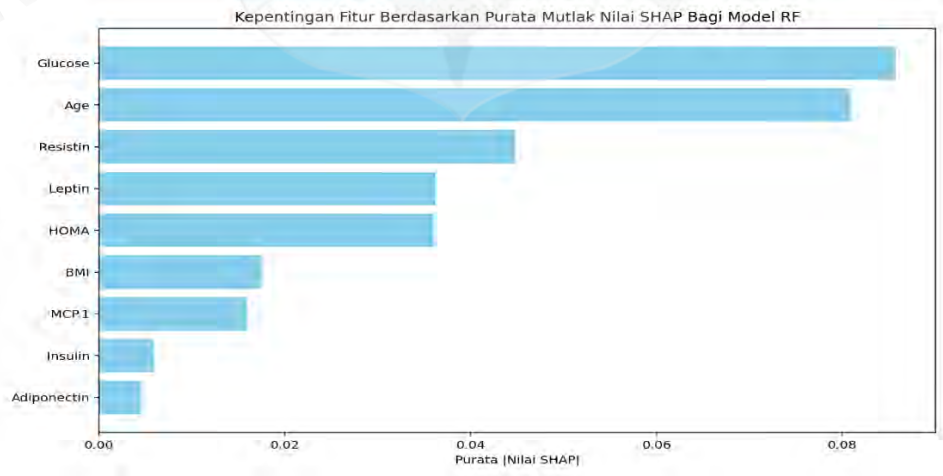
Rajah 5.9 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model XGB

Seterusnya, bagi set data BCC, purata nilai mutlak SHAP bagi setiap fitur dicatatkan dalam Jadual 5.4. Terdapat beberapa fitur menunjukkan nilai SHAP tertinggi pada ketiga-tiga model termasuk *Age*, *Glucose* dan *Resistin.*

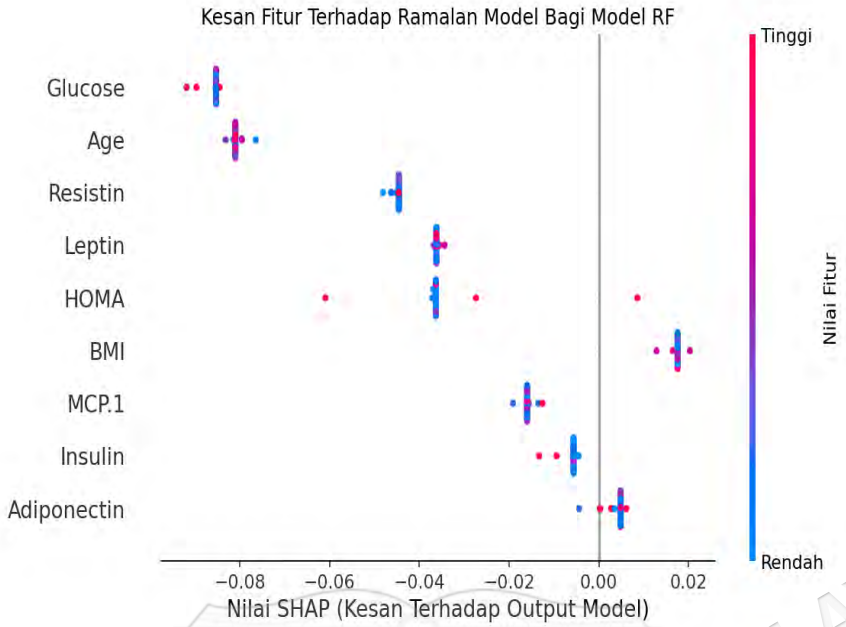
Jadual 5.4 Purata nilai mutlak SHAP bagi setiap model pengelasan bagi set data BCC

Fitur	Model Pengelasan		
	RF	**GB**	**XGB**
Age	**0.0809**	**4.0232**	**0.2191**
BMI	0.0175	0.9033	0.1212
Glucose	**0.0858**	**0.0853**	**1.3912**
Insulin	0.0059	1.8783	0.0277
HOMA	0.0361	2.3223	0.5739
Leptin	0.0362	5.1924	0.0327
Adiponectin	0.0046	1.7123	0.1618
Resistin	**0.0448**	**2.7901**	**0.3452**
MCP.1	0.0159	2.8007	0.0136

Rajah 5.10(a) menunjukkan fitur *Glucose*, *Age* dan *Resistin* paling penting dalam menentukan keputusan model RF. Rajah 5.11(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan individu sihat. Namun terdapat juga kes di mana nilai tinggi bagi fitur *Glucose* dan *Age* turut memberi kesan kepada pengelasan yang sama. Sebaliknya, fitur *Insulin* dan *Adiponectin* tidak menunjukkan sumbangan tinggi pada keputusan model RF.



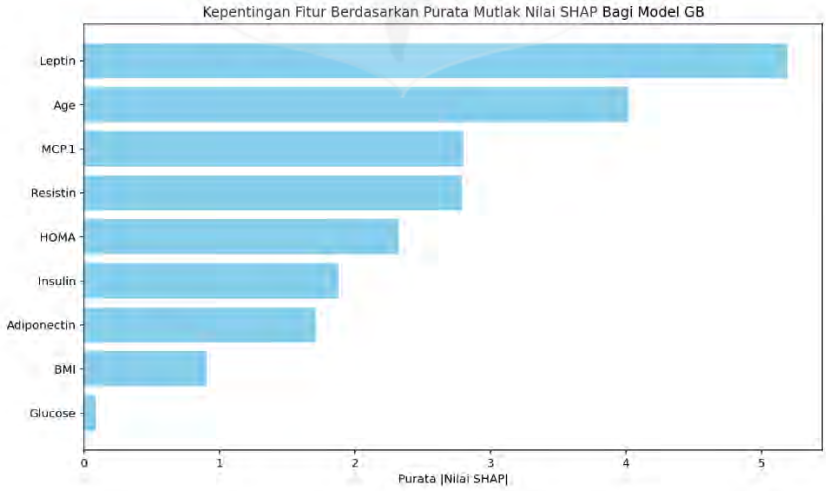
(a)



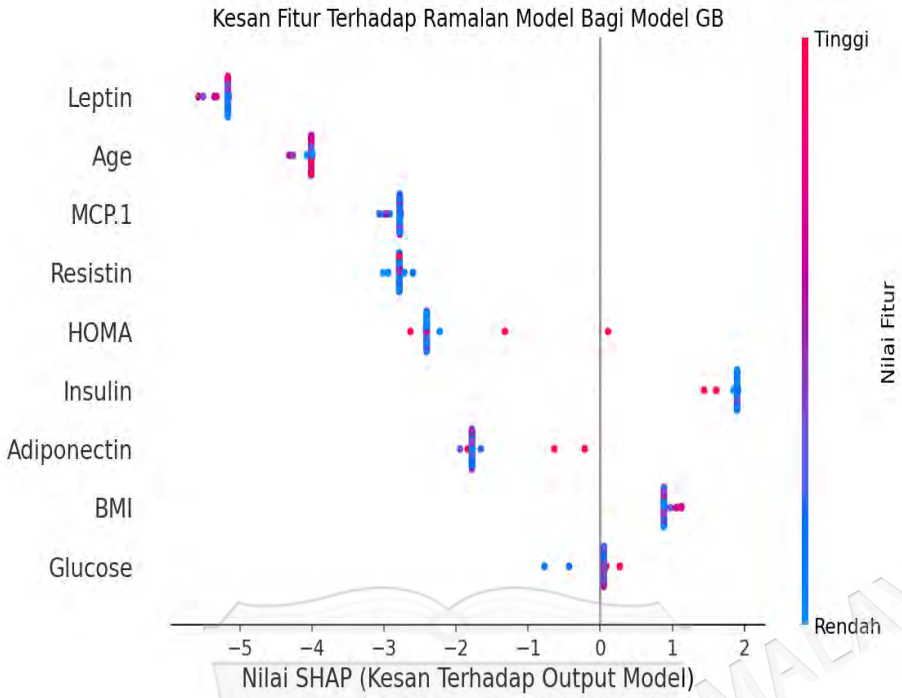
(b)

Rajah 5.10 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model RF

Rajah 5.11(a) menunjukkan fitur *Leptin*, *Age* dan *MCP.1* paling penting dalam menentukan keputusan model GB. Rajah 5.12(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan individu sihat. Namun terdapat juga kes di mana nilai tinggi bagi fitur *Leptin* dan *Age* turut memberi kesan kepada pengelasan yang sama. Sebaliknya, fitur *BMI* dan *Glucose* tidak menunjukkan sumbangan tinggi pada keputusan model GB.



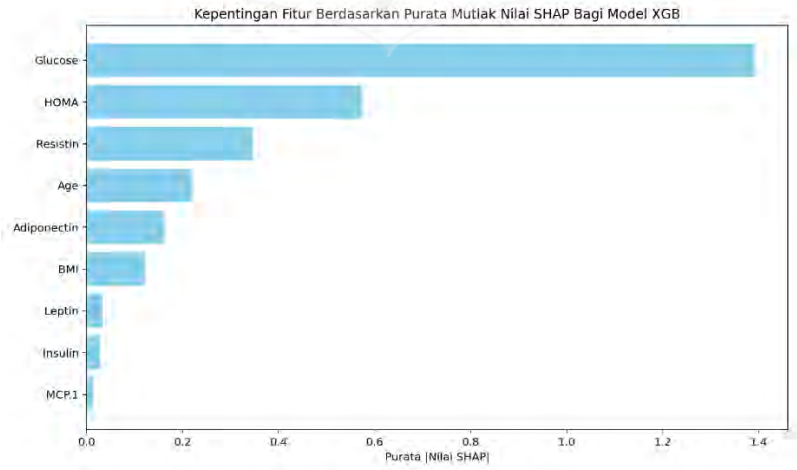
(a)



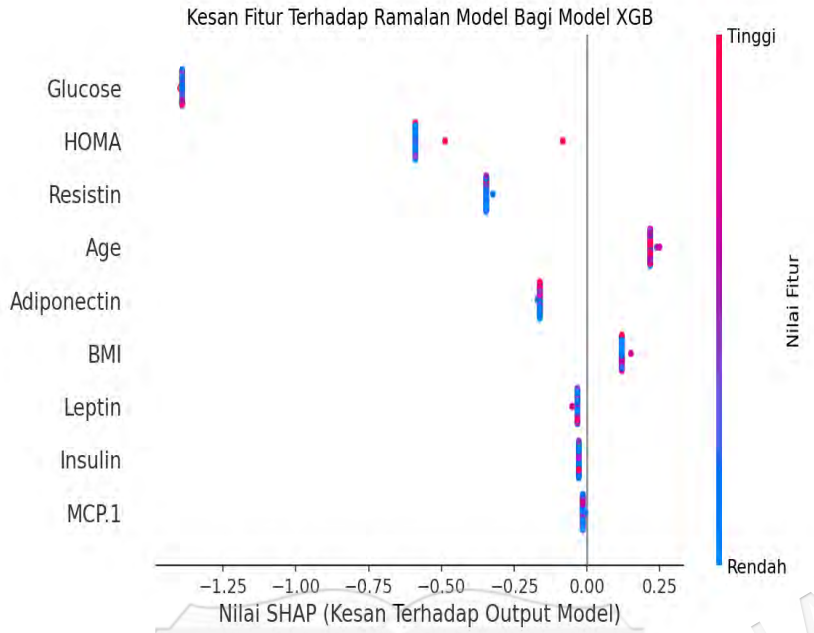
(b)

Rajah 5.11 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model GB

Rajah 5.12(a) menunjukkan fitur *Glucose, HOMA* dan *Resistin* paling penting dalam menentukan keputusan model XGB. Rajah 5.13(b) juga menunjukkan bahawa bagi nilai rendah bagi fitur-fitur ini cenderung kepada pengelasan individu sihat. Sebaliknya, fitur *Insulin* dan *MCP.1* tidak menunjukkan sumbangan tinggi pada keputusan model GB.



(a)



(b)

Rajah 5.12 Plot (a) kepentingan nilai SHAP (b) ringkasan SHAP bagi model XGB

5.2.3 Penjanaan Nilai Interaksi Fitur Menggunakan SHAP

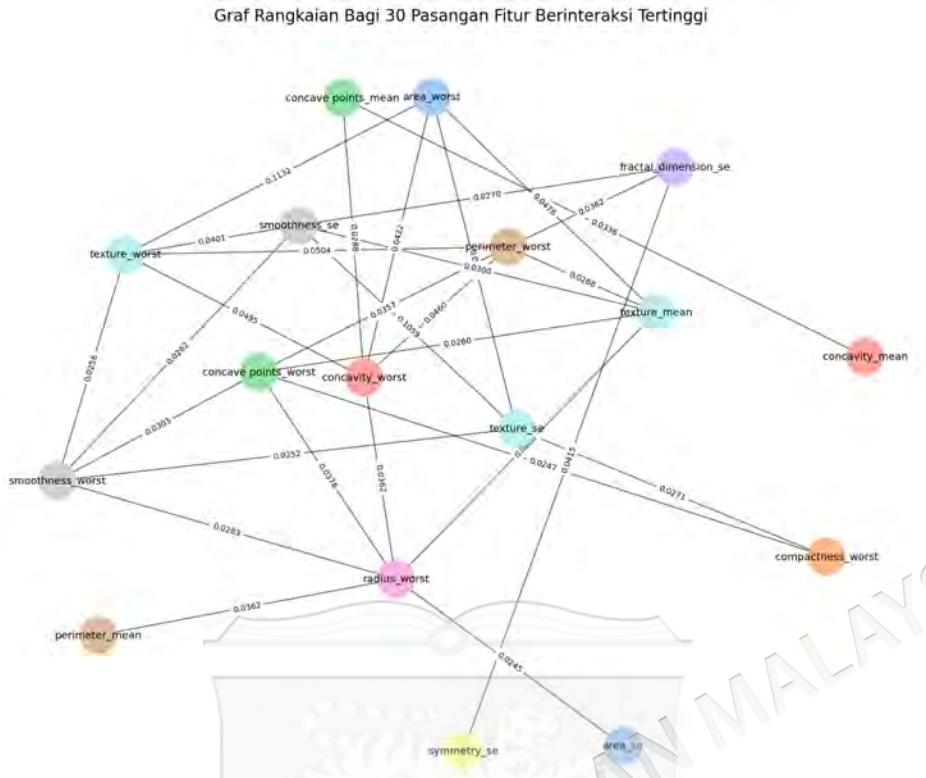
Nilai interaksi SHAP membolehkan pemahaman bagaimana dua fitur bersama-sama mempengaruhi ramalan model berbanding sumbangan setiap fitur secara berasingan. Dengan menilai interaksi ini, fitur yang kurang relevan secara individu tetapi memberikan impak yang besar apabila digabungkan dengan fitur lain dapat dikenal pasti yang selaras dengan objektif kedua kajian ini. Setiap model digunakan untuk menjana nilai interaksi SHAP secara berasingan bagi memastikan ketepatan dan kebolehpercayaan penilaian. Kemudian, purata interaksi SHAP daripada ketiga-tiga model dikira bagi setiap pasangan fitur untuk mendapatkan ukuran yang lebih stabil dan tidak berat sebelah kepada mana-mana model tertentu.

Jadual 5.5 menyenaraikan sepuluh pasangan fitur yang mencatatkan nilai purata interaksi tertinggi dalam set data WDBC. Pasangan fitur berinteraksi yang menonjol termasuk *texture_worst–area_worst*, *texture_se–smoothness_se*, *texture_mean–radius_worst* dan *texture_worst–perimeter_worst.*

Jadual 5.5 Pasangan fitur dengan purata nilai interaksi SHAP bagi set data WDBC

Pasangan Fitur		**Nilai Interaksi SHAP**			**Nilai Purata Mutlak Interaksi**
Fitur Pertama	**Fitur Kedua**	**RF**	**GB**	**XGB**	
texture_worst	***area_worst***	**0.0009**	**0.1547**	**0.1839**	**0.1132**
texture_se	***smoothness_se***	**-0.0015**	**-0.3163**	**0.0000**	**0.1059**
texture_mean	***radius_worst***	**0.0073**	**0.2528**	**0.0065**	**0.0888**
texture_worst	***perimeter_worst***	**0.0004**	**0.1294**	**-0.0214**	**0.0504**
texture_worst	*concavity_worst*	0.0013	0.0414	0.1059	0.0495
texture_mean	*area_worst*	0.0046	0.1136	0.0244	0.0476
perimeter_worst	*concavity_worst*	0.0011	0.0286	0.1082	0.0460
area_worst	*concavity_worst*	0.0057	0.0760	0.0449	0.0422
symmetry_se	*fractal_dimension_se*	-0.0002	0.1244	0.0000	0.0415
smoothness_se	*texture_worst*	0.0000	-0.1201	0.0000	0.0401

Rajah 5.13 memaparkan graf rangkaian interaksi antara fitur berdasarkan purata nilai interaksi SHAP daripada tiga model. Setiap nod mewakili fitur dan garisan penghubung menunjukkan kekuatan interaksi antara fitur. Berdasarkan Rajah 5.13, fitur *texture_worst* berinteraksi dengan beberapa fitur lain termasuk *area_worst*, *smoothness_worst*, *concavity_worst* dan *perimeter_worst* manakala fitur radius_worst berinteraksi dengan fitur *smoothness_worst, perimeter_mean*, *texture_mean* dan *area_se*.



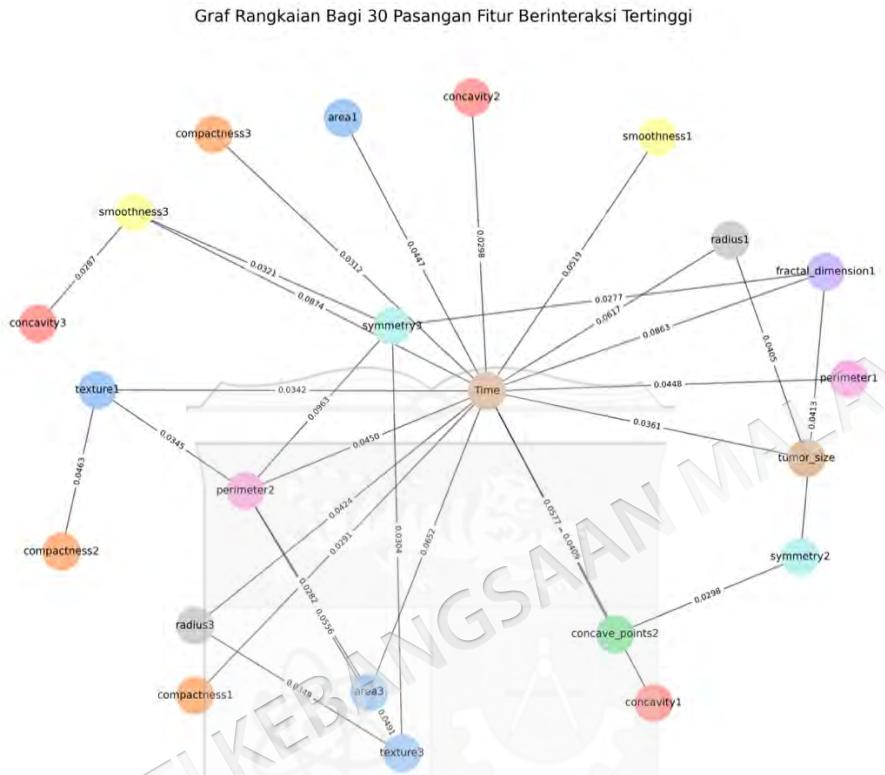
Rajah 5.13 Graf rangkaian bagi 30 pasangan fitur berinteraksi tertinggi dalam set data WDBC

Seterusnya, bagi set data WPBC, Jadual 5.6 menyenaraikan sepuluh pasangan fitur yang mencatatkan nilai purata interaksi tertinggi. Pasangan fitur berinteraksi yang menonjol termasuk *perimeter2–symmetry3*, *Time–smoothness3*, *Time–fractal_dimension1* dan *Time–area3.*

Jadual 5.6 Pasangan fitur dengan purata nilai interaksi SHAP bagi set data WPBC

Pasangan Fitur		**Nilai Interaksi SHAP**			**Nilai Purata Mutlak Interaksi**
Fitur Pertama	**Fitur Kedua**	**RF**	**GB**	**XGB**	
perimeter2	***symmetry3***	**0.0027**	**0.2862**	**0.0000**	**0.0963**
Time	***smoothness3***	**-0.0009**	**-0.2577**	**0.0035**	**0.0874**
Time	***fractal_dimension1***	**-0.0090**	**-0.2376**	**-0.0123**	**0.0863**
Time	***area3***	**-0.0033**	**-0.0878**	**-0.1045**	**0.0652**
Time	*radius1*	-0.0065	-0.1453	-0.0333	0.0617
Time	*concave_points2*	-0.0011	-0.1720	0.0000	0.0577
perimeter2	*texture3*	0.0001	-0.1668	0.0000	0.0556
Time	*smoothness1*	-0.0014	-0.1492	0.0052	0.0519
texture3	*area3*	0.0017	0.1454	0.0000	0.0491
texture1	*compactness2*	-0.0008	-0.1267	-0.0113	0.0463

Rajah 5.14 menunjukkan fitur *Time* berinteraksi dengan beberapa fitur lain termasuk *area1, concavity2*, *smoothness1* dan *radius1* manakala fitur *tumor_size* berinteraksi dengan fitur *symmetry2, fractal_dimension1, radius1* dan *Time*.



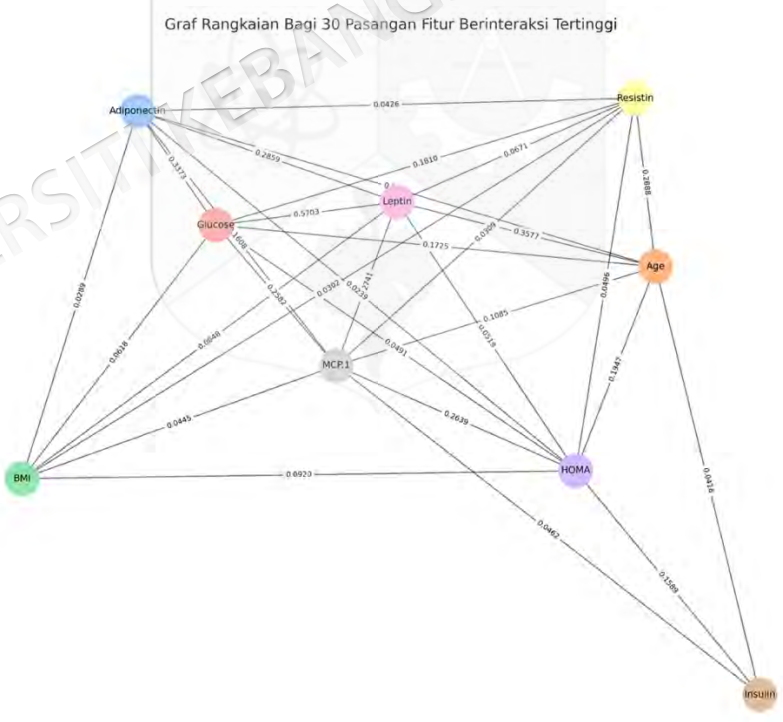
Rajah 5.14 Graf rangkaian bagi 30 pasangan fitur berinteraksi tertinggi bagi set data WPBC

Seterusnya, bagi set data BCC pula, Jadual 5.7 menyenaraikan sepuluh pasangan fitur yang mencatatkan nilai purata interaksi tertinggi. Pasangan fitur berinteraksi yang menonjol termasuk *Glucose–Leptin*, *Age–Leptin*, *Glucose–Adiponectin* dan *Age–Resistin.*

Jadual 5.7 Pasangan fitur dan purata nilai interaksi SHAP bagi set data BCC

Pasangan Fitur		Nilai Interaksi SHAP			Nilai Purata Mutlak Interaksi
Fitur Pertama	**Fitur Kedua**	**RF**	**GB**	**XGB**	
Glucose	***Leptin***	**0.0224**	**1.6679**	**0.0204**	**0.5703**
Age	***Leptin***	**0.0127**	**1.0573**	**0.0031**	**0.3577**
Glucose	***Adiponectin***	**0.0125**	**0.9951**	**-0.0042**	**0.3373**
Age	***Resistin***	**0.0108**	**0.7774**	**0.0781**	**0.2888**
Leptin	*Adiponectin*	0.0062	0.8515	0.0000	0.2859
Leptin	*MCP.1*	0.0035	0.8144	0.0044	0.2741
HOMA	*MCP.1*	0.0073	0.7827	-0.0016	0.2639
Glucose	*MCP.1*	0.0202	0.7270	0.0274	0.2582
Age	*HOMA*	0.0045	0.5710	-0.0087	0.1947
Glucose	*Resistin*	0.0017	-0.5340	0.0072	0.1810

Rajah 5.15 menunjukkan fitur *Glucose* berinteraksi dengan beberapa fitur lain termasuk *BMI, Adiponectin*, *Leptin* dan *MCP.1* manakala fitur *Age* berinteraksi dengan fitur *HOMA, Resistin, MCP.1* dan *Insulin*.



Rajah 5.15 Graf rangkaian bagi 30 pasangan fitur berinteraksi tertinggi bagi set data BCC

5.2.4 Penarafan Fitur Menggunakan Pengundian Majoriti Berwajaran

Proses penarafan fitur menggunakan kaedah pengundian majoriti berwajaran yang berdasarkan nilai kepentingan SHAP adalah bertujuan untuk menggabungkan kekuatan ketiga-tiga model bagi menilai kestabilan dan kepentingan fitur yang telah dikenal pasti sebelum ini. Jadual 5.8 memaparkan kedudukan awal setiap fitur berdasarkan nilai purata mutlak SHAP sebelum proses pengundian majoriti berwajaran dilakukan bagi set data WDBC.

Jadual 5.8 Kedudukan fitur berdasarkan purata nilai SHAP bagi set data WDBC

Fitur	**Model Pengelasan**		
	RF	**GB**	**XGB**
radius_mean	20	20	25
texture_mean	10	4	14
perimeter_mean	**16**	**24**	**29**
area_mean	**17**	**30**	**28**
smoothness_mean	21	25	20
compactness_mean	**29**	**28**	**13**
concavity_mean	11	13	17
concave points_mean	2	9	3
symmetry_mean	28	23	22
fractal_dimension_mean	25	22	18
radius_se	15	18	19
texture_se	26	2	27
perimeter_se	24	21	24
area_se	13	10	12
smoothness_se	12	8	21
compactness_se	27	27	8
concavity_se	23	19	29
concave points_se	30	26	10
symmetry_se	18	14	11
fractal_dimension_se	22	12	26
radius_worst	4	7	15
texture_worst	**8**	**5**	**1**
perimeter_worst	6	15	9
area_worst	5	17	2
smoothness_worst	**7**	**3**	**7**
compactness_worst	19	16	23
concavity_worst	**3**	**6**	**5**
concave points_worst	**1**	**1**	**4**
symmetry_worst	9	11	6
fractal_dimension_worst	14	29	16

Fitur yang paling konsisten berada dalam kedudukan tinggi ialah *concave_points_worst*, *concavity_worst*, *texture_worst* dan *smoothness_worst*. Sebaliknya, fitur seperti *area_mean*, *compactness_mean* dan *perimeter_mean* mempunyai kedudukan yang rendah dalam semua model menunjukkan sumbangannya terhadap model kurang signifikan.

Sebelum pengiraan dilakukan, skor AUC setiap model dinormalkan terlebih dahulu bagi memastikan keseragaman skala. Setelah itu, kedudukan akhir fitur berdasarkan nilai SHAP berwajaran diperoleh selepas proses pengundian majoriti berwajaran seperti pada Jadual 5.9. Fitur *texture_worst* berada di kedudukan pertama dengan nilai SHAP berwajaran tertinggi (1.2092) diikuti oleh *concave_points_worst* (1.1797), *smoothness_worst* (0.7767) dan *concavity_worst* (0.7725). Sebaliknya, fitur *fractal_dimension_worst*, *perimeter_mean* dan *area_mean* berada dalam kelompok terendah.

Jadual 5.9 Kedudukan fitur berdasarkan nilai SHAP berwajaran bagi set data WDBC

Fitur	Nilai purata mutlak SHAP			Nilai SHAP berwajaran	Kedudukan Fitur
	RF	GB	XGB		
texture_worst	**0.0252**	**1.6548**	**1.9461**	**1.2092**	**1**
concave points_worst	**0.0816**	**2.6828**	**0.7727**	**1.1797**	**2**
smoothness_worst	**0.0258**	**1.6848**	**0.6183**	**0.7767**	**3**
concavity_worst	**0.0425**	**1.5117**	**0.7622**	**0.7725**	**4**
texture_se	0.0044	2.0803	0.0286	0.7049	5
area_worst	0.0403	0.6501	1.3568	0.6826	6
texture_mean	0.0222	1.6795	0.3303	0.6777	7
concave points_mean	0.0544	0.9032	0.7857	0.5813	8
radius_worst	0.0404	1.4391	0.2481	0.5762	9
symmetry_worst	0.0252	0.8491	0.6780	0.5177	10
area_se	0.0192	0.8515	0.4564	0.4426	11
smoothness_se	0.0197	1.2024	0.0728	0.4319	12
perimeter_worst	0.0312	0.6925	0.5453	0.4232	13
symmetry_se	0.0133	0.6959	0.4598	0.3898	14
concavity_mean	0.0221	0.7364	0.2220	0.3270	15
fractal_dimension_se	0.0072	0.8430	0.0288	0.2932	16
compactness_se	0.0034	0.2070	0.5652	0.2586	17
concave points_se	0.0024	0.2610	0.4975	0.2537	18
radius_se	0.0160	0.5615	0.1610	0.2463	19

bersambung...

...sambungan

compactness_worst	0.0118	0.6519	0.0546	0.2396	20
fractal_dimension_mean	0.0050	0.3384	0.1880	0.1772	21
compactness_mean	0.0032	0.1658	0.3497	0.1730	22
concavity_se	0.0063	0.5009	0.0000	0.1692	23
smoothness_mean	0.0079	0.2886	0.1404	0.1457	24
radius_mean	0.0103	0.3605	0.0392	0.1368	25
perimeter_se	0.0057	0.3486	0.0424	0.1323	26
symmetry_mean	0.0032	0.3281	0.0587	0.1301	27
fractal_dimension_worst	**0.0176**	**0.1313**	**0.2363**	**0.1285**	**28**
perimeter_mean	**0.0154**	**0.3047**	**0.0000**	**0.1068**	**29**
area_mean	**0.0148**	**0.1201**	**0.0163**	**0.0504**	**30**

Bagi set data WPBC pula, fitur yang paling konsisten berada dalam kedudukan tinggi berdasarkan Jadual 5.10 ialah *Time*, *fractal_dimension1*, *symmetry3* dan *smoothness2.* Sebaliknya, fitur seperti *area1*, *concavity1* dan *area3* mempunyai kedudukan yang rendah dalam semua model menunjukkan sumbangannya terhadap model kurang signifikan.

Jadual 5.10 Kedudukan fitur berdasarkan purata nilai mutlak SHAP bagi set data WPBC

Fitur	**Model Pengelasan**		
	RF	**GB**	**XGB**
Time	**1**	**1**	**1**
radius1	11	19	24
texture1	3	32	14
perimeter1	12	33	17
area1	**29**	**16**	**21**
smoothness1	31	20	10
compactness1	33	17	29
concavity1	**16**	**8**	**31**
concave_points1	24	25	32
symmetry1	28	22	33
fractal_dimension1	**2**	**2**	**25**
radius2	17	10	19
texture2	19	30	28
perimeter2	18	26	30
area2	13	31	20
smoothness2	**4**	**5**	**2**
compactness2	7	4	15
concavity2	9	14	27

bersambung...

...sambungan

concave_points2	22	9	11
symmetry2	14	12	18
fractal_dimension2	26	11	4
radius3	30	6	9
texture3	8	13	16
perimeter3	21	27	26
area3	**5**	**23**	**7**
smoothness3	10	7	8
compactness3	20	29	23
concavity3	32	21	6
concave_points3	23	24	22
symmetry3	**6**	**3**	**3**
fractal_dimension3	25	18	13
tumor_size	27	15	12
lymph_node_status	15	28	5

Berdasarkan Jadual 5.11, selepas proses pengundian majoriti berwajaran dilakukan, fitur *Time* berada di kedudukan pertama dengan nilai SHAP berwajaran tertinggi (1.4861) diikuti oleh *fractal_dimension1* (0.7232), *symmetry3* (0.5704) dan *smoothness2* (0.4412). Sebaliknya, fitur *texture2, texture1* dan *perimeter1* berada dalam kelompok terendah.

Jadual 5.11 Kedudukan fitur berdasarkan nilai SHAP berwajaran bagi set data WPBC

Fitur	Nilai purata mutlak SHAP			Nilai SHAP berwajaran	Kedudukan Fitur
	RF	GB	XGB		
Time	**0.0644**	**3.4797**	**0.8898**	**1.4861**	**1**
fractal_dimension1	**0.0411**	**1.9777**	**0.1419**	**0.7232**	**2**
symmetry3	**0.0175**	**1.2007**	**0.4823**	**0.5704**	**3**
smoothness2	0.0205	0.7708	0.5226	0.4412	4
compactness2	0.0141	0.8667	0.2480	0.3784	5
smoothness3	0.0112	0.6021	0.4161	0.3457	6
radius3	0.0038	0.6049	0.4071	0.3412	7
fractal_dimension2	0.0052	0.5219	0.4708	0.3354	8
concave_points2	0.0063	0.5584	0.2879	0.2862	9
concavity3	0.0030	0.3435	0.4445	0.2661	10
area3	0.0194	0.3282	0.4327	0.2623	11
lymph_node_status	0.0073	0.2934	0.4545	0.2541	12
smoothness1	0.0036	0.3560	0.3903	0.2522	13
radius2	0.0070	0.5357	0.1647	0.2372	14
tumor_size	0.0047	0.4173	0.2746	0.2339	15

bersambung...

...sambungan

texture3	0.0132	0.4606	0.2233	0.2339	16
concavity1	0.0073	0.5839	0.0762	0.2235	17
fractal_dimension3	0.0058	0.3950	0.2632	0.2230	18
symmetry2	0.0083	0.4756	0.1794	0.2225	19
area1	0.0042	0.4015	0.1597	0.1897	20
concavity2	0.0121	0.4344	0.1055	0.1850	21
radius1	0.0102	0.3863	0.1421	0.1806	22
concave_points3	0.0063	0.3273	0.1570	0.1646	23
compactness1	0.0020	0.3978	0.0894	0.1640	24
compactness3	0.0066	0.2888	0.1550	0.1511	25
perimeter3	0.0065	0.3000	0.1340	0.1478	26
area2	0.0099	0.2370	0.1637	0.1378	27
perimeter2	0.0068	0.3089	0.0871	0.1350	28
concave_points1	0.0060	0.3204	0.0691	0.1326	29
symmetry1	0.0042	0.3299	0.0561	0.1307	30
texture2	**0.0066**	**0.2740**	**0.0955**	**0.1262**	**31**
texture1	**0.0258**	**0.0813**	**0.2561**	**0.1222**	**32**
perimeter1	**0.0101**	**0.0794**	**0.2062**	**0.0995**	**33**

Seterusnya, bagi set data BCC, berdasarkan Jadual 5.12, fitur yang paling konsisten berada dalam kedudukan tinggi ialah *Resistin*, *Age* dan *HOMA*. Sebaliknya, fitur seperti *MCP.1* dan *Adiponectin* mempunyai kedudukan yang rendah dalam semua model menunjukkan sumbangannya terhadap model kurang signifikan.

Jadual 5.12 Kedudukan fitur berdasarkan purata nilai mutlak SHAP bagi set data BCC

Fitur	**Model Pengelasan**		
	RF	**GB**	**XGB**
Age	**2**	**2**	**4**
BMI	6	8	6
Glucose	1	9	1
Insulin	8	6	8
HOMA	**5**	**5**	**2**
Leptin	4	1	7
Adiponectin	**9**	**7**	**5**
Resistin	**3**	**4**	**3**
MCP.1	**7**	**3**	**9**

Namun, selepas proses pengundian majoriti berwajaran dilakukan, Jadual 5.13 memperlihatkan fitur *Leptin* berada di kedudukan pertama dengan nilai SHAP berwajaran tertinggi (1.7512) diikuti oleh *Age* (1.4393), *Resistin* (1.0592) dan *HOMA*

(0.9772). Sebaliknya, fitur *Adiponectin, Glucose* dan *BMI* berada dalam kelompok terendah.

Jadual 5.13 Kedudukan fitur berdasarkan nilai SHAP berwajaran bagi set data BCC

Fitur	Nilai purata mutlak SHAP			Nilai SHAP berwajaran	Kedudukan Fitur
	RF	GB	XGB		
Leptin	**0.0362**	**5.1924**	**0.0327**	**1.7512**	**1**
Age	**0.0809**	**4.0232**	**0.2191**	**1.4393**	**2**
Resistin	**0.0448**	**2.7901**	**0.3452**	**1.0592**	**3**
HOMA	**0.0361**	**2.3223**	**0.5739**	**0.9772**	**4**
MCP.1	0.0159	2.8007	0.0136	0.9420	5
Insulin	0.0059	1.8783	0.0277	0.6364	6
Adiponectin	**0.0046**	**1.7123**	**0.1618**	**0.6256**	**7**
Glucose	**0.0858**	**0.0853**	**1.3912**	**0.5228**	**8**
BMI	**0.0175**	**0.9033**	**0.1212**	**0.3471**	**9**

5.2.5 Pengenalpastian Fitur Kurang Relevan Berdasarkan Nilai SHAP Berwajaran

Dalam seksyen ini, fitur-fitur yang mempunyai nilai SHAP berwajaran di bawah kuantil ke-75 dikenal pasti sebagai kurang relevan secara individu. Walau bagaimanapun, fitur ini tidak disingkirkan serta-merta kerana berpotensi memberikan kesan melalui interaksi dengan fitur lain. Jadual 5.14 menyenaraikan fitur-fitur yang kurang relevan dalam set data WDBC berdasarkan kuantil ke-75 nilai SHAP berwajaran (0.5800). Terdapat 22 fitur yang dikenal pasti di bawah ambang kuantil ke-75. Fitur-fitur ini termasuk *radius_worst, symmetry_worst*, *area_se fractal_dimension_worst, perimeter_mean* dan *area_mean.*

Jadual 5.14 Senarai fitur kurang relevan pada set data WDBC

Fitur	Nilai SHAP berwajaran	Kedudukan Fitur
radius_worst	**0.5762**	**9**
symmetry_worst	**0.5177**	**10**
area_se	**0.4426**	**11**
smoothness_se	0.4319	12
perimeter_worst	0.4232	13
symmetry_se	0.3898	14
concavity_mean	0.3270	15
fractal_dimension_se	0.2932	16

bersambung...

...sambungan

compactness_se	0.2586	17
concave points_se	0.2537	18
radius_se	0.2463	19
compactness_worst	0.2396	20
fractal_dimension_mean	0.1772	21
compactness_mean	0.1730	22
concavity_se	0.1692	23
smoothness_mean	0.1457	24
radius_mean	0.1368	25
perimeter_se	0.1323	26
symmetry_mean	0.1301	27
fractal_dimension_worst	**0.1285**	**28**
perimeter_mean	**0.1068**	**29**
area_mean	**0.0504**	**30**

Seterusnya, Jadual 5.15 menyenaraikan fitur-fitur yang kurang relevan dalam set data WPBC. Sebanyak 24 fitur dikenal pasti di bawah ambang kuantil ke-75 nilai SHAP berwajaran (0.2862). Fitur-fitur ini termasuk *concavity3, area3, lymph_node_status*, *texture2*, *texture1* dan *perimeter1*.

Jadual 5.15 Senarai fitur kurang relevan pada set data WPBC

Fitur	Nilai SHAP berwajaran	Kedudukan Fitur
concavity3	**0.2656**	**10**
area3	**0.2619**	**11**
lymph_node_status	**0.2537**	**12**
smoothness1	0.2517	13
radius2	0.2366	14
tumor_size	0.2334	15
texture3	0.2334	16
concavity1	0.2229	17
fractal_dimension3	0.2225	18
symmetry2	0.2220	19
area1	0.1893	20
concavity2	0.1846	21
radius1	0.1802	22
concave_points3	0.1643	23
compactness1	0.1636	24
compactness3	0.1508	25
perimeter3	0.1475	26
area2	0.1376	27
perimeter2	0.1347	28

bersambung...

...sambungan

concave_points1	0.1322	29
symmetry1	0.1304	30
texture2	**0.1259**	**31**
texture1	**0.1221**	**32**
perimeter1	**0.0994**	**33**

Jadual 5.16 menyenaraikan fitur-fitur yang kurang relevan dalam set data BCC berdasarkan kuantil ke-75 nilai SHAP berwajaran (1.0592). Terdapat 6 fitur yang dikenal pasti di bawah ambang kuantil ke-75 termasuk *HOMA, MCP.1*, *Insulin*, *Adiponectin*, *Glucose* dan *BMI.*

Jadual 5.16 Senarai fitur kurang relevan pada set data BCC

Fitur	Nilai SHAP berwajaran	Kedudukan Fitur
HOMA	0.9772	4
MCP.1	0.9420	5
Insulin	0.6364	6
Adiponectin	0.6256	7
Glucose	0.5228	8
BMI	0.3471	9

5.2.6 Penjanaan Nilai Interaksi SHAP bagi Fitur Kurang Relevan

Jadual 5.17 menyenaraikan sepuluh pasangan fitur yang mengandungi sekurang-kurangnya satu fitur kurang relevan yang mencatatkan nilai purata interaksi tertinggi dalam set data WDBC.

Jadual 5.17 Pasangan fitur kurang relevan dan purata nilai interaksi SHAP bagi set data WDBC

Pasangan Fitur		Nilai Interaksi SHAP			Nilai Purata Mutlak Interaksi	Jenis Interaksi
Fitur Pertama	Fitur Kedua	RF	GB	XGB		
texture_se	***smoothness_se***	**-0.0015**	**-0.3163**	**0.0000**	**0.1059**	**Kuat-Lemah**
texture_mean	***radius_worst***	**0.0073**	**0.2528**	**0.0065**	**0.0888**	**Kuat-Lemah**
texture_worst	*perimeter_worst*	0.0004	0.1294	-0.0214	0.0504	Kuat-Lemah
perimeter_worst	*concavity_worst*	0.0011	0.0286	0.1082	0.0460	Lemah-Kuat
symmetry_se	***fractal_dimension_se***	**-0.0002**	**0.1244**	**0.0000**	**0.0415**	**Lemah-Lemah**
smoothness_se	*texture_worst*	0.0000	-0.1201	0.0000	0.0401	Lemah-Kuat

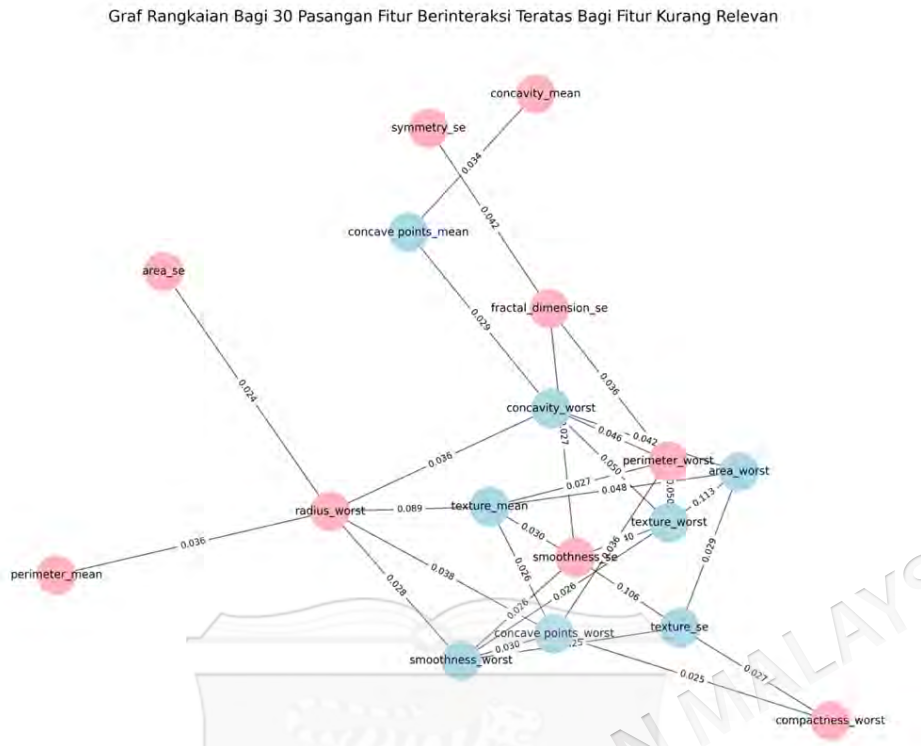
bersambung...

...sambungan

radius_worst	*concave points_worst*	0.0009	0.1078	0.0041	0.0376	Lemah-Kuat
fractal_dimension_se	***perimeter_worst***	**0.0004**	**0.0954**	**0.0129**	**0.0362**	**Lemah-Lemah**
perimeter_mean	*radius_worst*	0.0009	-0.1076	0.0000	0.0362	Lemah-Lemah
radius_worst	*concavity_worst*	0.0057	0.0742	0.0285	0.0362	Lemah-Kuat

Pasangan fitur berinteraksi yang mengandungi fitur kurang relevan termasuk *texture_se–smoothness_se* dan *texture_mean–radius_worst* menunjukkan bahawa terdapat fitur lemah secara individu namun masih mempunyai nilai apabila berinteraksi dengan fitur lain yang lebih dominan. Selain itu, pasangan seperti *symmetry_se–fractal_dimension_se* dan *fractal_dimension_se–perimeter_worst* melibatkan dua fitur kurang relevan tetapi mempunyai nilai interaksi yang tinggi.

Rajah 5.16 menunjukkan graf rangkaian bagi 30 pasangan fitur yang melibatkan fitur kurang relevan dengan purata nilai interaksi tertinggi dalam set data WDBC. Fitur lemah iaitu *radius_worst* berinteraksi dengan beberapa fitur kuat yang lain termasuk *texture_mean*, *smoothness_worst* dan *concave points_worst* serta fitur lemah seperti *perimeter_mean* dan *area_se*. Selain itu, fitur lemah *fractal_dimension_se* pula berinteraksi dengan fitur kuat seperti *concavity_worst* dan beberapa fitur lemah termasuk *symmetry_se* dan *perimeter_worst*.



Rajah 5.16 Graf rangkaian bagi 30 pasangan fitur kurang relevan bagi set data WDBC

Seterusnya, bagi set data WPBC dalam Jadual 5.18, pasangan fitur berinteraksi yang mengandungi fitur kurang relevan termasuk *perimeter2–symmetry3* dan *Time–area3* menunjukkan bahawa terdapat fitur lemah secara individu, namun masih mempunyai nilai apabila wujud dalam interaksi dengan fitur lain yang lebih dominan. Selain itu, pasangan seperti *perimeter2–texture3* dan *texture3–area3* melibatkan dua fitur kurang relevan tetapi mempunyai nilai interaksi yang tinggi.

Jadual 5.18 Pasangan fitur kurang relevan dan purata nilai interaksi SHAP bagi set data WPBC

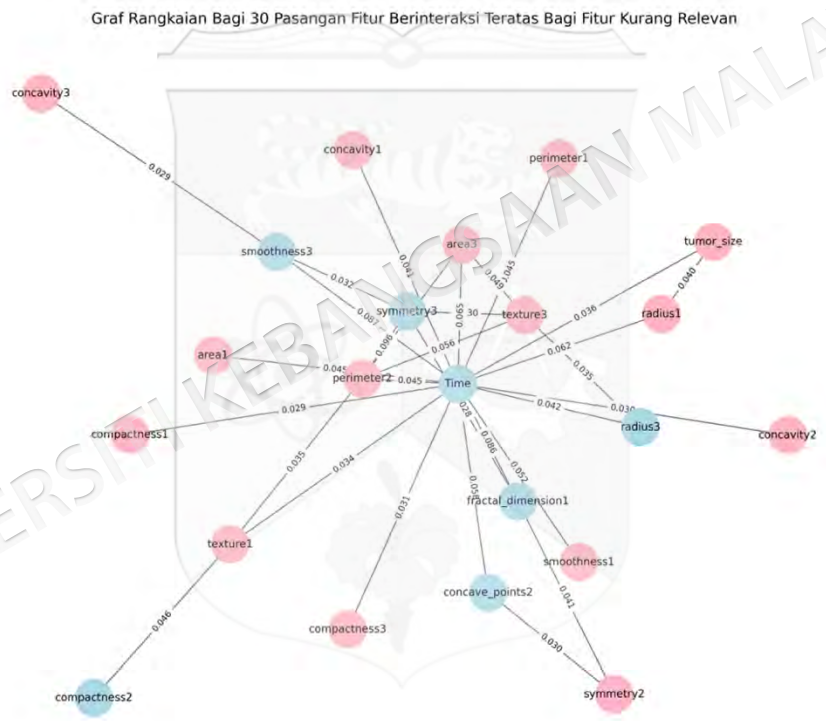
Pasangan Fitur		**Nilai Interaksi SHAP**			**Nilai Purata Mutlak Interaksi**	**Jenis Interaksi**
Fitur Pertama	**Fitur Kedua**	**RF**	**GB**	**XGB**		
perimeter2	***symmetry3***	**0.0027**	**0.2862**	**0.0000**	**0.0963**	**Lemah-Kuat**
Time	***area3***	**-0.0033**	**-0.0878**	**-0.1045**	**0.0652**	**Kuat-Lemah**
Time	*radius1*	-0.0065	-0.1453	-0.0333	0.0617	Kuat-Lemah
perimeter2	***texture3***	**0.0001**	**-0.1668**	**0.0000**	**0.0556**	**Lemah-Lemah**
Time	*smoothness1*	-0.0014	-0.1492	0.0052	0.0519	Kuat-Lemah
texture3	***area3***	**0.0017**	**0.1454**	**0.0000**	**0.0491**	**Lemah-Lemah**

bersambung...

...sambungan

texture1	*compactness2*	-0.0008	-0.1267	-0.0113	0.0463	Lemah-Kuat
Time	*perimeter2*	-0.0038	-0.0959	-0.0353	0.0450	Kuat-Lemah
Time	*perimeter1*	0.0010	-0.1334	0.0000	0.0448	Kuat-Lemah
Time	*area1*	-0.0008	-0.1335	0.0000	0.0447	Kuat-Lemah

Rajah 5.17 menunjukkan fitur lemah iaitu *perimeter2* berinteraksi dengan beberapa fitur kuat yang lain termasuk *Time* dan *symmetry3* serta fitur lemah seperti *area1, compactness1* dan *texture1*. Selain itu, fitur lemah *area3* pula berinteraksi dengan beberapa fitur kuat seperti *Time* dan *symmetry3* serta fitur lemah seperti *texture3*.



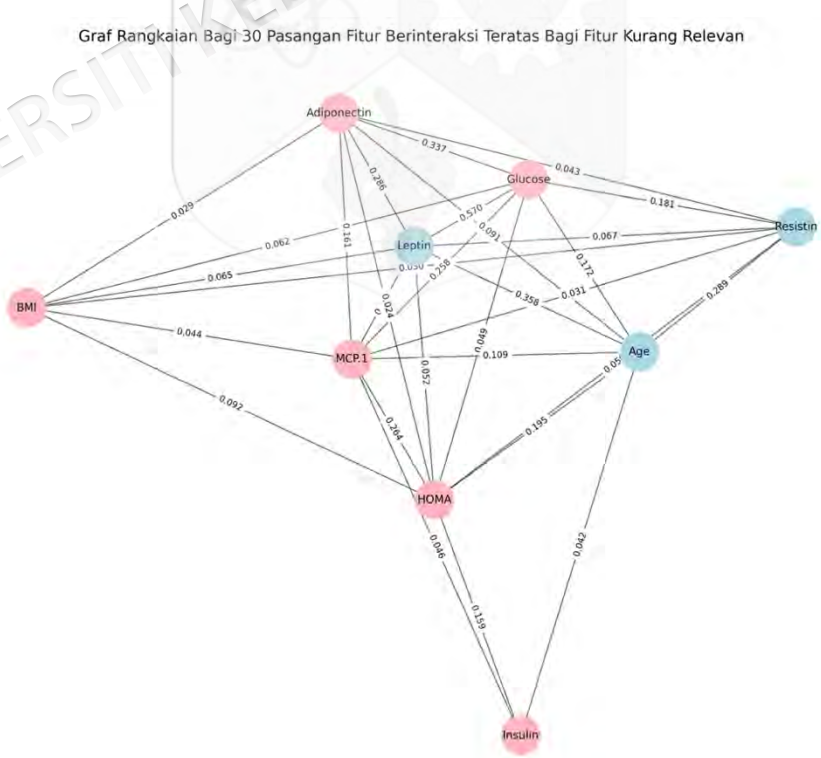
Rajah 5.17 Graf rangkaian bagi 30 pasangan fitur kurang relevan bagi set data WPBC

Bagi set data BCC, Jadual 5.19 memperlihatkan pasangan fitur berinteraksi yang mengandungi fitur kurang relevan termasuk *Glucose–Leptin* dan *Leptin–Adiponectin*. Selain itu, pasangan seperti *Glucose–Adiponectin* dan *HOMA–MCP.1* melibatkan dua fitur kurang relevan tetapi mempunyai nilai interaksi yang tinggi.

Jadual 5.19 Pasangan fitur kurang relevan dan purata nilai interaksi SHAP bagi set data BCC

Pasangan Fitur		Nilai Interaksi SHAP			Purata Nilai Interaksi	Jenis Interaksi
Fitur Pertama	**Fitur Kedua**	**RF**	**GB**	**XGB**		
Glucose	***Leptin***	**0.0224**	**1.6679**	**0.0204**	**0.5703**	**Lemah-Kuat**
Glucose	***Adiponectin***	**0.0125**	**0.9951**	**-0.0042**	**0.3373**	**Lemah-Lemah**
Leptin	***Adiponectin***	**0.0062**	**0.8515**	**0.0000**	**0.2859**	**Kuat-Lemah**
Leptin	*MCP.1*	0.0035	0.8144	0.0044	0.2741	Kuat-Lemah
HOMA	***MCP.1***	**0.0073**	**0.7827**	**-0.0016**	**0.2639**	**Lemah-Lemah**
Glucose	*MCP.1*	0.0202	0.7270	0.0274	0.2582	Lemah-Lemah
Age	*HOMA*	0.0045	0.5710	-0.0087	0.1947	Kuat-Lemah
Glucose	*Resistin*	0.0017	-0.5340	0.0072	0.1810	Lemah-Kuat
Age	*Glucose*	-0.0006	-0.4360	-0.0808	0.1725	Kuat-Lemah
Adiponectin	*MCP.1*	0.0040	0.4774	-0.0009	0.1608	Lemah-Lemah

Rajah 5.18 menunjukkan fitur lemah iaitu *MCP.1* berinteraksi dengan beberapa fitur kuat yang lain termasuk *Leptin*, *Resistin* dan *Age* serta fitur lemah seperti *BMI, Glucose, Insulin* dan *Adiponectin*. Selain itu, fitur lemah *Glucose* pula berinteraksi dengan beberapa fitur kuat seperti *Leptin, Age* dan *Resistin* serta fitur lemah seperti *HOMA, MCP.1 dan Adiponectin.*



Rajah 5.18 Graf rangkaian bagi 30 pasangan fitur kurang relevan bagi set data BCC

5.3 PENGESAHAN INTERAKSI FITUR KURANG RELEVAN BERDASARKAN MODEL SHAP

Seksyen ini mengesahkan sama ada subset fitur terpilih oleh kaedah FD-CARI merangkumi fitur-fitur yang menunjukkan interaksi ketara berdasarkan nilai SHAP. Justeru, subset fitur sebelum penyingkiran fitur bertindih yang diperoleh daripada kaedah FD-CARI seperti pada Jadual 4.16, 4.17 dan 4.18 digunakan sebagai perbandingan dengan fitur berinteraksi daripada model SHAP. Jadual 5.20 menyenaraikan tahap relevansi setiap fitur terpilih bagi setiap set data.

Jadual 5.20 Tahap relevansi fitur terpilih bagi setiap set data

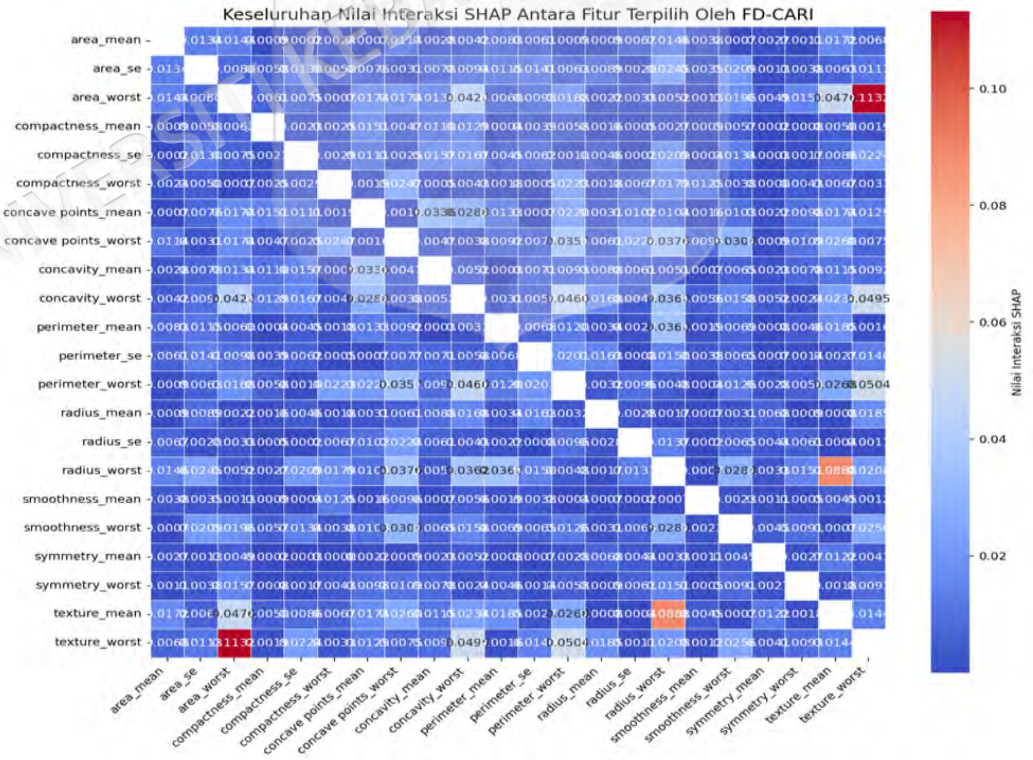
Set Data	Bilangan Fitur	Subset Fitur Sebelum Penyingkiran Fitur Bertindih	Fitur Sangat Relevan	Fitur Kurang Relevan
WDBC	22	*area_mean, concavity_mean, symmetry_worst, concavity_worst, radius_worst, perimeter_se, perimeter_worst, compactness_worst, smoothness_worst, symmetry_mean, texture_worst, area_worst, concave points_worst, concave points_mean, compactness_se, radius_se, compactness_mean, smoothness_mean, area_se, perimeter_mean, texture_mean, radius_mean*	*concavity_worst, smoothness_worst, texture_worst, area_worst, concave points_worst, concave points_mean, texture_mean*	***area_mean, concavity_mean, symmetry_worst, radius_worst, perimeter_se, perimeter_worst, compactness_worst, symmetry_mean, compactness_se, radius_se, compactness_mean, smoothness_mean, area_se perimeter_mean, radius_mean***
WPBC	20	*Time, area2, area3, compactness1, compactness2, compactness3, concave_points1, concave_points3, concavity3, lymph_node_status, perimeter1, perimeter3, radius1, radius2, radius3, smoothness3, symmetry3, texture1, texture3, tumor_size*	*Time, compactness2, radius3, smoothness3, symmetry3*	***area2, area3, compactness1, compactness3, concave_points1, concave_points3, concavity3, lymph_node_status, perimeter1, perimeter3, radius1, radius2, texture1, texture3, tumor_size***
BCC	7	*Age, Glucose, HOMA, Insulin, Leptin, MCP.1, Resistin*	*Age, Leptin, Resistin*	***HOMA, Insulin, MCP.1, Glucose***

Berdasarkan Jadual 5.20, set data WDBC mempunyai 22 fitur unik dengan 7 fitur sangat relevan dan 15 fitur kurang relevan. Untuk set data WPBC pula, 20 fitur dikenal pasti di mana 5 fitur sangat relevan dan 15 fitur kurang relevan. Bagi set data BCC, walaupun pada asalnya terdapat 11 fitur sebelum penyingkiran fitur bertindih,

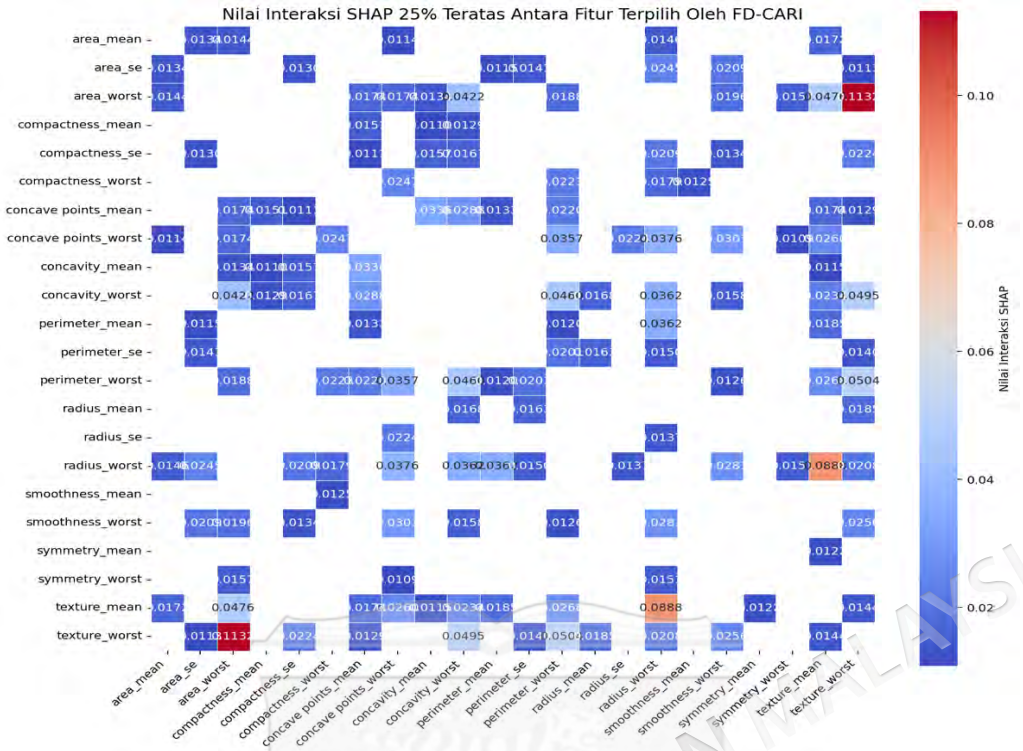
hanya 7 fitur unik digunakan kerana terdapat jenis fitur yang sama namun mempunyai nilai kategori yang berbeza seperti *'Glucose=Low'* dan *'Glucose=Medium'*. Terdapat 3 fitur sangat relevan dan 4 fitur kurang relevan pada set data ini.

5.2.1 Set Data WDBC

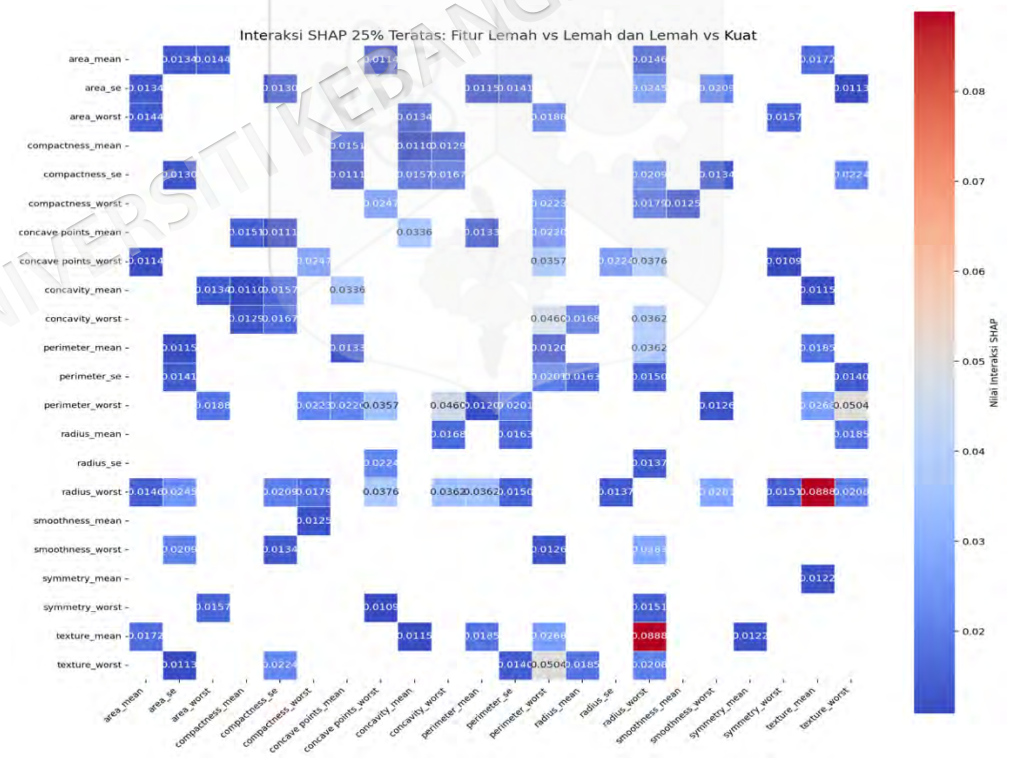
Pengesahan interaksi fitur kurang relevan dilakukan dengan menilai interaksi SHAP pada kuantil ke-75. Rajah 5.19 memaparkan tiga peta haba interaksi SHAP: (a) keseluruhan, (b) hanya kuantil ke -75 dan (c) khusus bagi fitur kurang relevan. Rajah 5.20 (a) dan Rajah 5.20 (b) menonjolkan beberapa pasangan dengan nilai interaksi tertinggi seperti *texture_worst–area_worst* (0.1132), *texture_mean–radius_worst* (0.0888) dan *texture_worst–perimeter_worst* (0.0504) serta *texture_worst–concavity_worst* (0.0495) dan *texture_mean–area_worst* (0.0476). Sementara itu, Rajah 5.20 (c) menunjukkan interaksi fitur kurang relevan dengan fitur dominan seperti *texture_mean–radius_worst*, *texture_worst–perimeter_worst* dan *perimeter_worst–concavity_worst* (0.0460).



(a)



(b)



(c)

Rajah 5.19 Peta haba interaksi bagi (a) keseluruhan (b) kuantil ke-75 (c) fitur kurang relevan bagi set data WDBC

Jadual 5.21 menyenaraikan 15 fitur kurang relevan yang tersenarai dalam kelompok kuantil ke-75 bersama pasangan interaksinya. Antara pasangan paling menonjol adalah *radius_worst–texture_mean* (0.0888), diikuti oleh *texture_worst–perimeter_worst* (0.0504), *perimeter_worst–concavity_worst* (0.0460), *radius_worst–concave_points_worst* (0.0376). Walaupun terdapat pasangan seperti *symmetry_worst–concave_points_worst* yang mencatatkan nilai terendah 0.0109, ia masih dianggap signifikan kerana melepasi ambang kuantil ke-75.

Jadual 5.21 Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-75 dalam set data WDBC

Fitur Kurang Relevan	Bilangan Pasangan	Pasangan Fitur		Nilai Purata Interaksi
		Fitur Pertama	Fitur Kedua	
area_mean	5	texture_mean	area_mean	0.0172
		area_mean	*radius_worst*	0.0146
		area_mean	*area_worst*	0.0144
		area_mean	*area_se*	0.0134
		area_mean	*concave points_worst*	0.0114
compactness_mean	3	*compactness_mean*	*concave points_mean*	0.0151
		compactness_mean	*concavity_worst*	0.0129
		compactness_mean	*concavity_mean*	0.0110
compactness_se	7	*compactness_se*	*texture_worst*	0.0224
		compactness_se	*radius_worst*	0.0209
		compactness_se	*concavity_worst*	0.0167
		concavity_mean	*compactness_se*	0.0157
		compactness_se	*smoothness_worst*	0.0134
		area_se	*compactness_se*	0.0130
		concave points_mean	*compactness_se*	0.0111
compactness_worst	4	*compactness_worst*	*concave points_worst*	0.0247
		perimeter_worst	*compactness_worst*	0.0223
		radius_worst	*compactness_worst*	0.0179
		smoothness_mean	*compactness_worst*	0.0125
concavity_mean	5	*concavity_mean*	*concave points_mean*	0.0336
		concavity_mean	*compactness_se*	0.0157
		concavity_mean	*area_worst*	0.0134
		texture_mean	*concavity_mean*	0.0115
		compactness_mean	*concavity_mean*	0.0110
perimeter_mean	5	*perimeter_mean*	*radius_worst*	0.0362
		texture_mean	*perimeter_mean*	0.0185
		perimeter_mean	*concave points_mean*	0.0133
		perimeter_mean	*perimeter_worst*	0.0120
		perimeter_mean	*area_se*	0.0115
perimeter_se	5	*perimeter_se*	*perimeter_worst*	0.0201
		radius_mean	*perimeter_se*	0.0163

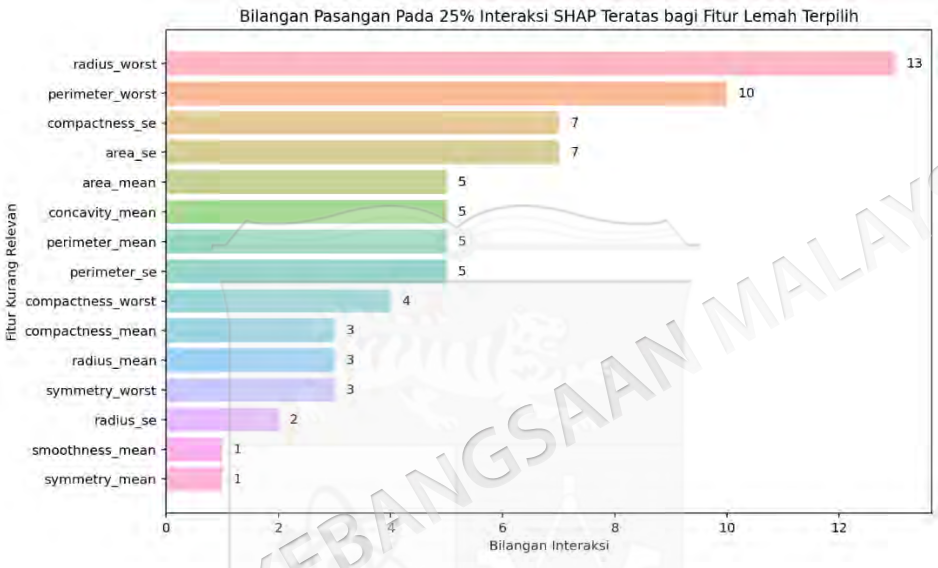
bersambung...

...sambungan

		perimeter_se	*radius_worst*	0.0150
		perimeter_se	*area_se*	0.0141
		perimeter_se	*texture_worst*	0.0140
radius_mean	3	*radius_mean*	*texture_worst*	0.0185
		radius_mean	*concavity_worst*	0.0168
		radius_mean	*perimeter_se*	0.0163
radius_se	2	*radius_se*	*concave points_worst*	0.0224
		radius_se	*radius_worst*	0.0137
smoothness_mean	1	*smoothness_mean*	*compactness_worst*	0.0125
symmetry_mean	1	*texture_mean*	*symmetry_mean*	0.0122
area_se	7	*area_se*	*radius_worst*	0.0245
		area_se	*smoothness_worst*	0.0209
		perimeter_se	*area_se*	0.0141
		area_mean	*area_se*	0.0134
		area_se	*compactness_se*	0.0130
		perimeter_mean	*area_se*	0.0115
		area_se	*texture_worst*	0.0113
perimeter_worst	10	***texture_worst***	***perimeter_worst***	**0.0504**
		perimeter_worst	***concavity_worst***	**0.0460**
		perimeter_worst	*concave points_worst*	0.0357
		texture_mean	*perimeter_worst*	0.0268
		perimeter_worst	*compactness_worst*	0.0223
		concave points_mean	*perimeter_worst*	0.0220
		perimeter_se	*perimeter_worst*	0.0201
		perimeter_worst	*area_worst*	0.0188
		perimeter_worst	*smoothness_worst*	0.0126
		perimeter_mean	*perimeter_worst*	0.0120
radius_worst	13	***texture_mean***	***radius_worst***	**0.0888**
		radius_worst	***concave points_worst***	**0.0376**
		perimeter_mean	*radius_worst*	0.0362
		radius_worst	*concavity_worst*	0.0362
		radius_worst	*smoothness_worst*	0.0283
		area_se	*radius_worst*	0.0245
		compactness_se	*radius_worst*	0.0209
		radius_worst	*texture_worst*	0.0208
		radius_worst	*compactness_worst*	0.0179
		radius_worst	*symmetry_worst*	0.0151
		perimeter_se	*radius_worst*	0.0150
		area_mean	*radius_worst*	0.0146
		radius_se	*radius_worst*	0.0137
symmetry_worst	3	*area_worst*	*symmetry_worst*	0.0157
		radius_worst	*symmetry_worst*	0.0151
		concave points_worst	***symmetry_worst***	**0.0109**

Rajah 5.20 menunjukkan jumlah pasangan interaksi bagi setiap fitur kurang relevan. Dapatan utama termasuk *radius_worst* mempunyai bilangan interaksi tertinggi

(13 pasangan), diikuti *perimeter_worst* (10), *compactness_se* dan *area_se* (7). Manakala a*rea_mean*, *concavity_mean*, *perimeter_mean*, dan *perimeter_se* masing-masingmuncul dalam 5 pasangan. Fitur dengan bilangan pasangan rendah seperti *radius_se*, *smoothness_mean*, dan *symmetry_mean* tetap menunjukkan interaksi signifikan kerana terlibat dengan fitur utama seperti *concave points_worst* dan *texture_mean*.

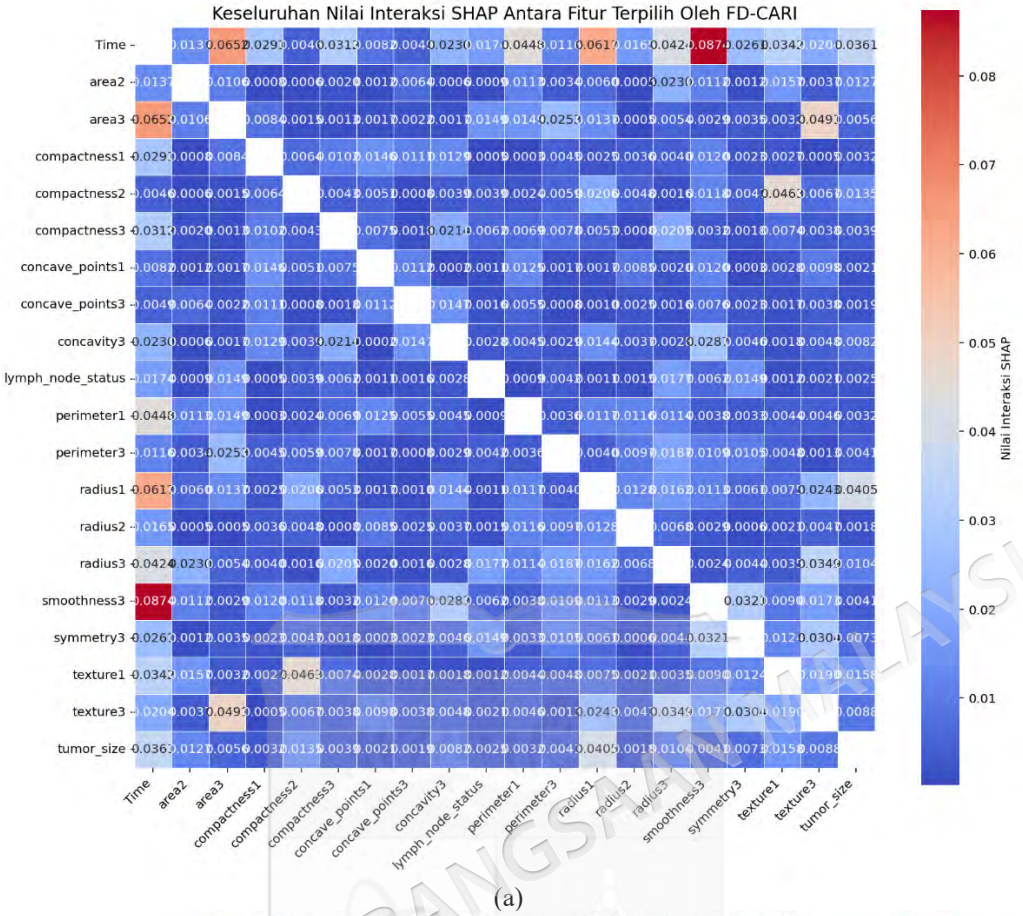


Rajah 5.20 Bilangan pasangan fitur kurang relevan pada kuantil ke-75 bagi set data WDBC

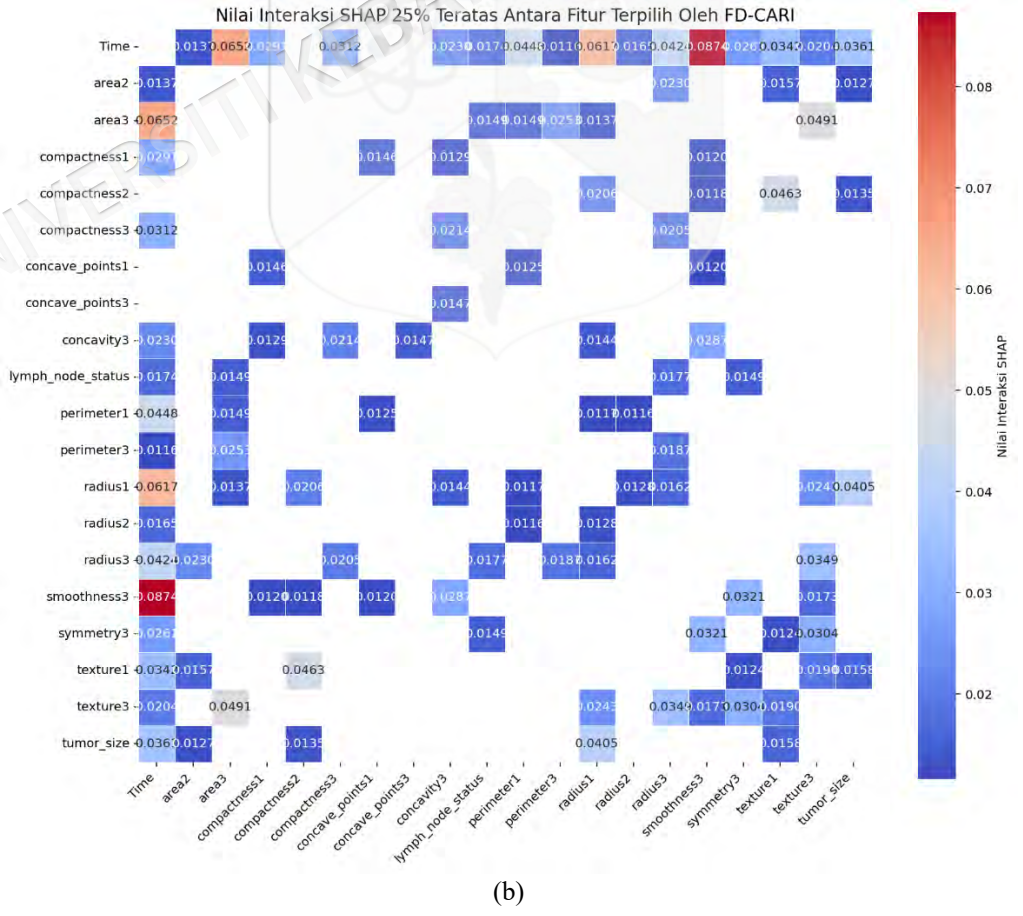
Justeru, setiap fitur kurang relevan yang dipilih oleh kaedah FD-CARI bagi set data WDBC menunjukkan sekurang-kurangnya satu interaksi signifikan dengan fitur lain dan mengesahkan bahawa pemilihannya adalah wajar dari perspektif interaksi.

5.2.2 Set Data WPBC

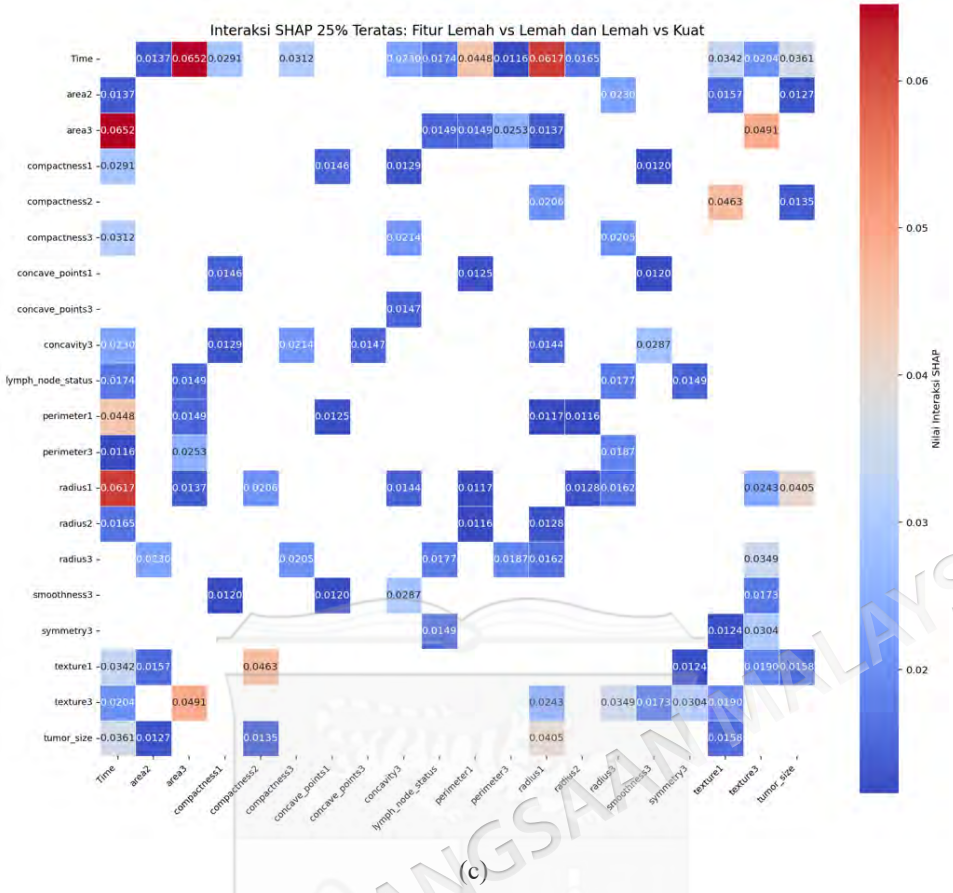
Rajah 5.21 (a) dan Rajah 5.21 (b) menunjukkan pasangan interaksi yang ketara tertumpu kepada *smoothness3–Time* dengan nilai 0.0874, *Time–area3* (0.0652) dan *radius1–Time* (0.0617). Selain itu, interaksi seperti *texture3–area3* (0.0491) dan *texture1–compactness2* (0.0463) turut menunjukkan nilai interaksi yang tinggi. Rajah 5.21 (c) menunjukkan interaksi fitur kurang relevan dengan fitur dominan seperti *radius1–Time* dan *Time–area3* serta pasangan interaksi bagi kedua-dua fitur kurang relevan iaitu *texture3–area3.*



(a)



(b)



(c)

Rajah 5.21 Peta haba interaksi bagi (a) keseluruhan (b) kuantil ke-75 (c) fitur kurang relevan bagi set data WPBC

Berdasarkan Jadual 5.22, nilai interaksi tertinggi dicatatkan oleh pasangan *Time–area3* dengan nilai 0.0652, diikuti oleh *texture3–area3* (0.0491), *Time–compactness3* (0.0312) dan *Time–compactness1* (0.0291). Sebaliknya, nilai interaksi terendah dalam kelompok ini ialah pasangan *perimeter1–radius2* dan *Time–perimeter3* dengan nilai 0.0116 yang masih dianggap signifikan kerana melepasi ambang pemilihan suku teratas.

Jadual 5.22 Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-75 dalam set data WPBC

Fitur Kurang Relevan	Bilangan Pasangan	Pasangan Fitur		Nilai Purata Interaksi
		Fitur Pertama	Fitur Kedua	
area2	4	*area2*	*radius3*	0.0230
		texture1	*area2*	0.0157
		Time	*area2*	0.0137
		area2	*tumor_size*	0.0127
area3	6	***Time***	***area3***	**0.0652**

bersambung...

...sambungan

		texture3	***area3***	**0.0491**
		perimeter3	*area3*	0.0253
		perimeter1	*area3*	0.0149
		area3	*lymph_node_status*	0.0149
		radius1	*area3*	0.0137
compactness1	4	***Time***	***compactness1***	**0.0291**
		compactness1	*concave_points1*	0.0146
		compactness1	*concavity3*	0.0129
		compactness1	*smoothness3*	0.0120
compactness3	3	***Time***	***compactness3***	**0.0312**
		compactness3	*concavity3*	0.0214
		radius3	*compactness3*	0.0205
concave_points1	3	*compactness1*	*concave_points1*	0.0146
		perimeter1	*concave_points1*	0.0125
		concave_points1	*smoothness3*	0.0120
concave_points3	1	*concavity3*	*concave_points3*	0.0147
concavity3	6	*smoothness3*	*concavity3*	0.0287
		Time	*concavity3*	0.0230
		compactness3	*concavity3*	0.0214
		concavity3	*concave_points3*	0.0147
		radius1	*concavity3*	0.0144
		compactness1	*concavity3*	0.0129
lymph_node_status	4	*radius3*	*lymph_node_status*	0.0177
		Time	*lymph_node_status*	0.0174
		symmetry3	*lymph_node_status*	0.0149
		area3	*lymph_node_status*	0.0149
perimeter1	5	*Time*	*perimeter1*	0.0448
		perimeter1	*area3*	0.0149
		perimeter1	*concave_points1*	0.0125
		radius1	*perimeter1*	0.0117
		perimeter1	***radius2***	**0.0116**
perimeter3	3	*perimeter3*	*area3*	0.0253
		radius3	*perimeter3*	0.0187
		Time	***perimeter3***	**0.0116**
radius1	9	*Time*	*radius1*	0.0617
		radius1	*tumor_size*	0.0405
		radius1	*texture3*	0.0243
		radius1	*compactness2*	0.0206
		radius1	*radius3*	0.0162
		radius1	*concavity3*	0.0144

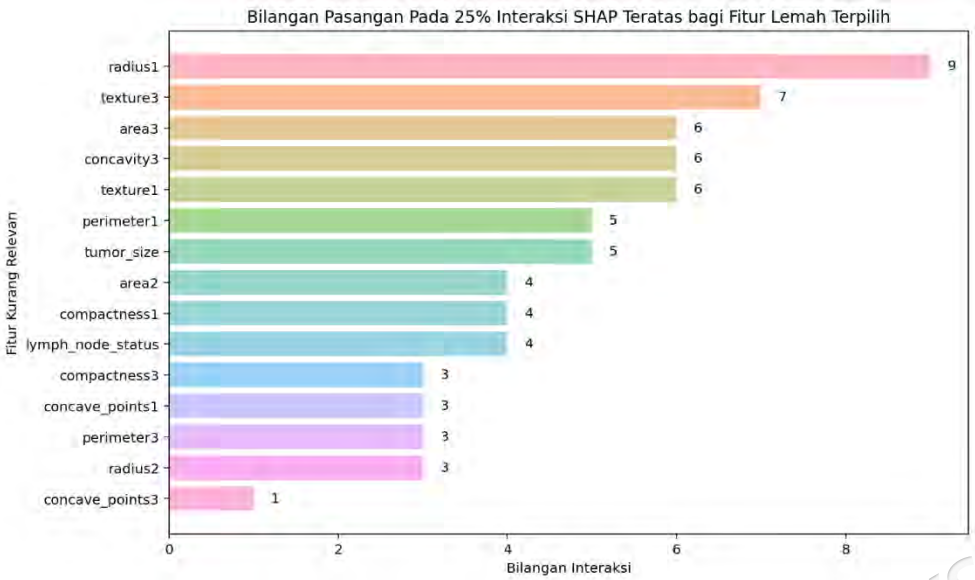
bersambung...

...sambungan

		radius1	*area3*	0.0137
		radius1	*radius2*	0.0128
		radius1	*perimeter1*	0.0117
radius2	3	*Time*	*radius2*	0.0165
		radius1	*radius2*	0.0128
		perimeter1	*radius2*	0.0116
texture1	6	*texture1*	*compactness2*	0.0463
		Time	*texture1*	0.0342
		texture1	*texture3*	0.0190
		texture1	*tumor_size*	0.0158
		texture1	*area2*	0.0157
		texture1	*symmetry3*	0.0124
texture3	7	*texture3*	*area3*	0.0491
		radius3	*texture3*	0.0349
		texture3	*symmetry3*	0.0304
		radius1	*texture3*	0.0243
		Time	*texture3*	0.0204
		texture1	*texture3*	0.0190
		texture3	*smoothness3*	0.0173
tumor_size	5	*radius1*	*tumor_size*	0.0405
		Time	*tumor_size*	0.0361
		texture1	*tumor_size*	0.0158
		compactness2	*tumor_size*	0.0135
		area2	*tumor_size*	0.0127

Rajah 5.22 menunjukkan setiap fitur kurang relevan mempunyai sekurang-kurangnya satu pasangan interaksi dalam kelompok kuantil ke-75. Fitur *radius1* berinteraksi dengan 9 fitur lain, diikuti *texture3* (7), *area3*, *concavity3*, *texture1* dan *perimeter1* (6). Fitur *tumor_size* dan *area2* muncul dalam 5 pasangan, manakala *area2, compactness1 dan lymph_node_status* masing-masing dalam 4. Fitur *compactness3*, *concave_points1*, *perimeter3* dan *radius2* pula masing-masing berinteraksi dengan 3 pasangan. Walaupun *concave_points3* hanya ada 1 pasangan, kehadirannya tetap penting kerana berinteraksi dengan fitur lain seperti *concavity3* serta menunjukkan kesan interaksi yang masih signifikan.

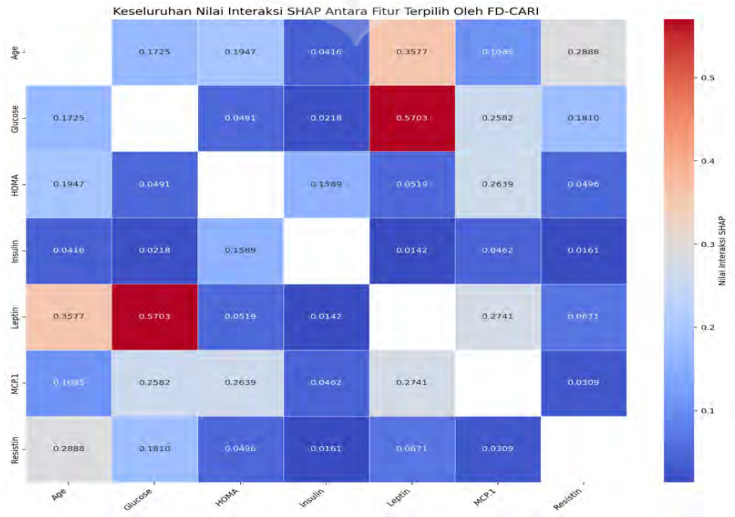
Dapatan ini mengesahkan semua fitur kurang relevan yang dipilih oleh kaedah FD-CARI berinteraksi dengan fitur-fitur lain dalam set data WPBC.



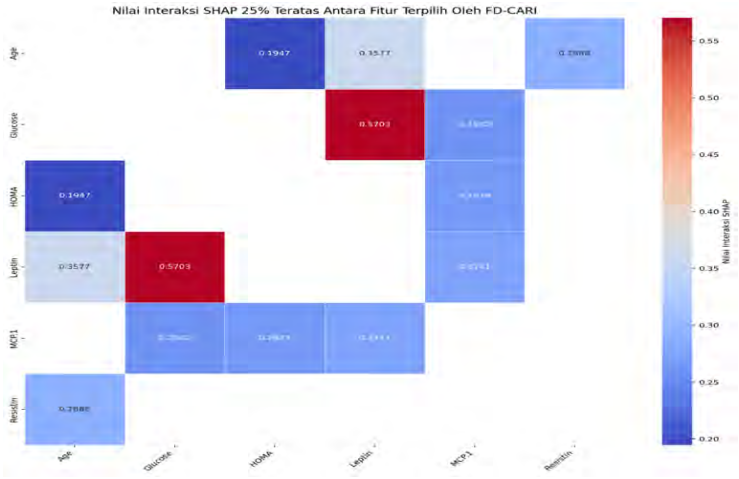
Rajah 5.22 Bilangan pasangan fitur kurang relevan pada kuantil ke-75 bagi set data WPBC

5.2.3 Set Data BCC

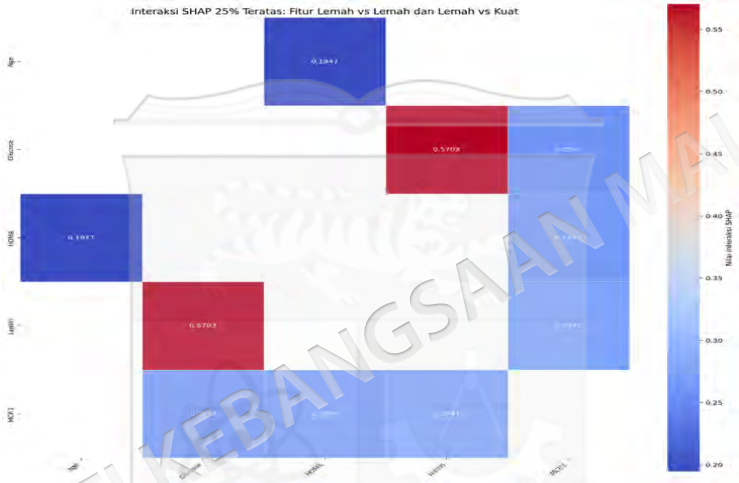
Rajah 5.23 (a) dan Rajah 5.23 (b) menunjukkan pasangan interaksi yang ketara tertumpu kepada *Glucose–Leptin* dengan nilai 0.5703, diikuti oleh *Age–Leptin* (0.3577) dan *HOMA–MCP.1* (0.2639). Selain itu, interaksi seperti *Leptin–MCP.1* (0.2741) dan *Adiponectin–Resistin* (0.2859) turut menunjukkan nilai interaksi yang tinggi. Rajah 5.23 (c) menunjukkan interaksi fitur kurang relevan dengan fitur dominan seperti *Leptin–Glucose* dan *Leptin–MCP.1* serta pasangan interaksi bagi kedua-dua fitur kurang relevan iaitu *HOMA–MCP.1*.



(a)



(b)



(c)

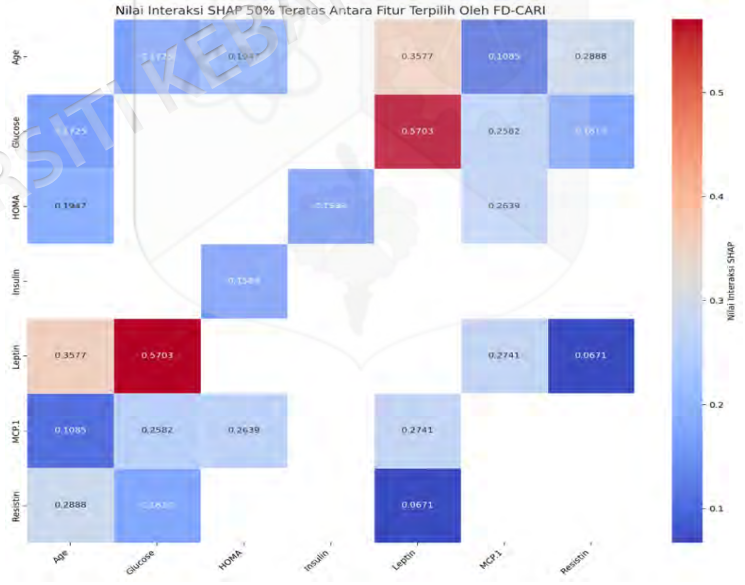
Rajah 5.23 Peta haba interaksi bagi (a) keseluruhan (b) kuantil ke-75 (c) fitur kurang relevan bagi set data BCC

Berdasarkan Jadual 5.23, nilai interaksi tertinggi dicatatkan oleh pasangan *Glucose–Leptin* dengan nilai 0.5703, diikuti oleh *Leptin–MCP.1* (0.2741), *HOMA–MCP.1* (0.2639) dan *Glucose–MCP.1* (0.2582). Sebaliknya, fitur *Insulin* tidak tersenarai dalam kelompok ini menunjukkan potensi interaksinya yang lebih rendah. Memandangkan fitur *Insulin* tidak mempunyai pasangan interaksi dalam kelompok teratas, analisis diteruskan dengan lebih mendalam di mana ambang diturunkan kepada kuantil ke-50 bagi menilai interaksi sederhana yang mungkin masih relevan.

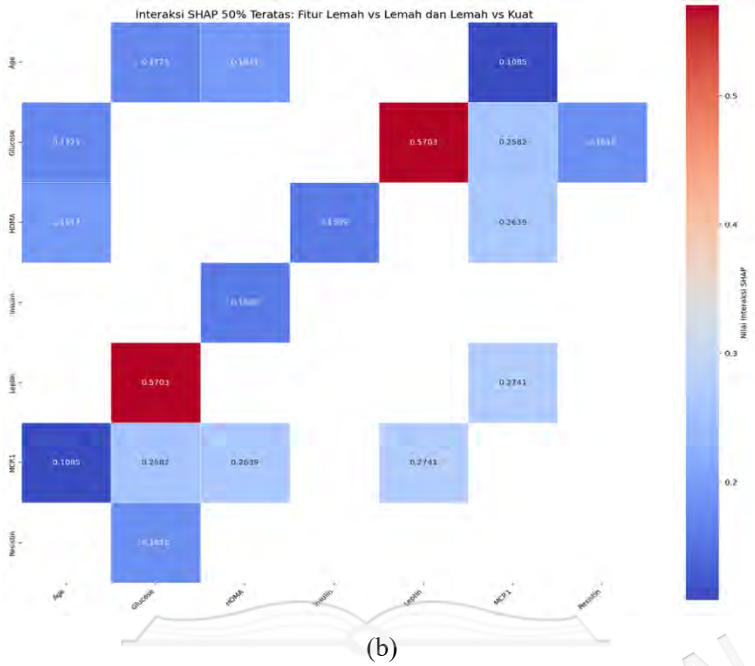
Jadual 5.23 Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-75 dalam set data BCC

Fitur Kurang Relevan	Bilangan Pasangan	Pasangan Fitur		Nilai Purata Interaksi
		Fitur Pertama	Fitur Kedua	
Glucose	2	*Glucose*	*Leptin*	0.5703
		Glucose	*MCP.1*	0.2582
HOMA	2	*HOMA*	*MCP.1*	0.2639
		Age	*HOMA*	0.1947
MCP.1	3	*Leptin*	*MCP.1*	0.2741
		HOMA	*MCP.1*	0.2639
		Glucose	*MCP.1*	0.2582

Rajah 5.24 (a) menunjukkan penambahan pasangan interaksi yang tidak dikenal pasti dalam kelompok kuantil ke-75 seperti *Resistin–Glucose* (0.1810), *Glucose–Age* (0.1725) dan *Insulin–HOMA* (0.1586). Rajah 5.24 (b) juga menunjukkan penambahan interaksi fitur kurang relevan dengan fitur dominan seperti *Resistin–Glucose* dan *Glucose–Age* serta pasangan interaksi bagi kedua-dua fitur kurang relevan iaitu *Insulin–HOMA*.



(a)



(b)

Rajah 5.24 Peta haba interaksi bagi (a) kuantil ke-50 (b) fitur kurang relevan bagi set data BCC

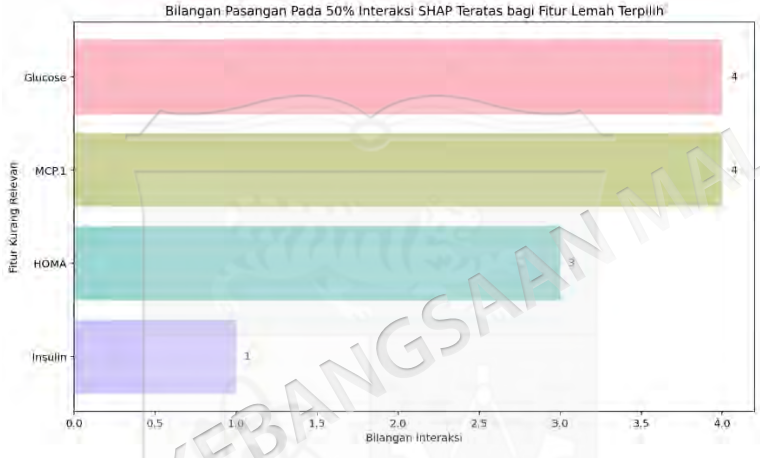
Berdasarkan Jadual 5.24, pasangan tambahan interaksi tertinggi dicatatkan oleh *Glucose–Resistin* (0.1810), *Age–Glucose* (0.1725) dan *Insulin–HOMA* (0.1589). Walaupun tidak tergolong dalam kelompok tertinggi, nilai-nilai ini masih berada dalam julat yang sederhana dan tidak terlalu rendah menunjukkan bahawa interaksi tersebut berpotensi menyumbang kepada prestasi model.

Jadual 5.24 Pasangan fitur melibatkan fitur kurang relevan pada kuantil ke-50 dalam set data BCC

Fitur Kurang Relevan	**Bilangan Pasangan**	**Pasangan Fitur**		**Nilai Purata Interaksi**
		Fitur Pertama	**Fitur Kedua**	
Glucose	4	*Glucose*	*Leptin*	0.5703
		Glucose	*MCP.1*	0.2582
		Glucose	***Resistin***	**0.1810**
		Age	***Glucose***	**0.1725**
HOMA	3	*HOMA*	*MCP.1*	0.2639
		Age	*HOMA*	0.1947
		Insulin	*HOMA*	0.1589
MCP.1	4	*Leptin*	*MCP.1*	0.2741
		HOMA	*MCP.1*	0.2639
		Glucose	*MCP.1*	0.2582
		Age	*MCP.1*	0.1085
Insulin	1	***Insulin***	***HOMA***	**0.1589**

Rajah 5.25 menunjukkan setiap fitur kurang relevan mempunyai sekurang-kurangnya satu pasangan interaksi dalam kelompok kuantil ke-50. Fitur *Glucose* dan *MCP.1* berinteraksi dengan 4 fitur lain, diikuti *HOMA* (3) dan *Insulin* (1). Walaupun *Insulin* hanya ada 1 pasangan, kehadirannya tetap penting kerana berinteraksi dengan fitur lain seperti *HOMA* serta menunjukkan kesan interaksi yang masih signifikan.

Dapatan ini mengesahkan semua fitur kurang relevan yang dipilih oleh kaedah FD-CARI berinteraksi dengan fitur-fitur lain dalam set data BCC.



Rajah 5.25 Bilangan pasangan fitur kurang relevan pada kuantil ke-50 bagi set data BCC

5.4 PERBINCANGAN

Keberkesanan kaedah FD-CARI dalam mengenal pasti fitur interaksi yang telah dianalisis dan disahkan melalui penilaian model SHAP dan teknik pengundian majoriti berwajaran. Hasil pengesahan menunjukkan bahawa FD-CARI berjaya mengekalkan fitur yang mungkin dianggap kurang relevan secara individu tetapi menunjukkan interaksi signifikan dalam kombinasi dengan fitur lain.

Sebagai contoh, set data WDBC menunjukkan interaksi tinggi *perimeter_worst–texture_worst* dan *radius_worst–concave_points_worst,* serta interaksi sesama fitur kurang relevan seperti *texture_mean–radius_worst.* Set data WPBC pula memaparkan interaksi penting antara *Time* dan fitur lemah seperti *area3, compactness3* dan *compactness1,* selain interaksi antara *texture3–area3* yang turut memberi sumbangan signifikan. Set data BCC pula menunjukkan pasangan *Glucose–*

Leptin dan *Leptin–MCP.1* sebagai antara interaksi tertinggi, manakala fitur seperti *Insulin* yang pada mulanya kurang menonjol juga menunjukkan kepentingan apabila dinilai dalam konteks interaksi pada kuantil ke-50.

Dapatan ini membuktikan bahawa fitur kurang relevan secara individu tidak boleh diabaikan sepenuhnya kerana kesan interaksi boleh memberi impak besar terhadap prestasi model. Pendekatan menggunakan ambang kuantil ke-75 dan ke-50 untuk pemilihan interaksi juga adalah selari dengan cadangan dalam kajian terdahulu serta membantu mengekalkan keseimbangan antara kekangan model dan pemeliharaan maklumat interaksi.

Secara keseluruhannya, kaedah FD-CARI terbukti konsisten merentas ketiga-tiga set data yang bukan sahaja dari segi keupayaan menghasilkan subset fitur padat, tetapi juga dalam mengekalkan fitur interaksi yang signifikan yang menyumbang kepada peningkatan prestasi pengelasan. Kaedah ini mengukuhkan justifikasi yang menyumbang kepada peningkatan proses pengelasan serta penggunaan FD-CARI sebagai kaedah pemilihan fitur yang berorientasikan interaksi yang berpotensi untuk digunakan dalam konteks dunia sebenar yang melibatkan fitur-fitur kompleks dan saling berkaitan.

5.5 KESIMPULAN

Bab ini membentangkan hasil analisis terhadap kaedah pengesahan yang digunakan untuk menilai keberkesanan FD-CARI dalam mengenal pasti fitur berinteraksi yang signifikan. Hasil eksperimen ke atas set data WDBC, WPBC dan BCC menunjukkan FD-CARI berjaya mengenal pasti fitur yang berpotensi membentuk interaksi bermakna dan menyumbang kepada prestasi model pengelasan. Melalui analisis nilai interaksi SHAP pada kuantil ke-75 dan ke-50, kajian mendapati bahawa semua fitur kurang relevan yang dipilih mempunyai sekurang-kurangnya satu pasangan interaksi yang kuat atau sederhana menandakan kepentingannya dalam konteks gabungan fitur. Secara keseluruhannya, dapatan ini mengesahkan keupayaan FD-CARI sebagai kaedah pemilihan fitur berasaskan interaksi yang berkesan terutamanya dalam bidang perubatan di mana hubungan antara fitur amat kritikal untuk membina model yang tepat dan boleh dipercayai.

BAB VI

PENAMBAHBAIKAN PRESTASI PEMILIHAN FITUR FD-CARI MELALUI PENGOPTIMUMAN HIPERPARAMETER BERASASKAN PENYEPUHLINDAPAN SIMULASI

6.1 PENGENALAN

Bab ini membentangkan hasil keputusan bagi Objektif 3 yang bertujuan untuk menilai keberkesanan proses penalaan hiperparameter menggunakan algoritma Penyepuhlindapan Simulasi (SA) dalam meningkatkan prestasi kaedah FD-CARI. Penalaan hiperparameter dilakukan ke atas nilai *minSupp*, *minConf*, *minLif, minCF*, *minConv* dan *minLev* agar kombinasi terbaik dapat dikenal pasti secara automatik. Algoritma SA dipilih kerana sifat ruang cariannya yang bernilai diskret dan sensitif terhadap perubahan nilai ambang. SA mempunyai struktur yang lebih ringkas dan tidak memerlukan populasi penyelesaian, berbeza dengan GA dan PSO yang merupakan algoritma metaheuristik berasaskan populasi. GA dan PSO lazimnya memerlukan lebih banyak parameter seperti saiz populasi, kadar mutasi, kadar pembiakan atau parameter inersia yang dapat meningkatkan kos pengiraan serta risiko terperangkap dalam optimum setempat apabila populasi tidak ditetapkan dengan tepat.

Dalam bab ini, hasil pemilihan fitur yang diperoleh melalui penalaan hiperparameter menggunakan algoritma SA (FD-CARI+SA) akan dibandingkan dengan kaedah asal (FD-CARI) daripada aspek:

- Bilangan subset fitur yang terhasil.
- Skor AUC prestasi pengelasan.

Perbandingan kaedah FD-CARI+SA dengan kaedah pemilihan fitur piawai serta ujian signifikan bagi menentukan perbezaan ketara antara kaedah FD-CARI+SA dengan kaedah pemilihan fitur terbaik juga dilakukan ke atas ketiga-tiga set data

WDBC, WPBC dan BCC bagi menilai keberkesanan dan kebolehgunaan kaedah tersebut dalam pelbagai senario pengelasan kanser payudara.

6.2 PENALAAN HIPERPARAMETER MENGGUNAKAN ALGORITMA PENYEPUHLINDAPAN SIMULASI

Bahagian ini melaporkan hasil pengoptimuman nilai α dan β serta *minLift, minCF*, *minConv* dan *minLev* menggunakan SA untuk setiap set data. Proses SA dijalankan secara iteratif untuk mengenal pasti kombinasi parameter yang menghasilkan prestasi pengelasan terbaik yang dinilai berdasarkan purata skor AUC daripada lima model pengelasan. Berikut adalah dapatan utama pada setiap 20 lelaran dan hasil hiperparameter terbaik yang diperoleh. Jadual 6.1, 6.2 dan 6.3 menunjukkan hasil lelaran proses pengoptimuman hiperparameter bagi set data WDBC, WPBC dan BCC masing-masing.

Berdasarkan Jadual 6.1, nilai suhu yang tinggi iaitu 85°C menurun secara berperingkat hingga 3.88°C yang menandakan proses penyejukan terkawal dalam algoritma SA. Dalam masa yang sama, purata skor AUC turut meningkat daripada 0.9785 pada lelaran pertama kepada nilai tertinggi pada lelaran ke-15 sebanyak 0.9860. Pada lelaran terbaik ini, nilai hiperparameter seperti *minSupp*=0.250, *minConf*=0.736, *minLift*=1.291, *minCF*=0.987, *minConv*=46.492 dan *minLev*=0.079 mencerminkan konfigurasi optimum dalam menjana peraturan yang berkualiti untuk membina subset fitur yang menyumbang kepada pengelasan yang tepat. Pelaporan nilai-nilai ini adalah penting untuk memahami bagaimana SA menyumbang kepada penjanaan subset fitur terbaik secara berstruktur dan bukannya secara rawak.

Jadual 6.1 Hasil lelaran proses pengoptimuman hiperparameter bagi set data WDBC

Lelaran	Suhu	*MinSupp*	*MinConf*	*MinLift*	*MinCF*	*MinConv*	*MinLev*	Purata Skor AUC		Masa (saat)
								Skor Semasa	Skor Terbaik	
1	85.00	0.289	0.717	1.242	0.903	46.093	0.073	0.9785	0.9785	24.86
2	72.25	0.265	0.809	1.222	0.959	49.047	0.050	0.9843	0.9843	24.2
3	61.41	0.270	0.768	1.216	0.996	46.683	0.054	0.9798	0.9843	23.32
4	52.20	0.260	0.861	1.273	0.954	49.866	0.067	0.9831	0.9843	27.69
5	44.37	0.283	0.824	1.286	0.958	48.523	0.052	0.9843	0.9843	26.18
6	37.72	0.229	0.716	1.223	0.910	46.390	0.079	0.9796	0.9843	43.52
7	32.06	0.221	0.753	1.294	0.965	48.046	0.058	0.9785	0.9843	64.14
8	27.25	0.238	0.898	1.264	0.956	48.423	0.088	0.9843	0.9843	34.01
9	23.16	0.223	0.706	1.232	0.927	46.055	0.092	0.9796	0.9843	56.28
10	19.69	0.266	0.779	1.291	0.946	46.324	0.061	0.9850	0.9850	28.51
11	16.73	0.226	0.817	1.290	0.940	46.097	0.095	0.9848	0.9850	40.6
12	14.22	0.205	0.722	1.263	0.979	47.111	0.053	0.9795	0.9850	107.27
13	12.09	0.253	0.894	1.286	0.901	48.604	0.081	0.9849	0.9850	35.19
14	10.28	0.227	0.828	1.211	0.943	47.269	0.093	0.9833	0.9850	40.62
15	**8.74**	**0.250**	**0.736**	**1.291**	**0.987**	**46.492**	**0.079**	**0.9860**	**0.9860**	**35.61**
16	7.43	0.215	0.853	1.254	0.978	47.652	0.050	0.9805	0.9860	46.58
17	6.31	0.293	0.876	1.283	0.931	45.290	0.090	0.9851	0.9860	30.97
18	5.36	0.209	0.797	1.207	0.976	48.829	0.056	0.9837	0.9860	61.09
19	4.56	0.227	0.874	1.242	0.921	47.696	0.083	0.9850	0.9860	40.39
20	3.88	0.231	0.899	1.265	0.944	47.588	0.055	0.9843	0.9860	41.79

Berdasarkan Jadual 6.2, purata skor AUC meningkat daripada 0.5843 (lelaran pertama) ke nilai tertinggi 0.7035 yang dicapai pada lelaran ke-6. Pada lelaran ini, nilai hiperparameter yang digunakan adalah *minSupp*=0.302, *minConf*=0.712, *minLift*=1.11, *minCF*=0.328, *minConv*=1.564 dan *minLev*=0.034. Gabungan nilai ini berjaya menghasilkan peraturan dan subset fitur yang menyumbang kepada peningkatan prestasi model.

Terdapat beberapa lelaran dengan skor AUC sebanyak 0.0000 yang menandakan tiada subset fitur yang dipilih disebabkan nilai hiperparameter yang terlalu ketat atau tidak serasi menyebabkan model gagal dibina. dari segi masa pemprosesan, kebanyakan lelaran mengambil masa antara 19 hingga 25 saat. Namun, lelaran ke-11 menunjukkan masa tertinggi iaitu 685.67 saat yang berkemungkinan besar akibat konfigurasi yang memerlukan pengiraan intensif atau bilangan peraturan yang terlalu banyak untuk diproses.

Jadual 6.2 Hasil lelaran proses pengoptimuman hiperparameter bagi set data WPBC

Lelaran	Suhu	*MinSupp*	*MinConf*	*MinLift*	*MinCF*	*MinConv*	*MinLev*	Purata Skor AUC		Masa (saat)
								Skor Semasa	Skor Terbaik	
1	85.00	0.318	0.704	1.142	0.303	1.522	0.035	0.5843	0.5866	19.28
2	72.25	0.313	0.727	1.122	0.359	1.581	0.03	0.5971	0.5971	21.12
3	61.41	0.314	0.717	1.116	0.396	1.534	0.031	0.5893	0.5971	22.66
4	52.20	0.312	0.74	1.173	0.354	1.597	0.034	0.5778	0.5971	20.92
5	44.37	0.312	0.743	1.158	0.37	1.505	0.032	0.5805	0.5971	20.3
6	**37.72**	**0.302**	**0.712**	**1.11**	**0.328**	**1.564**	**0.034**	**0.7035**	**0.7035**	**26.5**
7	32.06	0.304	0.713	1.194	0.365	1.561	0.032	0.0000	0.7035	10.3
8	27.25	0.308	0.749	1.164	0.356	1.568	0.038	0.5892	0.7035	22.04
9	23.16	0.305	0.702	1.132	0.327	1.521	0.039	0.5975	0.7035	24.5
10	19.69	0.306	0.733	1.14	0.391	1.546	0.033	0.5801	0.7035	24.11
11	16.73	0.305	0.729	1.19	0.34	1.522	0.04	0.5683	0.7035	686.67
12	14.22	0.301	0.705	1.163	0.379	1.542	0.031	0.5949	0.7035	23.99
13	12.09	0.32	0.726	1.197	0.386	1.501	0.037	0.0000	0.7035	4.63
14	10.28	0.305	0.732	1.111	0.343	1.545	0.04	0.5750	0.7035	22.87
15	8.74	0.305	0.725	1.118	0.391	1.587	0.033	0.6967	0.7035	23.99
16	7.43	0.312	0.708	1.176	0.354	1.578	0.035	0.5795	0.7035	20.51
17	6.31	0.3	0.746	1.188	0.383	1.531	0.031	0.5701	0.7035	23.85
18	5.36	0.302	0.724	1.107	0.376	1.577	0.031	0.7034	0.7035	25.72
19	4.56	0.311	0.713	1.187	0.342	1.521	0.035	0.0000	0.7035	7.43
20	3.88	0.306	0.75	1.165	0.344	1.552	0.031	0.5892	0.7035	22.96

Jadual 6.3 menunjukkan purata skor AUC meningkat daripada 0.7735 pada lelaran pertama hingga mencapai skor terbaik 0.8044 pada lelaran ke-3. Pada lelaran ini, nilai hiperparameter yang digunakan adalah *minSupp*=0.082, *minConf*=0.702, *minLift*=1.661, *minCF*=0.94, *minConv*=inf dan *minLev*=0.027. Nilai *minConv*=inf berlaku disebabkan nilai keyakinan=1 di mana konsekuen sepenuhnya bergantung kepada anteseden. Nilai infiniti ini berlaku disebabkan oleh nilai penyebut pada formula pada persamaan ...(3.18) menjadi 0. Masa pemprosesan bagi setiap lelaran adalah stabil iaitu di antara 13.85 saat hingga 18.84 saat,dengan majoriti berada dalam julat 15-17 saat. Namun, lelaran ke-15 mencatatkan masa yang sedikit panjang iaitu 18.84 saat, namun masih dalam had boleh terima.

Jadual 6.3 Hasil lelaran proses pengoptimuman hiperparameter bagi set data BCC

Lelaran	Suhu	*MinSupp*	*MinConf*	*MinLift*	*MinCF*	*MinConv*	*MinLev*	Purata Skor AUC		Masa (saat)
								Skor Semasa	Skor Terbaik	
1	85.00	0.069	0.968	1.517	0.884	inf	0.021	0.7735	0.7735	13.85
2	72.25	0.053	0.708	1.54	0.93	inf	0.031	0.6634	0.7735	14.48
3	**61.41**	**0.082**	**0.702**	**1.661**	**0.94**	**inf**	**0.027**	**0.8044**	**0.8044**	**16.44**
4	52.20	0.096	0.801	1.519	0.819	inf	0.037	0.6645	0.8044	16.99
5	44.37	0.074	0.861	1.695	0.876	inf	0.031	0.7420	0.8044	15.18
6	37.72	0.064	0.959	1.615	0.941	inf	0.021	0.7749	0.8044	15.47
7	32.06	0.032	0.724	1.547	0.82	inf	0.026	0.7907	0.8044	16.11
8	27.25	0.04	0.811	1.542	0.853	inf	0.039	0.7796	0.8044	15.35
9	23.16	0.021	0.919	1.533	0.876	inf	0.04	0.7795	0.8044	16.13
10	19.69	0.07	0.953	1.655	0.846	inf	0.021	0.7420	0.8044	15.95
11	16.73	0.025	0.983	1.675	0.863	inf	0.033	0.7749	0.8044	16.57
12	14.22	0.092	0.838	1.553	0.849	inf	0.031	0.8044	0.8044	16.76
13	12.09	0.061	0.969	1.58	0.844	inf	0.04	0.7843	0.8044	15.72
14	10.28	0.009	0.733	1.625	0.958	inf	0.028	0.7749	0.8044	17.22
15	8.74	0.1	0.859	1.694	0.972	inf	0.02	0.6702	0.8044	18.84
16	7.43	0.056	0.78	1.628	0.822	inf	0.029	0.6772	0.8044	15.75
17	6.31	0.096	0.963	1.553	0.9	inf	0.024	0.6997	0.8044	17.67
18	5.36	0.088	0.79	1.628	0.922	inf	0.023	0.8044	0.8044	16.56
19	4.56	0.056	0.934	1.606	0.8	inf	0.026	0.7786	0.8044	17.17
20	3.88	0.088	0.949	1.562	0.812	inf	0.038	0.8044	0.8044	15.75

6.3 PRESTASI MODEL PENGELASAN BERDASARKAN LELARAN

Jadual 6.4 memaparkan lima lelaran dengan purata AUC tertinggi bagi set data WDBC. Berdasarkan jadual tersebut, lelaran ke-15 mencatatkan purata AUC tertinggi 0.9860 dengan model RF mencatatkan skor tertinggi sebanyak 0.9920 ± 0.0048, diikuti oleh model NB (0.9916 ± 0.0054) dan SVM (0.9913 ± 0.0048). Model LR (0.9870 ± 0.0067) turut menunjukan prestasi kompetitif, manakala DT mencatatkan skor lebih rendah (0.9681 ± 0.0132), namun masih dalam julat berprestasi tertinggi. Hasil ini menyokong keputusan pemilihan lelaran ke-15 sebagai wakil prestasi terbaik untuk set data WDBC. Keseluruhan prestasi skor AUC bagi 20 lelaran boleh dirujuk pada Lampiran B.

Jadual 6.4 5 lelaran dengan skor AUC tertinggi bagi set data WDBC

Lelaran	**Model Pengelasan**					**Purata Skor AUC**
	DT	**RF**	**NB**	**LR**	**SVM**	
15	**0.9681 ± 0.0132**	**0.9920 ± 0.0048**	**0.9916 ± 0.0054**	**0.9870 ± 0.0067**	**0.9913 ± 0.0048**	**0.9860**
17	0.9618 ± 0.0114	0.9946 ± 0.0031	0.9891 ± 0.0039	0.9874 ± 0.0061	0.9927 ± 0.0033	0.9851
10	0.9675 ± 0.0155	0.9915 ± 0.0048	0.9873 ± 0.0054	0.9867 ± 0.0063	0.9919 ± 0.0047	0.9850
19	0.9660 ± 0.0110	0.9932 ± 0.0036	0.9861 ± 0.0043	0.9864 ± 0.0048	0.9935 ± 0.0032	0.9850
13	0.9658 ± 0.0108	0.9944 ± 0.0033	0.9844 ± 0.0051	0.9870 ± 0.0064	0.9927 ± 0.0045	0.9849

Seterusnya, Jadual 6.5 mencatatkan purata skor AUC tertinggi pada lelaran ke-6 dan 18 iaitu 0.7035 dengan model SVM mencatatkan skor tertinggi sebanyak 0.7429 ± 0.0568, diikuti oleh model LR (0.7400 ± 0.0552) dan RF (0.7069 ± 0.0890). Sebaliknya, model NB (0.7061 ± 0.0524) menunjukan prestasi yang setanding, manakala DT (0.6214 ± 0.0872) menunjukkan skor lebih rendah, namun masih memberi sumbangan pada purata keseluruhan. Keputusan ini memperkukuhkan pemilihan lelaran ke-18 sebagai konfigurasi terbaik untuk set data WPBC.

Jadual 6.5 5 lelaran dengan skor AUC tertinggi bagi set data WPBC

Lelaran	Model Pengelasan					Purata Skor AUC
	DT	RF	NB	LR	SVM	
6 dan 18	**0.6214 ± 0.0872**	**0.7069 ± 0.0890**	**0.7061 ± 0.0524**	**0.7400 ± 0.0552**	**0.7429 ± 0.0568**	**0.7035**
15	0.6171 ± 0.0924	0.6891 ± 0.0902	0.7031 ± 0.0710	0.7457 ± 0.0506	0.7288 ± 0.0644	0.6967
9	0.5240 ± 0.0862	0.5521 ± 0.0771	0.6080 ± 0.0462	0.6540 ± 0.0683	0.6497 ± 0.0674	0.5975
2	0.5217 ± 0.0556	0.5785 ± 0.0512	0.6191 ± 0.0714	0.6331 ± 0.0649	0.6329 ± 0.0745	0.5971
4	0.5243 ± 0.0439	0.5126 ± 0.0509	0.6352 ± 0.0573	0.6489 ± 0.0652	0.5680 ± 0.1538	0.5778
12	0.5104 ± 0.0984	0.4961 ± 0.0768	0.6463 ± 0.0580	0.6525 ± 0.0790	0.6695 ± 0.0658	0.5950

Jadual 6.6 mencatatkan purata skor AUC tertinggi pada lelaran ke-3, 12, 18 dan 20 iaitu 0.8044 dengan model SVM mencatatkan skor tertinggi sebanyak 0.8439 ± 0.0717, diikuti oleh model RF (0.8412 ± 0.0602) dan LR (0.7922 ± 0.0894). Sebaliknya, model NB (0.7773 ± 0.0861) dan DT (0.7676 ± 0.0712) menunjukkan prestasi sedikit lebih rendah berbanding model-model lain.

Jadual 6.6 5 lelaran dengan skor AUC tertinggi bagi set data BCC

Lelaran	Model Pengelasan					Purata Skor AUC
	DT	RF	NB	LR	SVM	
3,12,18 dan 20	**0.7676 ± 0.0712**	**0.8412 ± 0.0602**	**0.7773 ± 0.0861**	**0.7922 ± 0.0894**	**0.8439 ± 0.0717**	**0.8044**
7	0.7326 ± 0.0794	0.8183 ± 0.0715	0.7831 ± 0.0461	0.7922 ± 0.0623	0.8273 ± 0.0516	0.7907
13	0.7020 ± 0.0595	0.8201 ± 0.0779	0.7941 ± 0.0434	0.7947 ± 0.0723	0.8107 ± 0.0578	0.7843
8	0.7035 ± 0.0792	0.8222 ± 0.0781	0.7658 ± 0.0611	0.7872 ± 0.0785	0.8195 ± 0.0609	0.7796
9	0.7035 ± 0.0792	0.8222 ± 0.0781	0.7658 ± 0.0611	0.7872 ± 0.0785	0.8189 ± 0.0597	0.7795

6.4 BILANGAN SUBSET FITUR TERPILIH BERDASARKAN LELARAN TERBAIK

Jadual 6.7 hingga 6.9 menunjukkan bilangan fitur selepas pemurnian subset fitur bagi set data WDBC, WPBC dan BCC. Bagi set data WDBC, bilangan fitur yang dipilih berubah daripada 20 fitur awal kepada 3 fitur iaitu *concave points_worst, radius_worst*

dan *texture_worst* selepas proses pemurnian subset lelaran terbaik ke-15. Keseluruhan subset fitur akhir selepas pemurnian bagi 20 lelaran boleh dirujuk pada Lampiran B.

Jadual 6.7 Bilangan fitur selepas pemurnian subset fitur bagi set data WDBC

Lelaran	Bilangan Subset Fitur Awal	Subset Fitur Akhir Selepas Pemurnian Subset Fitur	
		Bilangan Fitur	Senarai Fitur
15	20	3	[*'concave points_worst', 'radius_worst', 'texture_worst'*]

Manakala, untuk set data WPBC, bilangan fitur adalah 4 iaitu *Time, radius2, radius3* dan *tumor_size* selepas pemurnian subset fitur leraian terbaik pada lelaran ke-6 dan 18.

Jadual 6.8 Bilangan fitur selepas pemurnian subset fitur bagi set data WPBC

Lelaran	Bilangan Subset Fitur Awal	Subset Fitur Akhir Selepas Pemurnian Subset Fitur	
		Bilangan Fitur	Senarai Fitur
6 dan 18	18	4	[*'Time', 'radius2', 'radius3', 'tumor_size'*]

Bagi set data BCC, lelaran terbaik adalah pada lelaran ke-3, 12, 18 dan 20. Daripada 9 fitur awal, 4 fitur yang sama telah dipilih dalam setiap lelaran terbaik ini iaitu *Age, Glucose, Leptin* dan *Resistin.*

Jadual 6.9 Bilangan fitur selepas pemurnian subset fitur bagi set data BCC

Lelaran	Bilangan Subset Fitur Awal	Subset Fitur Akhir Selepas Pemurnian Subset Fitur	
		Bilangan Fitur	Senarai Fitur
3, 12, 18 dan 20	11	4	['Age', 'Glucose', 'Leptin', 'Resistin']

6.5 PERBANDINGAN PRESTASI FD-CARI SEBELUM DAN SELEPAS PENGOPTIMUMAN HIPERPARAMETER MENGGUNAKAN SA

Seksyen ini membincangkan perbandingan prestasi kaedah FD-CARI dan kaedah yang dioptimumkan menggunakan algoritma SA (FD-CARI+SA).

6.5.1 Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman bagi Set Data WDBC

Berdasarkan jadual 6.10, kaedah FD-CARI memilih 4 fitur iaitu *concave points_worst, perimeter_mean, radius_worst* dan *texture_worst.* Sebaliknya, kaedah FD-CARI+SA hanya mengekalkan 3 fitur iaitu *concave points_worst, radius_worst* dan *texture_worst*. Walaupun bilangan fitur yang dipilih oleh FD-CARI+SA lebih sedikit, kaedah ini menunjukkan prestasi pengelasan yang lebih baik dengan purata skor AUC sebanyak 0.9860 berbanding 0.9857 pada kaedah asal. Ini menunjukkan bahawa proses pengoptimuman hiperparameter melalui SA berjaya menyaring subset fitur yang lebih padat tanpa menjejaskan, malah meningkatkan keupayaan model dalam membezakan antara kelas.

Jadual 6.10 Perbandingan prestasi bagi set data WDBC

Penilaian	FD-CARI	FD-CARI + SA
Bilangan fitur dipilih	4	3
Senarai fitur terpilih	*concave points_worst, perimeter_mean, radius_worst, texture_worst*	*concave points_worst, radius_worst, texture_worst*
Purata skor AUC	0.9857	0.9860
Hiperparameter terbaik	*minSupp* = 0.2, *minConf* =0.8, *minLift*=1.5, *minCF*=0.8, *minConv*=50, *minLev*=0.1	*minSupp* = 0.25, *minConf* =0.736, *minLift*=1.291, *minCF*=0.987, *minConv*=46.492, *minLev*=0.079

Berdasarkan Jadual 6.11, model RF menunjukkan peningkatan yang paling ketara dengan skor purata AUC kepada 0.9920 ± 0.0048, diikuti oleh model NB pada 0.9916 ± 0.0054. Namun, model DT, LR dan SVM mencatatkan skor AUC sedikit rendah pada kaedah FD-CARI+SA masing-masing sebanyak 0.9681 ± 0.0132, 0.9870 ± 0.0067 dan 0.9913 ± 0.0048. Ini berlaku disebabkan oleh kaedah FD-CARI+SA menggunakan subset fitur yang lebih kecil berbanding kaedah FD-CARI.

Jadual 6.11 Purata skor AUC bagi setiap model pengelasan bagi set data WDBC

Model Pengelasan	FD-CARI	FD-CARI + SA
DT	**0.9683 ± 0.0115**	0.9681 ± 0.0132
RF	0.9853 ± 0.0057	**0.9920 ± 0.0048**
NB	0.9911 ± 0.0049	**0.9916 ± 0.0054**
LR	**0.9919 ± 0.0049**	0.9870 ± 0.0067
SVM	**0.9917 ± 0.0049**	0.9913 ± 0.0048
Purata	0.9857	**0.9860**

6.5.2 Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman bagi Set Data WPBC

Berdasarkan Jadual 6.12, kaedah FD-CARI memilih 5 fitur iaitu *Time*, *area3*, *tumor_size*, *compactness2* dan *radius3* manakala kaedah FD-CARI+SA hanya memilih 4 fitur iaitu *Time, radius2, radius3* dan *tumor_size*. Walaupun bilangan fitur yang dipilih lebih sedikit, kaedah FD-CARI+SA masih memberikan purata skor AUC yang lebih tinggi iaitu 0.7035 berbanding 0.7015 untuk kaedah FD-CARI.

Jadual 6.12 Perbandingan prestasi bagi set data WPBC

Penilaian	FD-CARI	FD-CARI + SA
Bilangan fitur dipilih	5	4
Senarai fitur terpilih	*Time, area3, tumor_size, compactness2, radius3*	*Time, radius2, radius3, tumor_size*
Purata skor AUC	0.7015	0.7035
Hiperparameter terbaik	*minSupp* = 0.3, *minConf* =0.7, *minLift*=1.1, *minCF*=0.3, *minConv*=1.5, *minLev*=0.03	*minSupp* = 0.302, *minConf* =0.712, *minLift*=1.11, *minCF*=0.328, *minConv*=1.564, *minLev*=0.034

Berdasarkan Jadual 6.13, model RF menunjukkan peningkatan yang paling ketara dengan skor purata AUC kepada 0.7069 ± 0.0890, diikuti oleh model NB pada 0.7061 ± 0.0524, LR (0.7400 ± 0.0552) dan SVM (0.7429 ± 0.0568). Namun, model DT mencatatkan skor AUC sedikit rendah pada kaedah FD-CARI+SA sebanyak 0.6214 ± 0.0872. Ini berlaku disebabkan oleh kaedah FD-CARI+SA menggunakan subset fitur yang lebih kecil berbanding kaedah FD-CARI.

Jadual 6.13 Purata skor AUC bagi setiap model pengelasan bagi set data WPBC

Model Pengelasan	FD-CARI	FD-CARI + SA
DT	**0.6865 ± 0.0582**	0.6214 ± 0.0872
RF	0.6685 ± 0.0789	**0.7069 ± 0.0890**
NB	0.6789 ± 0.0839	**0.7061 ± 0.0524**
LR	0.7383 ± 0.0544	**0.7400 ± 0.0552**
SVM	0.7352 ± 0.0740	**0.7429 ± 0.0568**
Purata	0.7015	**0.7035**

6.5.3 Prestasi FD-CARI Sebelum dan Selepas Pengoptimuman bagi Set Data BCC

Berdasarkan Jadual 6.12, kaedah FD-CARI dan FD-CARI+SA memilih 4 fitur fitur yang sama *iaitu Age, Glucose, Leptin*. dan *Resistin*. Kedua-dua kaedah ini mencatatkan prestasi purata skor AUC yang sama iaitu 0.8044.

Jadual 6.14 Perbandingan prestasi bagi set data BCC

Penilaian	FD-CARI	FD-CARI + SA
Bilangan fitur dipilih	4	4
Senarai fitur terpilih	*Age, Glucose, Leptin, Resistin*	*Age, Glucose, Leptin, Resistin*
Purata skor AUC	0.8044	0.8044
Hiperparameter terbaik	*minSupp* = 0.05, *minConf* =0.7, *minLift*=1.5, *minCF*=1, *minConv*=inf, *minLev*=0.1	*minSupp* = 0.082, *minConf* =0.702, *minLift*=1.661, *minCF*=0.94, *minConv*=inf, *minLev*=0.027

Berdasarkan Jadual 6.15, model SVM mencatatkan skor purata AUC tertinggi iaitu 0.8439 ± 0.0717, diikuti oleh RF (0.8412 ± 0.0602), LR (0.7922 ± 0.0894) dan NB (0.7773 ± 0.0861). Sebaliknya model DT menunjukkan prestasi terendah dengan skor AUC 0.7676 ± 0.0712. Walaupun prestasi bagi kedua-dua kaedah adalah sama, kaedah FD-CARI+SA memberikan kelebihan dari segi kecekapan kerana tidak memerlukan proses penalaan manual yang memakan masa.

Jadual 6.15 Purata skor AUC bagi setiap model pengelasan bagi set data BCC

Model Pengelasan	FD-CARI	FD-CARI + SA
DT	0.7676 ± 0.0712	0.7676 ± 0.0712
RF	0.8412 ± 0.0602	0.8412 ± 0.0602
NB	0.7773 ± 0.0861	0.7773 ± 0.0861
LR	0.7922 ± 0.0894	0.7922 ± 0.0894
SVM	0.8439 ± 0.0717	0.8439 ± 0.0717
Purata	0.8044	0.8044

6.6 PERBANDINGAN MODEL PEMILIHAN FITUR

Seksyen ini membandingkan prestasi skor AUC kaedah FD-CARI+SA dengan beberapa kaedah pemilihan fitur piawai pada setiap set data.

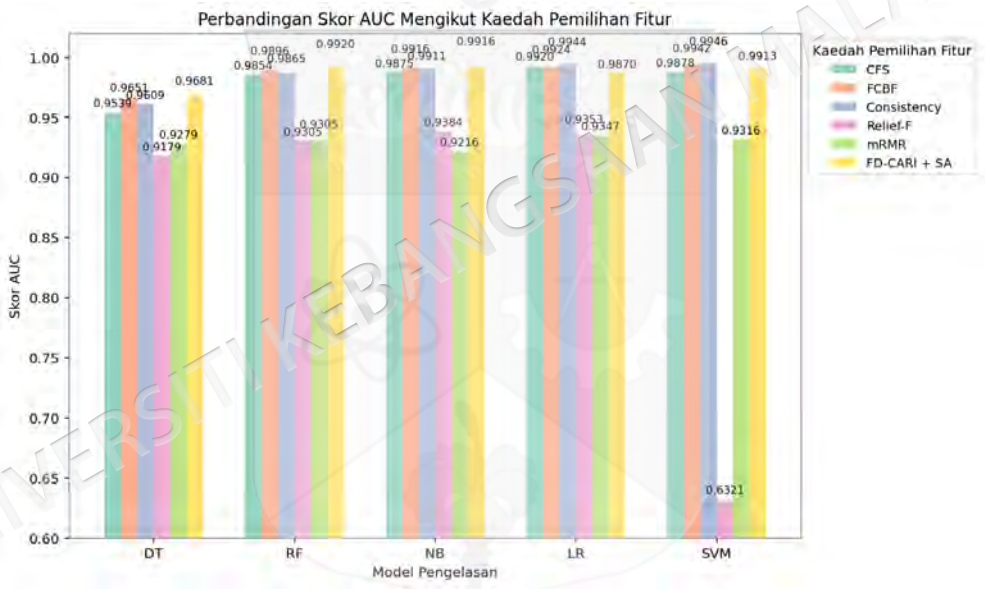
6.6.1 Perbandingan Prestasi FD-CARI+SA dengan Kaedah Pemilihan Fitur Piawai bagi Set Data WDBC

Keputusan perbandingan dalam Jadual 6.16 dan Rajah 6.1 menunjukkan kaedah FD-CARI+SA memberikan prestasi yang kompetetif berbanding kaedah pemilihan piawai yang lain, namun ia bukan sentiasa mencatatkan skor tertinggi dalam semua model pengelasan. Sebagai contoh, kaedah FD-CARI+SA (0.9860) adalah hampir setara dengan FCBF (0.9866) dan Consistency (0.9855) menunjukkan prestasi yang stabil dan boleh dipercayai. FD-CARI+SA menunjukkan kelebihan khususnya dalam model DT, RF dan NB di mana ia mencatatkan skor AUC tertinggi. Ini membuktikan kekuatan kaedah ini apabila digabungkan dengan model-model berasaskan struktur keputusan dan kebarangkalian. Walau bagaimanapun, dalam model LR dan SVM, kaedah Consistency mencatatkan prestasi yang lebih tinggi. Ini menunjukkan bahawa keberkesanan sesuatu kaedah pemilihan fitur boleh dipengaruhi oleh jenis model pengelasan yang digunakan.

Menariknya, kaedah Relief-F dan mRMR yang diuji menggunakan jumlah fitur yang sama seperti FD-CARI+SA mencatatkan purata skor AUC yang jauh lebih rendah (masing-masing 0.8708 dan 0.9293) sekali gus menunjukkan bahawa bilangan fitur yang kecil sahaja tidak mencukupi jika hubungan interaksi tidak dipertimbangkan secara tepat. Secara keseluruhan, FD-CARI+SA telah membuktikan prestasi yang seimbang, konsisten dan kompetitif merentas pelbagai model serta kelebihannya terserlah khususnya dalam pengendalian fitur berinteraksi yang merupakan sesuatu kriteria yang kurang dititikberatkan oleh kaedah pemilihan fitur piawai yang lain.

Jadual 6.16 Skor AUC setiap kaedah pemilihan fitur bagi set data WDBC

Kaedah Pemilihan Fitur	Bilangan Fitur Terpilih	Nilai Skor AUC ± SD					Purata AUC
		DT	RF	NB	LR	SVM	
CFS	11	0.9539 ± 0.0141	0.9854 ± 0.0060	0.9875 ± 0.0035	0.9920 ± 0.0035	0.9878 ± 0.0054	0.9813
FCBF	7	0.9651 ± 0.0124	0.9896 ± 0.0056	**0.9916 ± 0.0039**	0.9924 ± 0.0029	0.9942 ± 0.0035	**0.9866**
Consistency	8	0.9609 ± 0.0159	0.9865 ± 0.0048	0.9911 ± 0.0012	**0.9944 ± 0.0031**	**0.9946 ± 0.0027**	0.9855
Relief-F	3	0.9179 ± 0.0199	0.9305 ± 0.0209	0.9384 ± 0.0124	0.9353 ± 0.0205	0.6321 ± 0.0039	0.8708
mRMR	3	0.9279 ± 0.0131	0.9305 ± 0.0138	0.9216 ± 0.0063	0.9347 ± 0.0117	0.9316 ± 0.0145	0.9293
FD-CARI + SA	3	**0.9681 ± 0.0132**	**0.9920 ± 0.0048**	**0.9916 ± 0.0054**	0.9870 ± 0.0067	0.9913 ± 0.0048	0.9860



Rajah 6.1 Perbandingan skor AUC bagi set data WDBC

6.6.2 Perbandingan Prestasi FD-CARI+SA dengan Kaedah Pemilihan Fitur Piawai bagi Set Data WPBC

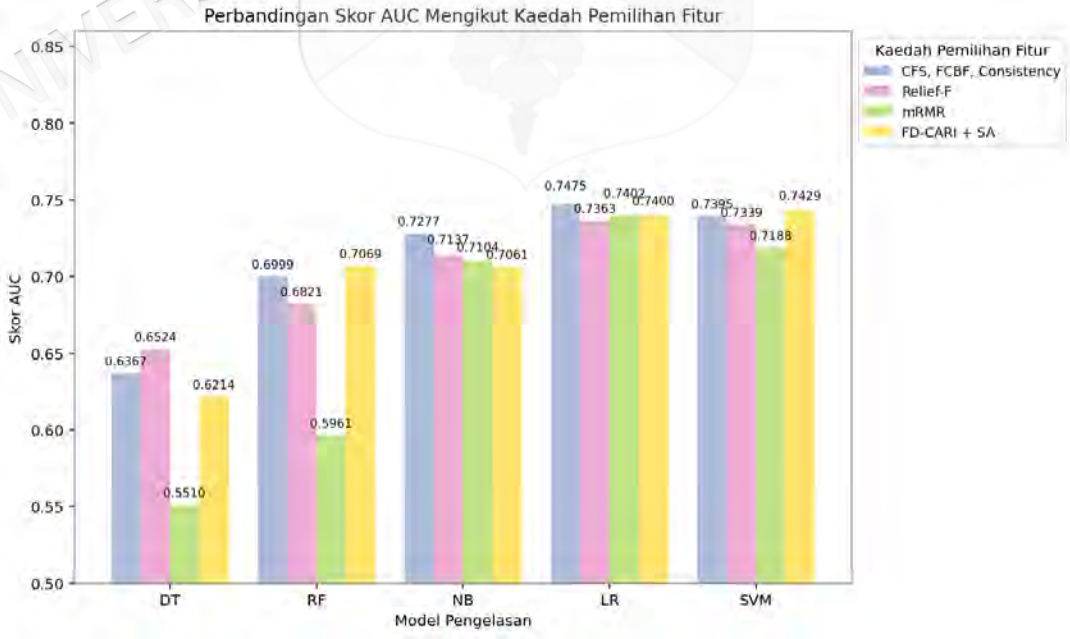
Kaedah FD-CARI+SA menunjukkan prestasi yang kompetitif berbanding kaedah pemilihan piawai bagi set data WPBC (lihat Jadual 6.17 dan Rajah 6.2). Walaupun bukan yang tertinggi secara purata (0.7035), kaedah ini mencatatkan skor AUC terbaik bagi model RF (0.7069 ± 0.0890) dan SVM (0.7429 ± 0.0568) yang membuktikan keupayaannya dalam menyokong model yang kompleks. Sebaliknya, kaedah CFS, FCBF dan Consistency mencatatkan purata skor AUC tertinggi (0.7103) dan

menunjukkan prestasi terbaik bagi model NB dan LR. Kaedah Relief-F pula menyerlah dalam model DT dengan skor AUC 0.6524 ± 0.0793.

Walaupun FD-CARI+SA menggunakan subset fitur yang lebih besar, ia tetap memberikan peningkatan yang signifikan kepada model-model tertentu. Ini menandakan kelebihan strategi interaksi fitur dalam set data yang mempunyai bilangan fitur yang lebih besar seperti WPBC.

Jadual 6.17 Skor AUC setiap kaedah pemilihan fitur bagi set data WPBC

Kaedah Pemilihan Fitur	Bilangan Fitur Terpilih	Nilai Skor AUC ± SD					Purata AUC
		DT	RF	NB	LR	SVM	
CFS	2	0.6367 ± 0.0721	0.6999 ± 0.0749	**0.7277 ± 0.0389**	**0.7475 ± 0.0465**	0.7395 ± 0.0520	**0.7103**
FCBF	2	0.6367 ± 0.0721	0.6999 ± 0.0749	**0.7277 ± 0.0389**	**0.7475 ± 0.0465**	0.7395 ± 0.0520	**0.7103**
Consistency	2	0.6367 ± 0.0721	0.6999 ± 0.0749	**0.7277 ± 0.0389**	**0.7475 ± 0.0465**	0.7395 ± 0.0520	**0.7103**
Relief-F	4	**0.6524 ± 0.0793**	0.6821 ± 0.0442	0.7137 ± 0.0404	0.7363 ± 0.0653	0.7339 ± 0.0656	0.7037
mRMR	4	0.5510 ± 0.0669	0.5961 ± 0.0806	0.7104 ± 0.0545	0.7402 ± 0.0491	0.7188 ± 0.0496	0.6633
FD-CARI + SA	4	0.6214 ± 0.0872	**0.7069 ± 0.0890**	0.7061 ± 0.0524	0.7400 ± 0.0552	**0.7429 ± 0.0568**	0.7035



Rajah 6.2 Perbandingan skor AUC bagi set data WPBC

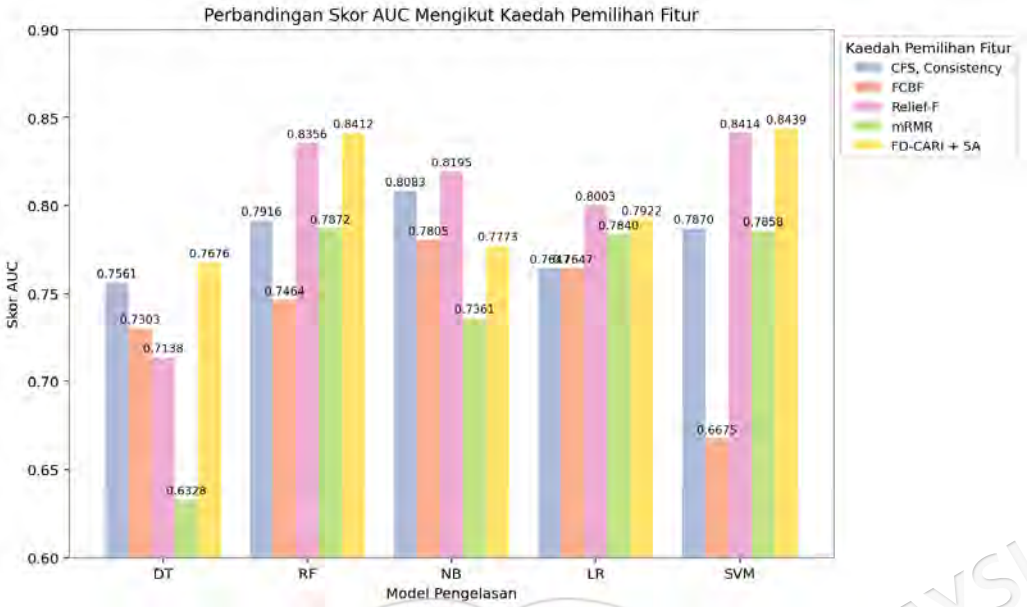
6.6.3 Perbandingan Prestasi FD-CARI+SA dengan Kaedah Pemilihan Fitur Piawai bagi Set Data BCC

Bagi set data BCC, kaedah FD-CARI+SA mencatatkan purata skor AUC tertinggi iaitu 0.8044 yang mengatasi semua kaedah pemilihan fitur piawai yang dibandingkan (rujuk Jadual 6.18 dan Rajah 6.3). Ia menunjukkan prestasi terbaik dalam model DT, RF dan SVM dengan skor AUC masing-masing sebanyak 0.7676 ± 0.0712, 0.8412 ± 0.0602 dan 0.8439 ± 0.0717. Kaedah Relief-F hampir setanding dengan skor AUC tertinggi sebanyak 0.8195 ± 0.0578 dan 0.8003 ± 0.0765 pada model NB dan LR masing-masing. Sementara itu, CFS dan Consistency merekodkan skor sederhana (0.7816), manakala FCBF dan mRMR mencatatkan prestasi paling rendah masing-masing 0.7452 dan 0.7379.

Walaupun FD-CARI+SA menggunakan lebih banyak fitur berbanding kaedah lain seperti CFS dan Consistency, keputusan menunjukkan kaedah ini berjaya meningkatkan prestasi model tertentu dan menegaskan kekuatan kaedah pemilihan fitur berasaskan interaksi dalam set data dengan bilangan fitur yang kecil seperti BCC.

Jadual 6.18 Skor AUC setiap kaedah pemilihan fitur bagi set data BCC

Kaedah Pemilihan Fitur	Bilangan Fitur Terpilih	Nilai Skor AUC ± SD					Purata AUC
		DT	RF	NB	LR	SVM	
CFS	3	0.7561 ± 0.0897	0.7916 ± 0.0787	0.8083 ± 0.0745	0.7647 ± 0.0765	0.7870 ± 0.1566	0.7816
FCBF	2	0.7303 ± 0.0788	0.7464 ± 0.0923	0.7805 ± 0.0630	0.7647 ± 0.0765	0.6675 ± 0.1607	0.7379
Consistency	3	0.7561 ± 0.0897	0.7916 ± 0.0787	0.8083 ± 0.0745	0.7647 ± 0.0765	0.7870 ± 0.1566	0.7816
Relief-F	4	0.7138 ± 0.0792	0.8356 ± 0.0667	**0.8195 ± 0.0578**	**0.8003 ± 0.0765**	0.8414 ± 0.0562	0.8021
mRMR	4	0.6328 ± 0.0815	0.7872 ± 0.0821	0.7361 ± 0.0647	0.7840 ± 0.0816	0.7858 ± 0.0811	0.7452
FD-CARI + SA	4	**0.7676 ± 0.0712**	**0.8412 ± 0.0602**	0.7773 ± 0.0861	0.7922 ± 0.0894	**0.8439 ± 0.0717**	**0.8044**



Rajah 6.3 Perbandingan skor AUC bagi set data BCC

6.7 UJIAN SIGNIFIKAN

Bagi menyokong dapatan prestasi yang diperoleh, ujian signifikan telah dijalankan ke atas skor AUC bagi setiap model pengelasan antara kaedah FD-CARI+SA dan kaedah pemilihan fitur piawai. Ujian ini bertujuan untuk menentukan sama ada perbezaan prestasi yang dicatatkan adalah signifikan secara statistik dan bukan berpunca daripada variasi rawak. Hipotesis nul (H_0) menyatakan bahawa tiada perbezaan signifikan antara FD-CARI+SA dan kaedah pemilihan fitur terbaik, manakala hipotesis alternatif (H_1) menyatakan bahawa terdapat perbezaan signifikan antara kedua-duanya. Lajur "Signifikan" dalam jadual menunjukkan sama ada perbezaan itu signifikan (+) atau tidak (-).

6.7.1 Keputusan Ujian Signifikan bagi Set Data WDBC

Jadual 6.19 membandingkan skor AUC kaedah FD-CARI+SA dan kaedah terbaik bagi setiap model pengelasan. Hasilnya menunjukkan bahawa FD-CARI+SA mencatatkan skor AUC tertinggi untuk model DT, RF dan NB manakala kaedah Consistency mencatatkan prestasi lebih baik bagi model LR dan SVM. Oleh itu, analisis statistik seterusnya difokuskan kepada dua model tersebut (LR dan SVM) bagi menentukan sama ada perbezaan yang diperoleh adalah signifikan.

Jadual 6.19 Skor AUC bagi setiap model pengelasan bagi set data WDBC

Model	Kaedah FD-CARI+SA Nilai AUC ± SD	Kaedah Terbaik	Nilai AUC ± SD
DT	0.9681 ± 0.0132	FD-CARI + SA	0.9681 ± 0.0132
RF	0.9920 ± 0.0048	FD-CARI + SA	0.9920 ± 0.0048
NB	0.9916 ± 0.0054	FD-CARI + SA dan FCBF	0.9916 ± 0.0054 / 0.9916 ± 0.0039
LR	0.9870 ± 0.0067	Consistency	0.9944 ± 0.0031
SVM	0.9913 ± 0.0048	Consistency	0.9946 ± 0.0027

Jadual 6.20 menunjukkan keputusan ujian *t*-berpasangan antara kaedah FD-CARI+SA dan kaedah pemilihan fitur terbaik bagi model LR dan SVM, iaitu Consistency. Keputusan analisis menunjukkan terdapat perbezaan yang signifikan secara statistik antara kedua-dua kaedah dengan nilai p = 0.0269 untuk model LR dan p= 0.0088 untuk model SVM. Ini menunjukan bahawa prestasi FD-CARI+SA adalah lebih rendah secara signifikan berbanding kaedah Consistency bagi kedua-dua model tersebut.

Bagi mendapatkan gambaran yang lebih menyeluruh, perbandingan tambahan dilakukan antara FD-CARI+SA dengan kaedah pemilihan fitur lain. Berdasarkan Jadual 6.21, didapati bahawa FD-CARI+SA menunjukkan prestasi yang lebih baik berbanding Relief-F dan mRMR dalam kedua-dua model LR dan SVM. Sementara itu, prestasinya adalah setanding dengan CFS dan FCBF memandangkan tiada perbezaan signifikan dikesan antara kaedah-kaedah ini.

Secara keseluruhan, meskipun FD-CARI+SA tidak mengatasi Consistency dalam model LR dan SVM, ia masih menunjukkan keupayaan kompetitif apabila dibandingkan dengan kebanyakan kaedah pemilihan fitur lain dengan kelebihan menghasilkan subset fitur yang lebih kecil yang meningkatkan keberkesanan dari segi kos pengiraan dan mampu memberikan interpretasi terhadap interaksi antara fitur.

Jadual 6.20 Hasil ujian signifikan pada kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan Prestasi
LR	0.0269	(+)	FD-CARI+SA < Consistency
SVM	0.0088	(+)	FD-CARI +SA < Consistency

Jadual 6.21 Hasil ujian signifikan pada selain kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan Prestasi
LR	0.9520	(-)	FD-CARI+SA = CFS
	0.8120	(-)	FD-CARI+SA = FCBF
	0.0002	(+)	FD-CARI+SA > Relief-F
	0.0000	(+)	FD-CARI+SA > mRMR
SVM	0.0516	(-)	FD-CARI+SA = CFS
	0.0985	(-)	FD-CARI+SA =FCBF
	0.0000	(+)	FD-CARI+SA > Relief-F
	0.0003	(+)	FD-CARI+SA > mRMR

6.7.2 Keputusan Ujian Signifikan bagi Set Data WPBC

Dapatan menunjukkan bahawa FD-CARI+SA mencatatkan prestasi terbaik bagi model RF dan SVM manakala model DT menunjukkan prestasi tertinggi dengan kaedah Relief-F. Sementara itu, bagi model NB dan LR, kaedah CFS, FCBF, Consistency menunjukkan skor AUC yang lebih tinggi berbanding FD-CARI+SA. Seperti yang ditunjukkan dalam Jadual 6.22, ketiga-tiga model ini mencatatkan tiada perbezaan signifikan dengan nilai *p* masing-masing sebanyak 0.4008, 0.4444 dan 0.7529. Ini membuktikan bahawa prestasi FD-CARI+SA adalah setanding dengan kaedah pemilihan fitur terbaik dalam ketiga-tiga model, sekali gus mengukuhkan keberkesanan kaedah ini dalam konteks set data WPBC.

Jadual 6.22 Skor AUC bagi setiap model pengelasan bagi set data WPBC

Model	Kaedah FD-CARI Nilai AUC ± SD	Kaedah Terbaik	Nilai AUC ± SD
DT	0.6214 ± 0.0872	Relief-F	0.6524 ± 0.0793
RF	0.7069 ± 0.0890	FD-CARI+SA	0.7069 ± 0.0890
NB	0.7061 ± 0.0524	CFS, FCBF, Consistency	0.7277 ± 0.0389
LR	0.7400 ± 0.0552	CFS, FCBF, Consistency	0.7475 ± 0.0465
SVM	0.7429 ± 0.0568	FD-CARI+SA	0.7395 ± 0.0520

Jadual 6.23 Hasil ujian signifikan pada kaedah pemilihan fitur terbaik

Model	Nilai *p*	Signifikan	Penjelasan
DT	0.4008	(-)	FD-CARI+SA = Relief-F
NB	0.4444	(-)	FD-CARI+SA = CFS, FCBF, Consistency
LR	0.7529	(-)	FD-CARI+SA = CFS, FCBF, Consistency

6.7.3 Keputusan Ujian Signifikan bagi Set Data BCC

Memandangkan keputusan yang diperoleh daripada FD-CARI+SA dan penalaan manual FD-CARI adalah konsisten pada set data BCC, tiada ujian signifikan dijalankan bagi set data ini. Prestasi yang dicatatkan adalah sama membawa kepada hasil ujian signifikan yang sama seperti yang telah dibincangkan dalam Seksyen 4.6.3.

6.8 PERBINCANGAN

Bab ini membentangkan dapatan bagi objektif ketiga iaitu menilai keberkesanan strategi pengoptimuman FD-CARI menggunakan algoritma SA (FD-CARI+SA) dalam menghasilkan subset fitur yang lebih padat dan meningkatkan prestasi pengelasan. Hasil kajian menunjukkan bahawa pendekatan ini telah berjaya mengenal pasti kombinasi hiperparameter yang menghasilkan subset fitur optimum serta skor AUC yang kompetitif atau lebih baik berbanding kaedah FD-CARI asal dan kaedah pemilihan fitur piawai.

Berdasarkan kajian yang telah dijalankan ke atas tiga set data, FD-CARI+SA berjaya memilih subset fitur yang lebih kecil, namun masih mengekalkan atau meningkatkan prestasi pengelasan. Sebagai contoh, bagi set data WDBC, FD-CARI+SA memilih hanya 3 fitur iaitu *concave points_worst, radius_worst* dan *texture_worst* berbanding 4 fitur dalam FD-CARI, tetapi mencatatkan skor AUC purata 0.9860 sedikit lebih tinggi. Bagi set data WPBC pula, FD-CARI+SA mencatatkan skor 0.7035 yang setanding atau lebih tinggi berbanding kebanyakan kaedah pemilihan fitur lain. Sementara itu, bagi set data BCC, FD-CARI+SA mencatatkan skor AUC tertinggi iaitu 0.8044 dan menjadikannya kaedah terbaik bagi set data tersebut.

Perbandingan prestasi dengan kaedah piawai seperti FCBF, CFS, Relief-F dan mRMR menunjukkan bahawa FD-CARI+SA tidak semestinya mengatasi semua kaedah dalam setiap model pengelasan. Walau bagaimanapun, ujian signifikan menunjukkan bahawa prestasi FD-CARI+SA adalah setanding dengan kaedah terbaik dalam kebanyakan kes. Malah, dalam sesetengah model seperti DT, RF dan SVM, FD-CARI+SA mencatatkan prestasi tertinggi secara konsisten. Dapatan ini menunjukkan bahawa kaedah FD-CARI+SA bukan sahaja mampu memilih subset fitur yang padat,

tetapi juga mempertimbangkan elemen interaksi antara fitur dan penjanaan hiperparameter pada peraturan secara menyeluruh. Keupayaan algoritma SA dalam meneroka ruang carian secara global membolehkan pemilihan kombinasi hiperparameter yang sesuai untuk pelbagai jenis data.

Secara keseluruhan, algoritma SA memberikan keseimbangan yang efektif antara proses eksplorasi dan eksploitasi semasa penalaan ambang ARM, sekali gus meningkatkan kecekapan proses pemilihan fitur dalam FD-CARI+SA. Keupayaan SA meneroka ruang carian secara meluas pada peringkat awal, tetapi secara adaptif mengecilkan carian ke kawasan yang lebih baik pada peringkat seterusnya yang membolehkan ia mengelakkan perangkap minimum setempat iaitu suatu isu lazim dalam penalaan ambang ARM. Gabungan kaedah ini turut membuktikan bahawa kerangka ini relevan untuk domain kesihatan yang memerlukan interpretasi peraturan yang jelas serta proses pengekstrakan pengetahuan yang stabil dan boleh dipercayai daripada data perubatan. Kaedah ini bukan sahaja mengekalkan kebolehjelasan model, tetapi juga memastikan hiperparameter ditetapkan secara automatik dengan cara yang konsisten merentas set data yang berbeza.

6.9 KESIMPULAN

Bab ini membentangkan penilaian keberkesanan strategi pengoptimuman FD-CARI menggunakan algoritma SA selaras dengan objektif ketiga kajian. Melalui kaedah FD-CARI+SA, proses pemilihan fitur telah ditambah baik dengan mencari kombinasi hiperparameter yang menghasilkan subset fitur lebih padat tanpa menjejaskan prestasi model pengelasan. Hasil kajian ke atas tiga set data menunjukkan bahawa kaedah ini berjaya mencapai skor AUC yang kompetitif dan dalam beberapa kes, ia lebih tinggi berbanding FD-CARI asal serta kaedah pemilihan fitur piawai lain seperti CFS, FCBF, Consistency, Relief-F dan mRMR. Ujian signifikan turut menunjukkan bahawa FD-CARI+SA memiliki prestasi yang setanding atau lebih baik berbanding kaedah-kaedah sedia ada terutamanya dalam model-model utama seperti RF dan SVM. Ini membuktikan kebolehlaksanaan dan kelebihan kaedah FD-CARI+SA ini dalam mengenal pasti subset fitur yang relevan, berinteraksi dan tidak bertindih walaupun dalam persekitaran data yang pelbagai.

BAB VII

KESIMPULAN DAN CADANGAN

7.1 RUMUSAN KAJIAN

Bab ini merumuskan keseluruhan kajian dengan menekankan pencapaian utama bagi setiap objektif serta membincangkan sumbangan kajian kepada teori, metodologi dan aplikasi dalam bidang pengelasan dan diagnosis kanser payudara. Kajian ini telah memperkenalkan dan menilai keberkesanan kaedah FD-CARI (*Fuzzy Discretization and Class Association Rule Mining with Interaction*) dalam pemilihan fitur berasaskan interaksi bagi pengelasan kanser payudara. Objektif kajian telah dicapai melalui tiga fasa utama yang masing-masing ditumpukan dalam Bab 4, Bab 5 dan Bab 6. Secara keseluruhan, dapatan kajian ini menyumbang secara signifikan kepada pengukuhan metodologi pemilihan fitur yang mempertimbangkan interaksi yang bukan hanya kebergantungan individu terhadap kelas sasaran.

Pertama, kajian ini telah berjaya membuktikan bahawa kaedah FD-CARI menghasilkan subset fitur yang lebih padat dan relevan berbanding kaedah piawai. Sebagai contoh, dalam set data WDBC, FD-CARI hanya memilih 4 daripada 30 fitur dengan purata skor AUC sebanyak 0.9857 yang setanding dengan model yang menggunakan bilangan fitur lebih besar. Pendekatan modular dalam Bab 4 juga menunjukkan penarafan fitur berdasarkan interaksi awal berjaya menyingkirkan fitur tidak relevan berdasarkan tiga kaedah penaraf tanpa menjejaskan prestasi pengelasan.

Kedua, penilaian menerusi teknik XAI iaitu model SHAP dalam Bab 5 mengesahkan bahawa fitur kurang relevan yang dipilih oleh FD-CARI menunjukkan interaksi signifikan dengan fitur dominan. Sebagai contoh, dalam set data WDBC, fitur seperti *radius_worst* dan *perimeter_worst* membentuk 13 dan 10 interaksi masing-masing dalam kuantil ke-75 interaksi SHAP. Analisis ini memperkukuhkan justifikasi

bahawa fitur yang kelihatan lemah secara individu sebenarnya memainkan peranan penting dalam interaksi model.

Ketiga, integrasi algoritma SA seperti dibentangkan dalam Bab 6 menunjukkan peningkatan prestasi melalui pengoptimuman hiperparameter. Bagi set data WDBC, skor AUC meningkat daripada 0.9857 kepada 0.9860, manakala jumlah fitur berjaya dikurangkan secara signifikan iaitu daripada 4 kepada 3 fitur sahaja daripada jumlah 30 fitur. Ujian signifikan juga menunjukkan prestasi kaedah FD-CARI+SA setanding atau lebih baik daripada kaedah piawai dalam kebanyakan model pengelasan.

Dapatan-dapatan ini menunjukkan sumbangan kajian ini dalam menghasilkan satu kerangka pemilihan fitur berasaskan interaksi yang konsisten berasaskan bukti yang telah dioptimumkan. Berbeza dengan pendekatan terdahulu yang lebih tertumpu kepada pemilihan fitur individu, FD-CARI memberi tumpuan kepada kebergantungan fitur-fitur dalam bentuk hubungan kelas, lalu menyumbang kepada pembangunan model pengelasan yang lebih telus dan boleh dijelaskan.

7.2 PERBINCANGAN DAN PERBANDINGAN KRITIS

Seksyen ini membincangkan dapatan kajian secara kritikal dengan membuat perbandingan terhadap pendekatan sedia ada dari aspek pemilihan fitur, pengesahan interaksi dan prestasi model pengelasan. Perbincangan ini membuktikan bahawa pendekatan yang dicadangkan bukan sahaja memenuhi objektif kajian, malah memberikan nilai tambah yang signifikan kepada bidang perubatan.

7.2.1 Keunikan FD-CARI Berbanding Teknik Sedia Ada

Model FD-CARI menggabungkan pendiskretan kabur dan kaedah CARM untuk mengenal pasti fitur yang bukan sahaja relevan, tetapi juga berinteraksi dan tidak bertindih. Ini berbeza dengan kaedah seperti:

- CFS memilih subset berdasarkan korelasi namun tidak mengesan interaksi fitur.
- FCBF menggunakan *Symmetrical Uncertainty* (SU) untuk menilai hubungan dengan kelas dan tidak menilai interaksi fitur.

- Consistency mempertimbangkan kestabilan subset dengan meminimumkan ketidakselarasan (*inconsistency*) dalam pengelasan namun tidak menilai interaksi fitur.
- Relief-F menggunakan konsep jiran terdekat untuk mengukur kepentingan fitur namun hanya menumpu pada 2 hala interaksi fitur.
- mRMR memilih fitur yang relevan dengan kelas sambil meminimumkan pertindihan namun tidak menilai interaksi antara fitur.

Dapatan kajian menunjukkan bahawa FD-CARI mencapai purata skor AUC 0.9860 untuk WDBC dengan hanya 4 fitur, manakala mRMR dan Relief-F gagal mengekalkan prestasi apabila jumlah fitur dikurangkan ke tahap yang sama.

7.2.2 Perbandingan Pengesahan Interaksi Fitur

Kajian ini menggunakan model SHAP untuk mengesahkan bahawa fitur kurang relevan yang dipilih oleh FD-CARI sebenarnya berinteraksi secara signifikan dengan fitur utama. Ini menjawab satu kekangan utama dalam banyak kaedah pemilihan fitur terdahulu yang mengabaikan fitur 'lemah' hanya kerana prestasi individu yang rendah. Antara contoh kritikal ialah fitur *area_se* dan *compactness_se* dalam WDBC tidak muncul sebagai penting secara tunggal, tetapi mempunyai nilai SHAP interaksi tinggi dengan *radius_worst* dan *texture_worst* menjelaskan sumbangan tidak langsung mereka. Tiada kajian terdahulu yang menggabungkan pendiskretan kabur berasaskan CARM yang mempertimbangkan interaksi dengan pengesahan berasaskan SHAP serta pengundian majoriti berwajaran menjadikan kaedah ini baharu dan boleh dijelaskan.

7.2.3 Pengoptimuman FD-CARI Menggunakan Algoritma SA

Walaupun FD-CARI terbukti berkesan, pelarasan hiperparameter secara manual mungkin tidak menghasilkan konfigurasi terbaik. Justeru, SA digunakan untuk mengoptimumkan gabungan nilai *minSupp*, *minConf, minLift*, *minConv, minCF dan minLev*. Bukti kuantitatif dapat dilihat di mana FD-CARI+SA mencatatkan peningkatan skor AUC daripada 0.9857 (FD-CARI) kepada 0.9860 (FD-CARI+SA) untuk set data WDBC dan lebih ketara untuk WPBC iaitu daripada 0.7015 kepada 0.7035. Tambahan

pula, FD-CARI+SA berjaya mengurangkan bilangan fitur terpilih dan menunjukkan kecekapan tanpa menjejaskan ketepatan.

7.2.4 Perbandingan dengan Kajian Kesusasteraan

Jadual 7.1 menunjukkan perbandingan antara kajian ini dengan beberapa kajian utama kesusasteraan.

Jadual 7.1 Perbandingan model FD-CARI dengan kajian kesusasteraan

Kajian/Model	Interaksi Fitur	Boleh Dijelaskan	Pengoptimuman Hiperparameter	Saiz Subset Fitur	Prestasi AUC		
					WDBC	WPBC	BCC
CFS	×	×	×	2-11	0.9813	0.7103	0.78156
FCBF	×	×	×	2-7	0.9866	0.7103	0.7379
Consistency	×	×	×	3-8	0.9855	0.7103	0.7816
Relief-F	√	×	×	3-4	0.8708	0.7037	0.8021
mRMR	×	×	×	3-4	0.9293	0.6633	0.7452
FD-CARI+SA	√	√	√	3-4	0.9860	0.7035	0.8044

Jadual 7.1 menunjukkan kajian ini menyumbang kepada perkembangan kaedah pemilihan fitur dengan memperkenalkan teknik berasakan interaksi, boleh dijelaskan dan dioptimumkan secara algoritma sekali gus menjawab kelemahan dalam kajian terdahulu.

7.3 IMPLIKASI KAJIAN

Kajian ini memberi implikasi penting dalam tiga dimensi utama, teori, metodologi dan praktikal yang bersama-sama menyumbang kepada pengembangan pengetahuan dalam bidang pemilihan fitur, XAI dan pengelasan diagnosis perubatan.

7.3.1 Implikasi terhadap Teori

Kajian ini mengembangkan pemahaman bahawa fitur yang dilihat sebagai kurang relevan secara individu boleh menyumbang secara signifikan apabila digabungkan melalui interaksi yang kuat. Ini memperluaskan konsep tradisional dalam pemilihan fitur yang hanya menekankan kekuatan individu fitur. Melalui pengesahan nilai interaksi SHAP, kajian ini membuktikan secara kuantitatif bahawa interaksi antara fitur harus diambil kira sebagai sebahagian daripada teori pemilihan fitur moden. Kajian ini

menyatukan teori pemilihan fitur dengan model XAI dan menunjukkan potensi teori baharu yang menggabungkan kedua-dua bidang ini.

7.3.2 Implikasi terhadap Metodologi

Kajian ini memperkenalkan FD-CARI sebagai satu model berasaskan CARM yang disesuaikan secara khusus untuk mempertimbangkan interaksi dan set fitur yang tidak bertindih serta fitur yang kurang diberi perhatian dalam kebanyakan teknik penapis sedia ada. Penggunaan algoritma SA pula dalam melaraskan hiperparameter FD-CARI menunjukkan gabungan inovatif antara pendekatan metaheuristik dan pemilihan fitur berasaskan peraturan yang boleh membuka ruang untuk penyelidikan lanjutan dalam bidang pemilihan fitur berpandukan pengoptimuman. Kajian ini membuktikan juga bahawa model SHAP bukan sahaja berguna untuk interpretasi model, tetapi juga berkesan dalam mengesahkan interaksi fitur terpilih dan menjadikan ia penting dalam membentuk kaedah pemilihan fitur masa depan.

7.3.3 Implikasi terhadap Amalan Praktikal

Kajian ini memberi impak praktikal kepada penghasilan sistem pengelasan kanser payudara dengan bilangan fitur minimum, namun masih mengekalkan prestasi tinggi (AUC>0.98), justeru menjadikannya lebih mudah diterjemahkan dalam sistem sokongan keputusan klinikal. Pemilihan fitur yang berinteraksi dan boleh dijelaskan melalui nilai SHAP menjadikan model lebih telus dan dipercayai, lantas memudahkan doktor memahami hubungan antara fitur perubatan. Selain itu, dengan subset fitur yang kecil tetapi berkesan, sistem ini boleh digunakan dalam persekitaran klinikal sebenar untuk mengurangkan kos ujian dan mempercepatkan proses pengelasan tanpa kehilangan ketepatan.

7.4 SUMBANGAN KAJIAN

Kajian ini memberikan sumbangan yang signifikan dalam tiga aspek utama iaitu sumbangan teori, metodologi dan praktikal. Setiap sumbangan disokong oleh bukti kuantitatif dan dijajarkan dengan objektif kajian yang telah dicapai sepenuhnya (lihat Jadual 7.2).

Jadual 7.2 Pemetaan objektif, dapatan utama dan sumbangan kajian

Objektif Kajian	Dapatan Utama	Sumbangan Kajian
Merumuskan kaedah pemilihan fitur berasaskan kaedah CARM iaitu FD-CARI yang dapat mengenal pasti subset fitur yang relevan, berinteraksi dan tidak bertindih	FD-CARI menghasilkan subset fitur padat (4-5 fitur) dengan prestasi tinggi (AUC >0.98 untuk WDBC). Prestasi setanding atau lebih baik berbanding kaedah piawai	▪ Kaedah pemilihan fitur baru yang menggabungkan kebergantungan fitur, pendiskretan kabur dan CARM ▪ Sumbangan metodologi kepada bidang pemilihan fitur berasaskan interaksi
Mereka bentuk kerangka pengesahan fitur berinteraksi terpilih menggunakan teknik XAI berasaskan model SHAP dan kaedah pengundian majoriti berwajaran untuk meningkatkan ketelusan dan kebolehpercayaan kaedah FD-CARI	Semua fitur kurang relevan mempunyai sekurang-kurangnya satu pasangan interaksi yang signifikan berdasarkan SHAP, peta haba dan jaringan interaksi menyokong dapatan	▪ Sumbangan teoretikal dalam memantapkan pemahaman interaksi fitur tersembunyi ▪ Pelaksanaan proses pengesahan berlapis
Membangunkan kaedah hibrid FD-CARI+SA yang menggabungkan kaedah CARM dengan algoritma SA bagi penalaan hiperparameter secara automatik dan peningkatan prestasi model pengelasan kanser payudara	FD-CARI+SA menunjukkan peningkatan skor AUC dalam kebanyakan set data, dan meningkatkan prestasi dengan subset fitur yang lebih kecil (3-4 fitur)	▪ Sumbangan metodologi melalui gabungan pemilihan fitur dan SA ▪ Menyediakan laluan praktikal untuk perlaksanaan sistem ramalan yang ringan dan efisien

7.4.1 Sumbangan Teori

Kajian ini membuktikan secara empirikal bahawa fitur yang kurang relevan secara individu boleh membentuk interaksi yang kuat dan signifikan. Ini menambah nilai kepada teori pemilihan fitur dengan memperluaskan pemahaman terhadap sinergi fitur dan pemilihan fitur berasaskan interaksi. Selain itu, kajian ini juga membina jambatan antara dua domain penting iaitu pemilihan fitur dan XAI dengan menunjukkan bahawa pemilihan fitur berinteraksi boleh disahkan dan diperjelaskan melalui nilai SHAP sekali gus menyokong perkembangan teori baharu.

7.4.2 Sumbangan Metodologi

Kaedah FD-CARI yang dibangunkan mengintegrasikan pendiskretan kabur, CARM dan konsep interaksi fitur. Kaedah ini berbeza daripada kaedah pemilihan fitur sedia ada kerana ia menekankan pada pemilihan fitur yang relevan, berinteraksi dan tidak bertindih secara serentak. Kaedah ini turut menggabungkan FD-CARI dengan SA sebagai teknik pengoptimuman hiperparameter untuk menghasilkan subset fitur optimum. Tiada kajian sebelum ini menunjukkan pendekatan ini dalam konteks pengelasan kanser payudara. Selain itu, kajian ini juga memperkenalkan satu prosedur sistematik untuk mengesahkan interaksi fitur melalui nilai SHAP interaksi kuantil ke-75 serta kaedah pengundian majoriti berwajaran untuk memperhalusi kepentingan fitur. Pendekatan ini menyumbang kaedah baru kepada penilaian prestasi model yang menjurus kepada kebolehfahaman.

7.4.3 Sumbangan Praktikal

Kajian ini menghasilkan subset kecil iaitu 3 hingga 5 fitur sahaja dengan prestasi AUC melebihi 0.98 membuktikan keberkesanannya untuk sistem pengelasan klinikal yang ringan dan pantas. Selain itu, pemilihan fitur yang boleh dijelaskan juga memberi kelebihan praktikal dari segi kos ujian makmal yang lebih rendah serta mudah difahami oleh pengamal klinikal sekali gus memperkukuh penerimaan sistem dalam amalan sebenar. Walaupun kajian ini berfokuskan kepada data kanser payudara, kaedah FD-CARI+SA bersifat generik dan boleh diguna pakai dalam domain lain seperti genomik, kewangan dan kesihatan awam yang menjadikannya bersifat kebolehskalaan dan boleh diadaptasi.

7.5 KETERBATASAN DAN CADANGAN KAJIAN LANJUTAN

Walaupun kajian ini berjaya membangunkan dan mengesahkan teknik pemilihan fitur berasaskan interaksi (FD-CARI) yang berprestasi tinggi, terdapat beberapa keterbatasan yang wajar diakui sebagai panduan untuk penyelidikan lanjut. Kajian ini menggunakan tiga set data kanser payudara yang tersedia secara umum. Meskipun sesuai untuk tujuan pengesahan awal, set data ini mempunyai saiz sampel yang relatif kecil dan tidak merangkumi variasi konteks klinikal sebenar seperti data dunia sebenar di hospital atau

multi modal seperti gabungan data imej dan teks. Selain itu, hanya lima model pembelajaran mesin konvensional digunakan untuk penilaian prestasi. Walaupun mencukupi untuk penilaian awal, keterbatasan ini boleh menjejaskan generalisasi keputusan terhadap model pembelajaran mendalam. Penilaian analisis SHAP digunakan untuk interpretasi fitur juga tidak melibatkan pakar klinikal secara langsung.

Sehubungan dengan itu, terdapat beberapa cadangan kajian lanjutan seperti berikut: (i) menguji kerangka FD-CARI+SA pada set data kanser atau penyakit lain bagi menilai kebolehumuman model, (ii) meneroka algoritma pengoptimuman alternatif seperti algoritma PSO atau GA untuk perbandingan yang lebih komprehensif, dan (iii) mengintegrasikan model ini ke dalam sistem sokongan keputusan perubatan bagi menilai potensi penggunaannya dalam membantu diagnosis awal di persekitaran klinikal sebenar. Malah, sesi *Delphi* atau bengkel fokus berkumpulan (FGD) bersama pakar klinikal juga boleh dijalankan bagi mengesahkan semula kepentingan dan kesahihan fitur yang dipilih serta interpretasi SHAP dalam konteks diagnostik sebenar.

7.6 KESIMPULAN

Kajian ini telah membangunkan, mengesahkan dan menilai secara kritis satu kaedah pemilihan fitur baru iaitu FD-CARI yang memfokuskan kepada interaksi antara fitur dalam konteks diagnosis kanser payudara. Sepanjang 3 objektif kajian, kaedah ini telah terbukti berkesan dalam menghasilkan subset fitur yang padat, bermakna dan berprestasi tinggi dalam pengelasan serta meningkatkan interpretasi model melalui nilai SHAP dan analisis interaksi. Objektif pertama berjaya dicapai melalui pembentukan FD-CARI yang menggabungkan pendiskretan kabur dan kaedah CARM bagi menghasilkan subset fitur interaksi yang relevan dan tidak bertindih. Objektif kedua disahkan melalui penilaian interaksi menggunakan nilai SHAP dan pengundian majoriti berwajaran membuktikan bahawa fitur yang dianggap kurang relevan turut mempunyai nilai interaksi yang signifikan. Objektif ketiga melibatkan proses pengoptimuman menggunakan algoritma SA yang berjaya menambahbaik prestasi pengelasan dan mengurangkan bilangan fitur dengan menyesuaikan gabungan nilai hiperparameter. Secara keseluruhan, dapatan kajian ini memperkukuh pemahaman bahawa pemilihan fitur tidak boleh bersandarkan kepada kepentingan fitur individu semata-mata, sebaliknya perlu mengambil kira interaksi tersirat antara fitur untuk menghasilkan

model pengelasan yang bukan sahaja tepat, tetapi juga mudah difahami dan boleh dijelaskan kepada domain pakar seperti bidang klinikal. Sumbangan utama kajian ini terletak pada (i) aspek teori melalui pemahaman baru terhadap pengaruh interaksi fitur dalam tugas pengelasan, (ii) aspek metodologi melalui penghasilan kerangka FD-CARI yang menggabungkan pendiskretan kabur, CARM, SHAP dan SA serta (iii) aspek praktikal melalui potensi integrasi dalam sistem diagnosis klinikal yang memerlukan subset fitur minimum namun bermakna.

RUJUKAN

Aamir, S., Rahim, A., Aamir, Z., Abbasi, S.F., Khan, M.S., Alhaisoni, M., Khan, M.A., Khan, K. & Ahmad, J. 2022. Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques. *Computational and Mathematical Methods in Medicine 2022*

Abrantes, J. & Rouzrokh, P. 2024. Explaining explainability: The role of XAI in medical imaging. *European Journal of Radiology 173*: 111389. https://www.sciencedirect.com/science/article/pii/S0720048X24001050.

Adadi, A. & Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access 6*: 52138–52160.

Adam, N. & Wieder, R. 2024. Temporal Association Rule Mining: Race-Based Patterns of Treatment-Adverse Events in Breast Cancer Patients Using SEER–Medicare Dataset. *Biomedicines 12*(6)

Agrawal, R., Imieliński, T. & Swami, A. 1993. Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Record 22*(2): 207–216.

Ahmed, H.K., Abdullah, D.A. & Al-Tuwaijari, J.M. 2019. Breast Cancer Classification: Using Hybrid Technique (Association Rules and Deep Neural Network). *Journal of Physics: Conference Series 1294*(4)

Ahn, J.S., Shin, S., Yang, S.-A., Park, E.K., Kim, K.H., Cho, S.I., Ock, C.-Y. & Kim, S. 2023. Artificial intelligence in breast cancer diagnosis and personalized medicine. *Journal of breast cancer 26*(5): 405–435.

Ahsan, M.M., Mahmud, M.A.P., Saha, P.K., Gupta, K.D. & Siddique, Z. 2021. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies 9*(3): 5–9.

Akbas, K.E., Kivrak, M., Arslan, A.K., Yakinbas, T., Korkmaz, H., Önalan, E. & Çolak, C. 2022. Assessment of Association Rule Mining Using Interest Measures on the Gene Data. *Medical Records 4*(3): 286–292.

Al Nuaimi, N., Masud, M.M., Serhani, M.A. & Zaki, N. 2019. Streaming Feature Selection Algorithms for Big Data: A Survey. *Applied Computing and Informatics*

Al Rahman, E.A., Ruhaiyem, N.I.R., Bouchahma, M. & Musa, K.I. 2023. Framework for a Computer-Aided Treatment Prediction (CATP) System for Breast Cancer. *Intelligent Automation and Soft Computing 36*(3): 3007–3028.

Alam, T.M., Shaukat, K., Hameed, I.A., Khan, W.A., Sarwar, M.U., Iqbal, F. & Luo, S. 2021. A Novel Framework for Prognostic Factors Identification of Malignant Mesothelioma Through Association Rule Mining. *Biomedical Signal Processing and Control 68*(May 2021): 102726. https://doi.org/10.1016/j.bspc.2021.102726.

Albahri, A.S., Duhaim, A.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L., Albahri, O.S., Alamoodi, A.H., Bai, J., Salhi, A., Santamaría, J., Ouyang, C., Gupta, A., Gu, Y. & Deveci, M. 2023. A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion. *Information Fusion 96*: 156–191. https://www.sciencedirect.com/science/article/pii/S1566253523000891.

Al-Fahaidy, F.A.K., Al-Fuhaidi, B., Al-Darouby, I., Al-Abady, F., Al-Qadry, M. & Al-Gamal, A. 2022. A Diagnostic Model of Breast Cancer Based on Digital Mammogram Images Using Machine Learning Techniques. *Applied Computational Intelligence and Soft Computing 2022*

Alharith, R., Khalil, M., Ibrahim, A.O. & Babiker, S.H. 2024. Extraction of Association Rules from Cancer Patient's Records using F-P Growth Algorithm. *ITM Web of Conferences 63*: 01017.

Alhussan, A.A., Eid, M.M., Towfek, S.K. & Khafaga, D.S. 2023. Breast Cancer Classification Depends on the Dynamic Dipper Throated Optimization Algorithm. *Biomimetics 8*(2): 163.

Alibrahim, H. & Ludwig, S.A. 2021. Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. *2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings* 1551–1559.

Aljehani, S.S. & Alotaibi, Y.A. 2024. Preserving privacy in association rule mining using metaheuristic-based algorithms: A systematic literature review. *IEEE Access 12*(December 2023): 21217–21236.

Alsabry, A., Algabri, M., Ahsan, A.M., Mosleh, M.A.A., Ahmed, A.A. & Qasem, H.A. 2023. Enhancing prediction models' performance for breast cancer using SMOTE technique. *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)* hlm. 1–8

Al-Thelaya, K., Gilal, N.U., Alzubaidi, M., Majeed, F., Agus, M., Schneider, J. & Househ, M. 2023. Applications of Discriminative and Deep Learning Feature Extraction Methods for Whole Slide Image Analysis: A Survey. *Journal of Pathology Informatics 14*(July): 100335. https://doi.org/10.1016/j.jpi.2023.100335.

Alwidian, J., Hammo, B.H. & Obeid, N. 2018. WCBA: Weighted Classification Based on Association Rules Algorithm for Breast Cancer Disease. *Applied Soft Computing Journal 62*: 536–549. http://dx.doi.org/10.1016/j.asoc.2017.11.013.

Amann, J., Blasimme, A., Vayena, E., Frey, D. & Madai, V.I. 2020. Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Medical Informatics and Decision Making 20*(1): 1–9. https://doi.org/10.1186/s12911-020-01332-6.

Amin, M.N., Kamal, R., Farouk, A., Gomaa, M., Rushdi, M.A. & Mahmoud, A.M. 2023. An efficient hybrid computer-aided breast cancer diagnosis system with wavelet packet transform and synthetically-generated contrast-enhanced

spectral mammography images. *Biomedical Signal Processing and Control 85*(March): 104808. https://doi.org/10.1016/j.bspc.2023.104808.

Amoroso, N., Pomarico, D., Fanizzi, A., Didonna, V., Giotta, F., La Forgia, D., Latorre, A., Monaco, A., Pantaleo, E., Petruzzellis, N., Tamborra, P., Zito, A., Lorusso, V., Bellotti, R. & Massafra, R. 2021. A Roadmap Towards Breast Cancer Therapies Supported by Explainable Artificial Intelligence. *Applied Sciences (Switzerland) 11*(11): 1–17.

Ansyari, M.R., Mazdadi, M.I., Indriani, F., Kartini, D. & Saragih, T.H. 2023. Implementation of Random Forest and Extreme Gradient Boosting in the Classification of Heart Disease using Particle Swarm Optimization Feature Selection. *Journal of Electronics, Electromedical Engineering, and Medical Informatics 5*(4): 250–260.

Arslan, A.K., Tunc, Z., Cicek, I.B. & Colak, C. 2020. A Novel Interpretable Web-Based Tool on the Associative Classification Methods : An Application on Breast Cancer Dataset *5*(1): 33–40.

Avcı, H. & Karakaya, J. 2023. A Novel Medical Image Enhancement Algorithm for Breast Cancer Detection on Mammography Images Using Machine Learning. *Diagnostics (Basel, Switzerland) 13*(3)

Ayoola, J. & Ogunfunmi, T. 2022. A Comparative Analysis of Regression Algorithms with Genetic Algorithm in the Prediction of Breast Cancer Tumors. *2022 IEEE Global Humanitarian Technology Conference, GHTC 2022* 143–149.

Azizah, A., Hashimah, B., Nirmal, K., Siti Zubaidah, A., Puteri, N., Nabihah, A., Sukumaran, R., Balqis, B., Nadia, S. & Sharifah, S. 2019. Malaysia National cancer registry report (MNCR). *National Cancer Institute, Ministry of Health: Putrajaya, Malaysia*

Bai, S., Nasir, S., Khan, R.A., Arif, S., Meyer, A. & Konik, H. 2024. Breast Cancer Diagnosis: A Comprehensive Exploration of Explainable Artificial Intelligence (XAI) Techniques http://arxiv.org/abs/2406.00532.

Barrios, C.H. 2022. Global challenges in breast cancer detection and treatment. *Breast 62*(S1): S3–S6. https://doi.org/10.1016/j.breast.2022.02.003.

Bartschat, A., Reischl, M. & Mikut, R. 2019. Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9*(4): 1–14.

Battiti, R. 1994. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural Networks 5*(4): 537–550.

Beausoleil, R.P. 2020. Associative Classification With Multiobjective Tabu Search. *Revista De Matemática: Teoría Y Aplicaciones 27*(2): 333–354.

Benesty, J., Chen, J., Huang, Y. & Cohen, I. 2009. Pearson Correlation Coefficient. *Noise Reduction in Speech Processing* hlm. 1–4. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5.

Bennasar, M., Setchi, R. & Hicks, Y. 2013. Feature Interaction Maximisation. *Pattern Recognition Letters 34*: 1630–1635.

Binder, M., Moosbauer, J. & Thomas, J. 2020. Multi-Objective Hyperparameter Tuning and Feature Selection using Filter Ensembles 471–479.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.L., Deng, D. & Lindauer, M. 2023. Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13*(2)

Boumaraf, S., Liu, X., Ferkous, C. & Ma, X. 2020. A New Computer-Aided Diagnosis System with Modified Genetic Feature Selection for BI-RADS Classification of Breast Masses in Mammograms. *BioMed Research International 2020*

Breiman, L. 2001. Random Forests. *Machine Learning 45*: 5–32.

Carvalho, R.H.D.O., Martins, A.S., Neves, L.A. & Do Nascimento, M.Z. 2020. Analysis of Features for Breast Cancer Recognition in Different Magnifications of Histopathological Images. *International Conference on Systems, Signals, and Image Processing 2020-July*: 39–44.

Castronuovo, G., Favia, G., Telesca, V. & Vammacigno, A. 2023. Analyzing the Interactions between Environmental Parameters and Cardiovascular Diseases Using Random Forest and SHAP Algorithms. *Reviews in Cardiovascular Medicine 24*(11)

Çetinel, G., Mutlu, F. & Gül, S. 2020. Decision support system for breast lesions via dynamic contrast enhanced magnetic resonance imaging. *Physical and Engineering Sciences in Medicine 43*(3): 1029–1048. https://doi.org/10.1007/s13246-020-00902-2.

Chakraborty, D., Ivan, C., Amero, P., Khan, M., Rodriguez-Aguayo, C., Başağaoğlu, H. & Lopez-Berestein, G. 2021. Explainable Artificial Intelligence Reveals Novel Insight Into Tumor Microenvironment Conditions Linked with Better Prognosis in Patients with Breast Cancer. *Cancers 13*(14)

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T. & Tsaneva-Atanasova, K. 2019. Artificial Intelligence, Bias and Clinical Safety. *BMJ Quality and Safety 28*(3): 231–237.

Chamola, V., Hassija, V., Sulthana, A.R., Ghosh, D., Dhingra, D. & Sikdar, B. 2023. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access 11*(Ml): 78994–79015.

Chawla, N. V, Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Int. Res. 16*(1): 321–357.

Cheng, X. 2024. A Comprehensive Study of Feature Selection Techniques in Machine Learning Models. *Insights in Computer, Signals and Systems 1*: 1–14.

Chhillar, I. & Singh, A. 2025. An improved soft voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing. *Discover Artificial Intelligence 5*(1)

Connolly, P., Flanagan, K. & Fallon, E. 2024. Parameter reduction optimisation for analysis of e-commerce consumer purchase patterns. *Proceedings of the 2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2024* 341–347.

Cortes, C. & Vapnik, V. 1995. Support-vector networks. *Machine Learning 297*(20): 273–297.

Cruz-Ramos, C., García-Avila, O., Almaraz-Damian, J.-A., Ponomaryov, V., Reyes-Reyes, R. & Sadovnychiy, S. 2023. Benign and malignant breast tumor classification in ultrasound and mammography images via fusion of deep learning and handcraft features. *Entropy 25*(7): 991.

Dalwinder, S., Birmohan, S. & Manpreet, K. 2020. simultaneous feature weighting and parameter determination of neural networks using ant lion optimization for the classification of breast cancer. *Biocybernetics and Biomedical Engineering 40*(1): 337–351.

Daoud, M.I., Abdel-Rahman, S. & Alazrai, R. 2019. Breast ultrasound image classification using a pre-trained convolutional neural network. *Proceedings - 15th International Conference on Signal Image Technology and Internet Based Systems, SISITS 2019* 167–171.

Daoud, M.I., Abdel-Rahman, S., Bdair, T.M., Al-Najar, M.S., Al-Hawari, F.H. & Alazrai, R. 2020. Breast tumor classification in ultrasound images using combined deep and handcrafted features. *Sensors (Switzerland) 20*(23): 1–20.

Dash, M. & Liu, H. 1997. Feature selection for classification. *Intelligent Data Analysis 1*(3): 131–156.

Dash, Manoranjan & Liu, H. 2003. Consistency-based search in feature selection. *Artificial Intelligence 151*(1–2): 155–176.

Deisy C, Subramanian, B., Ramaraj, N., Koori, J. & Jeevanandam, P. 2010. A novel information theoretic-interact algorithm (it-in) for feature selection using three machine learning algorithms. *Expert Systems with Applications 37*: 7589–7597.

Demircioğlu, A. 2024. Applying oversampling before cross-validation will lead to high bias in radiomics. *Scientific Reports 14*(1): 1–11. https://doi.org/10.1038/s41598-024-62585-z.

Deshpande, D.S., Rajurkar, A.M. & Manthalkar, R.R. 2015. Texture based associative classifier - An application of data mining for mammogram classification. *Computational Intelligence in Data Mining 31*

Dhanaseelan, F.R. & Sutha, M.J. 2021. Detection of breast cancer based on fuzzy frequent itemsets mining. *IRBM 42*(3): 198–206. https://doi.org/10.1016/j.irbm.2020.05.002.

Di Carlo, F., Nezami, N., Anahideh, H. & Asudeh, A. 2023. FairPilot: An explorative system for hyperparameter tuning through the lens of fairness (April) https://arxiv.org/abs/2304.04679v1.

Duarte, M.A., Pereira, W.C.A. & Alvarenga, A.V. 2020. Calculating texture features from mammograms and evaluating their performance in classifying clusters of microcalcification. *XV Mediterranean Conference on Medical and Biological Engineering and Computing, IFMBE Proceedings* hlm. 322–332 10.1007/978-3-030-31635-8_39.

Dutta, M., Mehedi Hasan, K.M., Akter, A., Rahman, M.H. & Assaduzzaman, M. 2024. An interpretable machine learning-based breast cancer classification using XGBoost, SHAP, and LIME. *Bulletin of Electrical Engineering and Informatics 13*(6): 4306–4315.

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.

Elkorany, A.S., Marey, M., Almustafa, K.M. & Elsharkawy, Z.F. 2022. Breast cancer diagnosis using support vector machines optimized by whale optimization and dragonfly algorithms. *IEEE Access 10*(June): 69688–69699.

Engel, S., Jacobsen, H.B. & Reme, S.E. 2023a. A cross-sectional study of fear of surgery in female breast cancer patients: Prevalence, severity, and sources, as well as relevant differences among patients experiencing high, moderate, and low fear of surgery. *PLoS ONE 18*(6 JUNE): 1–19.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2): 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research 5* 1531–1555.

Geng, X., Yang, Z., Jiao, L., Zhou, Z.-J. & Ma, Z. 2025. Association rule-based classification: A comprehensive review of methodologies and applications. *Expert Systems with Applications 280*: 127454. https://www.sciencedirect.com/science/article/pii/S0957417425010760.

Ghaemi, Z., Tran, T.T.D. & Smith, A.D. 2022. Comparing classical and metaheuristic methods to optimize multi-objective operation planning of district energy systems considering uncertainties. *Applied Energy 321*

Golberg, D.E. 1989. Genetic algorithms in search, optimization, and machine learning. *Addion wesley*

Gu, X., Guo, J., Wei, H. & He, Y. 2020. Spatial-domain steganalytic feature selection based on three-way interaction information and KS test. *Soft Computing 24*(1): 333–340. https://doi.org/10.1007/s00500-019-03910-x.

Gunning, D. & Aha, D.W. 2019. DARPA's explainable artificial intelligence program. *AI Magazine 40*(2): 44–58.

Guo, Q., Liu, K., Xu, T., Wang, P. & Yang, X. 2024. Fuzzy feature factorization machine: Bridging feature interaction, selection, and construction. *Expert Systems with Applications 255*: 124600. https://www.sciencedirect.com/science/article/pii/S0957417424014672.

Guo, Z., Xie, J., Wan, Y., Zhang, M., Qiao, L., Yu, J., Chen, S., Li, B. & Yao, Y. 2022. A Review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis. *Open Life Sciences 17*(1): 1600–1611.

Guyon, I. & Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research 3*: 1157–1182.

Hall, M.A. 2000. Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)* hlm. 1–16

Harikumar, S., Dilipkumar, D.U. & Kaimal, M.R. 2017. Efficient attribute selection strategies for association rule mining in high dimensional data. *International Journal of Computational Science and Engineering 15*(3–4): 201–213.

Hasan, R. & Shafi, A.S.M. 2023. Feature selection based breast cancer prediction. *International Journal of Image, Graphics and Signal Processing 15*(2): 13–23.

Hassan, N.M., Hamad, S. & Mahar, K. 2022. Mammogram breast cancer CAD systems for mass detection and classification: A review. *Multimedia Tools and Applications 81*.

Hastie, T., Tibshirani, R. & Friedman, J. 2009. *The Elements of Statistical Learning*. *Data Mining, Inference, and Prediction*

Herath, H., Herath, H., Madusanka, N. & Lee, B.-I. 2025. A systematic review of medical image quality assessment. *Journal of Imaging 11*(4): 100.

Hooker, G., Mentch, L. & Zhou, S. 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing 31*(6): 1–22.

Hotelling, H. 1933. Analysis of a complex statistical variables into principal components. *Journal of Educational Psychology 24*: 498–520.

Hotelling, H. 1936. relations between two sets of variates. *Biometrika 28*: 3/4.

Hussain, F., Hammad, M. & Ksantini, R. 2021. Application of artificial intelligence in digital breast tomosynthesis and mammography. *2021 International Conference*

on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2021 638–645.

Indira, K. & Kanmani, S. 2015. Association rule mining through adaptive parameter control in particle swarm optimization. *Computational Statistics 30*(1): 251–277.

Islam, T., Sheakh, M.A., Tahosin, M.S., Hena, M.H., Akash, S., Bin Jardan, Y.A., FentahunWondmie, G., Nafidi, H.A. & Bourhia, M. 2024. Predictive modeling for breast cancer classification in the context of bangladeshi patients by use of machine learning approach with explainable AI. *Scientific Reports 14*(1): 1–17. https://doi.org/10.1038/s41598-024-57740-5.

J. Bergstra & Y. Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research 13*: 281–305.

Jain, P., Patel, D., Verma, J.P. & Tanwar, S. 2021. Computer-aided-diagnosis system for symptom detection of breast and cervical cancer. *Lecture Notes in Networks and Systems 203 LNNS*: 743–758.

Jain, R. & Xu, W. 2021. HDSI: High dimensional selection with interactions algorithm on feature selection and testing. *PLoS ONE 16*(2 February): 1–17.

Jajroudi, M., Enferadi, M., Bagherpour, Z. & Reiazi, R. 2024. Association rule mining-based radiomics in breast cancer diagnosis. *Iranian Journal of Medical Physics 21*(2): 113–120.

Jakulin, A. & Bratko, I. 2003. Analyzing attribute dependencies. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science) 2838*: 229–240.

Jakulin, A. & Bratko, I. 2004. Testing the significance of attribute interactions. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004* (September 2004): 409–416.

Jiang, F., Yuen, K.K.R. & Lee, E.W.M. 2020. Analysis of motorcycle accidents using association rule mining-based framework with parameter optimization and GIS technology. *Journal of Safety Research 75*(September): 292–309. https://doi.org/10.1016/j.jsr.2020.09.004.

Jimenez, F., Martinez, C., Marzano, E., Palma, J.T., Sanchez, G. & Sciavicco, G. 2019. Multiobjective evolutionary feature selection for fuzzy classification. *IEEE Transactions on Fuzzy Systems 27*(5): 1085–1099.

John, G.H., Kohavi, R. & Pfleger, K. 1994. Irrelevant feature and the subset selection problem. *Proceedings of 11th International Conference on Machine Learning, New Brunswick, 10-13 July 1994* hlm. 121–129

Jovanovic, L., Zivkovic, M., Antonijevic, M., Jovanovic, D., Ivanovic, M. & Jassim, H.S. 2022. An emperor penguin optimizer application for medical diagnostics.

2022 IEEE Zooming Innovation in Consumer Technologies Conference, ZINC 2022 191–196.

Kabane, S. 2024. Impact of sampling techniques and data leakage on XGBoost performance in credit card fraud detection 1–19. http://arxiv.org/abs/2412.07437.

Kabir, M.F., Ludwig, S.A. & Abdullah, A.S. 2018. Rule discovery from breast cancer risk factors using association rule mining. *Proceedings 2018 IEEE International Conference on Big Data, Big Data 2018* 2433–2441.

Kalita, D.J., Singh, P.V. & Kumar, V. 2022. Two way threshold based intelligent water drops feature selection algorithm for accurate detection of breast cancer. *Soft Computing 26:* 2277-2305

Kanani, D. S. & K. Mishra, S. 2015. An optimized association rule mining using genetic algorithm. *International Journal of Computer Applications 119*(14): 11–15.

Kaoungku, N., Kerdprasop, K. & Kerdprasop, N. 2017. A method to clustering the feature ranking on data classification using an ensemble feature selection. *International Journal of Future Computer and Communication 6*(3): 81–85.

Kaoungku, N., Suksut, K., Chanklan, R., Kerdprasop, K. & Kerdprasop, N. 2017. Data classification based on feature selection with association rule mining. *Lecture Notes in Engineering and Computer Science 2227*: 321–326.

Kaoungku, N., Suksut, K., Chanklan, R., Kerdprasop, K. & Kerdprasop, N. 2018. The silhouette width criterion for clustering and association mining to select image features. *International Journal of Machine Learning and Computing 8*(1): 69–73.

Kapoor, S. & Narayanan, A. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns 4*(9): 100804. https://doi.org/10.1016/j.patter.2023.100804.

Kaushik, M., Sharma, R., Fister, I. & Draheim, D. 2023. Numerical association rule mining: a systematic literature review *1*(1) http://arxiv.org/abs/2307.00662.

Keen, J.D., Keen, J.M. & Keen, J.E. 2018. Utilization of computer-aided detection for digital screening mammography in the United States, 2008 to 2016. *Journal of the American College of Radiology : JACR 15*(1 Pt A): 44–48.

Ketabi, H., Ekhlasi, A. & Ahmadi, H. 2021. A computer-aided approach for automatic detection of breast masses in digital mammogram via spectral clustering and support vector machine. *Physical and Engineering Sciences in Medicine 44*(1): 277–290.

Keyvanpour, M.R., Barani Shirzad, M. & Mahdikhani, L. 2021. WARM: A new breast masses classification method by weighting association rule mining. *Signal, Image and Video Processing* https://doi.org/10.1007/s11760-021-01989-0.

Khanna, M., Singh, L.K., Shrivastava, K. & singh, R. 2024. An enhanced and efficient approach for feature selection for chronic human disease prediction: A breast cancer study. *Heliyon 10*(5): e26799. https://doi.org/10.1016/j.heliyon.2024.e26799.

Khanna, P., Sahu, M. & Kumar Singh, B. 2021. Improving the classification performance of breast ultrasound image using deep learning and optimization algorithm. *2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society, TRIBES 2021*

Kharya, S., Onyema, E.M., Zafar, A., Wajid, M.A., Afriyie, R.K., Swarnkar, T. & Soni, S. 2022. Weighted bayesian belief network: A computational intelligence approach for predictive modeling in clinical datasets. *Computational Intelligence and Neuroscience 2022*

Kianmehr, K., Alshalalfa, M. & Alhajj, R. 2010. Fuzzy clustering-based discretization for gene expression classification. *Knowledge and Information Systems 24*(3): 441–465.

Kim, J., Harper, A., McCormack, V., Sung, H., Houssami, N., Morgan, E., Mutebi, M., Garvey, G., Soerjomataram, I. & Fidler-Benaoudia, M.M. 2025. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature Medicine 31*(4): 1154–1162. http://dx.doi.org/10.1038/s41591-025-03502-3.

Kishor, P. & Porika, S. 2023. A novel association rule mining model for generating positive and negative association rules with hybridized meta-heuristic development. *International Journal of Intelligent Engineering and Systems 16*(2): 125–141.

Kliegr, T. & Kuchař, J. 2019. Tuning hyperparameters of classification based on associations (CBA). *CEUR Workshop Proceedings 2473*: 9–16.

Koemel, N.A., Shah, S., Senior, A.M., Severi, G., Mancini, F.R., Gill, T.P., Simpson, S.J., Raubenheimer, D., Boutron-Ruault, M.C., Laouali, N. & Skilton, M.R. 2024. macronutrient composition of plant-based diets and breast cancer risk: The E3N prospective cohort study. *European Journal of Nutrition 63*(5): 1771–1781.

Kohavi, R. & John, G.H. 1997. Wrappers for feature subset selection. *Artificial Intelligence 97*(1): 273–324.

Kok, I., Okay, F.Y., Muyanli, O. & Ozdemir, S. 2023. Explainable artificial intelligence (XAI) for internet of things: A survey. *IEEE Internet of Things Journal 10*(16): 14764–14779.

Koller, D. & Sahami, M. 1996. Toward optimal feature selection. *International Conference on Machine Learning* 284–292.

Kumar, V. 2014. Feature selection: A literature review. *The Smart Computing Review 4*(3)

Kuo, C.F.J., Chen, H.Y., Barman, J., Ko, K.H. & Hsu, H.H. 2023. complete, fully automatic detection and classification of benign and malignant breast tumors based on CT images using artificial intelligent and image processing. *Journal of Clinical Medicine 12*(4)

Kusumaningrum, R., Nisa, I.Z., Nawangsari, R.P. & Wibowo, A. 2021. Sentiment analysis of indonesian hotel reviews: From classical machine learning to deep learning. *International Journal of Advances in Intelligent Informatics 7*(3): 292–303.

Ladbury, C., Zarinshenas, R., Semwal, H., Tam, A., Vaidehi, N., Rodin, A.S., Liu, A., Glaser, S., Salgia, R. & Amini, A. 2022. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: A narrative review. *Translational Cancer Research 11*(10): 3853–3868.

Lagree, A., Shiner, A., Alera, M.A., Fleshner, L., Law, E., Law, B., Lu, F.I., Dodington, D., Gandhi, S., Slodkowska, E.A., Shenfield, A., Jerzak, K.J., Sadeghi-Naini, A. & Tran, W.T. 2021. Assessment of digital pathology imaging biomarkers associated with breast cancer histologic grade. *Current Oncology 28*(6): 4298–4316.

Lamb, L.R., Fonseca, M.M., Verma, R. & Seely, J.M. 2020. Missed breast cancer: Effects of subconscious bias and lesion characteristics. *Radiographics 40*(4): 941–960.

Lazard, A.J., Nicolla, S., Vereen, R.N., Pendleton, S., Charlot, M., Tan, H.-J., DiFranzo, D., Pulido, M. & Dasgupta, N. 2023. Exposure and reactions to cancer treatment misinformation and advice: Survey study. *JMIR cancer 9*: e43749.

Le Cessie, S. & Van Houwelingen, J.C. 1992. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 41*(1): 191–201. http://www.jstor.org/stable/2347628.

Lewis, D.D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. Dlm. Nédellec, C. & Rouveirol, C. (pnyt). *Machine Learning: ECML-98* hlm. 4–15. Berlin, Heidelberg: Springer Berlin Heidelberg.

Li, A., Liu, L., Ullah, A., Wang, R., Ma, J., Huang, R., Yu, Z. & Ning, H. 2019. Association rule-based breast cancer prevention and control system. *IEEE Transactions on Computational Social Systems 6*(5): 1106–1114.

Li, F., Shang, C., Li, Y. & Shen, Q. 2020. Interpretable mammographic mass classification with fuzzy interpolative reasoning. *Knowledge-Based Systems 191*(xxxx)

Li, H., Whitney, H.M., Yu, J., Edwards, A., John, P., Peifang, L. & Giger, M.L. 2022. Impact of continuous learning on diagnostic breast MRI AI: Evaluation on an independent clinical dataset. *Journal of Medical Imaging*

Li, Jia, Shi, J., Chen, J., Du, Z. & Huang, L. 2023. Self-attention random forest for breast cancer image classification. *Frontiers in Oncology 13*(February): 1–14.

Li, Jundong, Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. & Liu, H. 2017. Feature selection: A data perspective. *ACM Computing Surveys 50*(6)

Liang, J., Hou, L., Luan, Z. & Huang, W. 2019. Feature selection with conditional mutual information considering feature interaction. *Symmetry*

Liew, X.Y., Hameed, N. & Clos, J. 2021. A review of computer-aided expert systems for breast cancer diagnosis. *Cancers 13*(11)

Liewlom, P. 2021. Class-association-rules pruning by the profitability-of-interestingness measure: Case study of an imbalanced class ratio in a breast cancer dataset. *Journal of Advances in Information Technology 12*(3): 246–252.

Lim, Y.X., Lim, Z.L., Ho, P.J. & Li, J. 2022. Breast cancer in asia: Incidence, mortality, early detection, mammography programs, and risk-based screening initiatives. *Cancers 14*(17): 1–21.

Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. 2021. Explainable AI: A review of machine learning interpretability methods. *Entropy 23*(1): 1–45.

Liu, H., Wang, J., Gao, J., Liu, S., Liu, X., Zhao, Z., Guo, D. & Dan, G. 2020. A comprehensive hierarchical classification based on multi-features of breast DCE-MRI for cancer diagnosis. *Medical and Biological Engineering and Computing 58*(10): 2413–2425.

Lundberg, S.M. & Lee, S.I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 2017-Decem*(Section 2): 4766–4775.

Luo, S. & Chen, Z. 2021. Sequential interaction group selection by the principle of correlation search for high-dimensional interaction models. *Statistica Sinica 31*(1): 197–221.

Mahesh, T.R., Vinoth Kumar, V., Vivek, V., Karthick Raghunath, K.M. & Sindhu Madhuri, G. 2024. Early predictive model for breast cancer classification using blended ensemble learning. *International Journal of System Assurance Engineering and Management 15*(1): 188–197. https://doi.org/10.1007/s13198-022-01696-0.

Mamdouh Farghaly, H. & Abd El-Hafeez, T. 2023. A high-quality feature selection method based on frequent and correlated items for text classification. *Soft Computing 27*(16): 11259–11274.

Manikandan, L.C. & Selvakumar, R.K. 2022. A review on data mining concepts and tools. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology 3307*: 600–605.

Marino, A. Di, Bevilacqua, V., Ciaramella, A., De Falco, I. & Sannino, G. 2025. Ante-hoc methods for interpretable deep models: A survey *57*(10)

Meyer, P.E., Schretter, C. & Bontempi, G. 2008. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal on Selected Topics in Signal Processing 2*(3): 261–274.

Michael, E., Ma, H., Li, H., Kulwa, F. & Li, J. 2021. Breast cancer segmentation methods: current status and future potentials. *BioMed research international 2021*: 9962109.

Michaels, E., Worthington, R.O. & Rusiecki, J. 2023. Breast cancer: risk assessment, screening, and primary prevention. *The Medical clinics of North America 107*(2): 271–284.

Miglioretti, D.L., Abraham, L., Sprague, B.L., Lee, C.I., Bissell, M.C.S., Ho, T.-Q.H., Bowles, E.J.A., Henderson, L.M., Hubbard, R.A., Tosteson, A.N.A. & Kerlikowske, K. 2024. Association between false-positive results and return to screening mammography in the breast cancer Surveillance Consortium Cohort. *Annals of internal medicine 177*(10): 1297–1307.

Minnoor, M. & Baths, V. 2023. Diagnosis of breast cancer using random forests. *Procedia Computer Science 218*(2022): 429–437. https://doi.org/10.1016/j.procs.2023.01.025.

Mishra, A.K., Roy, P., Bandyopadhyay, S. & Das, S.K. 2021. Breast ultrasound tumour classification: A machine learning—radiomics based approach. *Expert Systems 38*(7): 1–12.

Mohammed, S.A., Darrab, S., Noaman, S.A. & Saake, G. 2020. Analysis of breast cancer detection using different machine learning techniques. *Communications in Computer and Information Science 1234*. http://dx.doi.org/10.1007/978-981-15-7205-0_10.

Mohd Sagheer, S. V. & George, S.N. 2020. A review on medical image denoising algorithms. *Biomedical Signal Processing and Control 61*: 102036. https://doi.org/10.1016/j.bspc.2020.102036.

Muhtadi, S. 2022. Breast tumor classification using intratumoral quantitative ultrasound descriptors. *Computational and Mathematical Methods in Medicine 2022*

Mumtahina, U., Alahakoon, S. & Wolfs, P. 2024. Hyperparameter Tuning of Load-Forecasting Models Using Metaheuristic Optimization Algorithms—A Systematic Review. *Mathematics 12*(21)

Murugesan, S., Bhuvaneswaran, R.S., Khanna Nehemiah, H., Keerthana Sankari, S. & Nancy Jane, Y. 2021. Feature selection and classification of clinical datasets using bioinspired algorithms and super learner. *Computational and Mathematical Methods in Medicine 2021*

Mutlu, F., Çetinel, G. & Gül, S. 2020. A fully-automated computer-aided breast lesion detection and classification system. *Biomedical Signal Processing and Control 62*(August)

Nguyen, V. 2019. Bayesian optimization for accelerating hyper-parameter tuning. *Proceedings - IEEE 2nd International Conference on Artificial Intelligence and Knowledge Engineering, AIKE 2019* 302–305.

Nikolaev, A.G. & Jacobson, S. 2010. Simulated annealing. *Handbook of Metaheuristics* Vol. 146, hlm. 1–39

Ogedengbe, M., Junaidu, S. & Kana, D. 2024. Adaptive minimum support threshold for association rule mining. *Indonesian Journal of Data and Science 5*(2): 101–108.

Oladipupo, O., Olajide, O., Adubi, S., Oyelade, J. & Omogbadegun, Z. 2021. An interval type-2 fuzzy association rule mining approach to pattern discovery in breast cancer dataset. *Journal of Computer Science 17*(3): 330–348.

Olmo, J.L., Luna, J.M., Romero, J.R. & Ventura, S. 2013. Mining association rules with single and multi-objective grammar guided ant programming. *Integrated Computer-Aided Engineering 20*(3): 217–234.

Omana, S.P. & Jesu, J.C. 2017. JFIM: A novel filter feature selection approach using joint feature interaction maximization. *International Journal of Intelligent Engineering and Systems 10*(6): 203–210.

Ono, Y. & Mitani, Y. 2022. Evaluation of feature extraction methods with ensemble learning for breast cancer classification. *LifeTech 2022 - 2022 IEEE 4th Global Conference on Life Sciences and Technologies* 194–195.

Oussous, A., Benjelloun, F.Z., Ait Lahcen, A. & Belfkih, S. 2018. Big data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences 30*(4): 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001.

Pacheco, S.A.B. 2022. Breast cancer prediction using ensemble technique. *3rd IEEE 2022 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2022* 823–828.

Pacheco-Ortiz, J., Rodríguez-Mazahua, L., Mejía-Miranda, J., Machorro-Cano, I. & Juárez-Martínez, U. 2020. Towards association rule-based item selection strategy in computerized adaptive testing. *Revista Perspectiva Empresarial 7*(2–1): 19–30.

Patial, P., Kaur, K., Bansal, S. & Kaur, R. 2025. Breast cancer prediction based on SMOTE and ensemble classifier. Dlm. Mohanty, S.N., Rocha, Á. & Dutta, P.K. (pnyt). hlm. 253–264. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-94302-7_17.

Patrcio, M., Pereira, J., Crisstomo, J., Matafome, P., Seia, R. & Caramelo, F. 2018. Breast Cancer Coimbra. *UCI Machine Learning Repository*

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11): 559–572.

Peng, H., Long, F. & Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*(8): 1226–1238.

Pramanik, P., Mukhopadhyay, S., Mirjalili, S. & Sarkar, R. 2023. Deep feature selection using local search embedded social ski-driver optimization algorithm for breast cancer detection in mammograms. *Neural Computing and Applications 35*(7): 5479–5499.

Prasetiyowati, M.I., Maulidevi, N.U. & Surendro, K. 2021. determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data 8*(1) https://doi.org/10.1186/s40537-021-00472-4.

Prowal, P., Singh, A.S., Thirunavukkarasu, K., Khan, S. & Qader, M.R. 2021. Machine learning algorithms based on feature selection method used for the prediction of breast cancer. *2021 International Conference on Data Analytics for Business and Industry, ICDABI 2021* 100–106.

Quinlan, J.R. 1986. Induction of decision trees. *Machine Learning 1*(1): 81–106.

Rahman, F., Mehejabin, T., Yeasmin, S. & Sarkar, M. 2020. A comprehensive study of machine learning approach on cytological data for early breast cancer detection. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*

Rajeswari, K. 2015. Feature selection by mining optimized association rules based on Apriori algorithm. *International Journal of Computer Applications 119*(20): 30–34.

Rehman, M.Z.U., Ahmad, J., Jaha, E.S., Ali, A.M., Alzain, M.A. & Saeed, F. 2023. An efficient automated technique for classification of breast cancer using deep ensemble model. *Computer Systems Science and Engineering 46*(1): 897–911.

Remeseiro, B. & Bolon-Canedo, V. 2019. A review of feature selection methods in medical applications. *Computers in Biology and Medicine 112*(February): 103375. https://doi.org/10.1016/j.compbiomed.2019.103375.

Reshma, V.K., Arya, N., Ahmad, S.S., Wattar, I., Mekala, S., Joshi, S. & Krah, D. 2022. Detection of breast cancer using histopathological image classification dataset with deep learning techniques. *BioMed Research International 2022*

Rezaeipanah, A. & Ahmadi, G. 2022. Breast cancer diagnosis using multi-stage weight adjustment in the mlp neural network. *Computer Journal 65*(4): 788–804.

Ribeiro, M.T., Singh, S. & Guestrin, C. 2016. "Why should i trust you?" Explaining the predictions of any classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session* 97–101.

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S. & Scheinost, D. 2024. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications 15*(1): 1–15.

Sadeghi, Z., Alizadehsani, R., CIFCI, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladík, M., Nahavandi, S. & Pardalos, P.M. 2024. A review of explainable artificial intelligence in healthcare. *Computers and Electrical Engineering 118*(PA): 109370. https://doi.org/10.1016/j.compeleceng.2024.109370.

Samee, N.A., Atteia, G., Meshoul, S., Al-antari, M.A. & Kadah, Y.M. 2022. Deep learning cascaded feature selection framework for breast cancer classification: hybrid CNN with univariate-based approach. *Mathematics 10*(19)

Sannasi Chakravarthy, S.R., Rajaguru, H. & Chidambaram, S. 2022. Processing of wisconsin breast cancer data using ebola optimization algorithm with mixture kernel SVM. *Proceedings - 2nd International Conference on Smart Technologies, Communication and Robotics 2022, STCR 2022* (December): 1–4.

Sha, Z., Hu, L. & Rouyendegh, B.D. 2020. Deep learning and optimization algorithms for automatic breast cancer detection. *International Journal of Imaging Systems and Technology 30*(2): 495–506.

Shakhovska, N., Shebeko, A. & Prykarpatskyy, Y. 2024. A novel explainable AI model for medical data analysis. *Journal of Artificial Intelligence and Soft Computing Research 14*(2): 121–137.

Shanti, M.A. 2024. Optimizing predictive accuracy : A comparative study of feature selection strategies in the healthcare domain *15*: 217–229.

Shi, L., Rahman, R., Melamed, E., Gwizdka, J., Rousseau, J.F. & Ding, Y. 2023. Using explainable AI to cross-validate socio-economic disparities among covid-19 patient mortality. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2023*: 477–486.

Shia, W.C., Lin, L.S. & Chen, D.R. 2021. Classification of malignant tumours in breast ultrasound using unsupervised machine learning approaches. *Scientific Reports 11*(1): 1–11. https://doi.org/10.1038/s41598-021-81008-x.

Shishkin, A., Bezzubtseva, A., Drutsa, A., Shishkov, I., Gladkikh, E., Gusev, G. & Serdyukov, P. 2016. Efficient high-order interaction-aware feature selection based on conditional mutual information. *Advances in Neural Information Processing Systems* (Nips): 4644–4652.

Šikonja, M.-R. & Kononenko, I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning 53*: 23–69. http://lkm.fri.uni-lj.si/xaigor/slo/clanki/MLJ2003-FinalPaper.pdf.

Silva-Aravena, F., Núñez Delafuente, H., Gutiérrez-Bahamondes, J.H. & Morales, J. 2023. A hybrid algorithm of ML and XAI to prevent breast cancer: A strategy to support decision making. *Cancers 15*(9): 1–18.

Silveira, C.L.B., Tabares, A., Faria, L.T. & Franco, J.F. 2021. Mathematical optimization versus metaheuristic techniques: A performance comparison for reconfiguration of distribution systems. *Electric Power Systems Research*

Simon, S., Kolyada, N., Akiki, C., Potthast, M., Stein, B. & Siegmund, N. 2023. Exploring hyperparameter usage and tuning in machine learning research. *2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN)* hlm. 68–79

Singh, Amitojdeep, Sengupta, S. & Lakshminarayanan, V. 2020. Explainable deep learning models in medical image analysis. *Journal of Imaging 6*(6): 1–19.

Singh, Anoop & Sivakkumar, M. 2021. Breast cancer detection based on feature optimization and pulse coupled neural network model. *Proceedings of International Conference on Advances in Technology, Management and Education, ICATME 2021* 54–58.

Singh, H., Sharma, V. & Singh, D. 2022. Comparative analysis of proficiencies of various textures and geometric features in breast mass classification using K-Nearest Neighbor. *Visual Computing for Industry, Biomedicine, and Art 5*(3)

Siswanto, B., Soeparno, H., Sianipar, N.F. & Budiharto, W. 2024. SDFP-Growth algorithm as a novelty of association rule mining optimization. *IEEE Access 12*(January): 21491–21502.

Skalak, D.B. 1994. Prototype and feature selection by sampling and random mutation hill climbing algorithms. *Proceedings of the 11th International Conference on Machine Learning, ICML 1994* 293–301.

Snoek, J., Larochelle, H. & Adams, R.P. 2012. Practical bayesian optimization of machine learning algorithms. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12 hlm. 2951–2959. Red Hook, NY, USA: Curran Associates Inc.

Sobhan, M. & Mondal, A. 2023. Evaluating SHAP's robustness in precision medicine : Effect of filtering and normalization *Proceedings IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* hlm. 3157–3164.

Song, W., Duan, Z., Xu, Y., Shi, C., Zhang, M., Xiao, Z. & Tang, J. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. *International Conference on Information and Knowledge Management, Proceedings* 1161–1170.

Sowan, B., Eshtay, M., Dahal, K., Qattous, H. & Zhang, L. 2023. Hybrid PSO feature selection based association classification approach for breast cancer detection. *Neural Computing and Applications 35*: 5291–5317.

Srinivasu, P.N., Jaya Lakshmi, G., Gudipalli, A., Narahari, S.C., Shafi, J., Woźniak, M. & Ijaz, M.F. 2024. XAI-driven CatBoost multi-layer perceptron neural network for analyzing breast cancer. *Scientific Reports 14*(1): 1–19.

Stephan, P., Stephan, T., Kannan, R. & Abraham, A. 2021. A hybrid artificial bee colony with whale optimization algorithm for improved breast cancer diagnosis. *Neural Computing and Applications 33*(20): 13667–13691.

Su, T., Xu, H. & Zhou, X. 2019. Particle swarm optimization-based association rule mining in big data environment. *IEEE Access 7*: 161008–161016.

Subasi, A. 2020. *Machine learning techniques*.

Sylvester, S., Sagehorn, M., Gruber, T., Atzmueller, M. & Schöne, B. 2024. SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI Methods. *Behavior Research Methods 56*(6): 6067–6081.

Tang, J., Alelyani, S. & Liu, H. 2014. Feature selection for classification: A review. *Data Classification: Algorithms and Applications* 37–64.

Tangherloni, A., Cazzaniga, P., Stranieri, N., Buffa, F.M. & Nobile, M.S. 2024. A fast feature selection for interpretable modeling based on fuzzy inference systems. *21st IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2024*.

Thakur, N., Kumar, P. & Kumar, A. 2024. A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities. *Multimedia Tools and Applications 83*(12): 35849–35942. https://doi.org/10.1007/s11042-023-16634-w.

Tharwat, A. 2018. Classification assessment methods. *Applied Computing and Informatics 17*(1): 168–192.

Thatha, V.N., Chalichalamala, S., Pamula, U., Krishna, D.P., Chinthakunta, M., Mantena, S.V., Vahiduddin, S. & Vatambeti, R. 2025. Optimized machine learning mechanism for big data healthcare system to predict disease risk factor. *Scientific Reports 15*(1): 1–20.

Thawkar, S. & Ingolikar, R. 2020. Classification of masses in digital mammograms using the genetic ensemble method. *Journal of Intelligent Systems 29*(1): 831–845.

Thawkar, S., Singh, L.K. & Khanna, M. 2021. Multi-objective techniques for feature selection and classification in digital mammography. *Intelligent Decision Technologies 15*(1): 115–125.

Theng, D. & Bhoyar, K.K. 2024. Feature selection techniques for machine learning: a survey of more than two decades of research. *Knowledge and Information Systems 66*(3): 1575–1637.

Triantali, D.G., Kypriadis, A.D. & Parsopoulos, K.E. 2022. Optimization-based association rule mining using weighted transactions: A case study. *Proceedings - 2022 International Conference on Frontiers of Communications, Information System and Data Science, CISDS 2022* 41–45.

Tunali, O. & Bayrak, A.T. 2021. Generic association rule engine optimization using item support projection. *Proceedings - 2021 2nd International Conference on Computing and Data Science, CDS 2021* 524–529.

Uddin, K.M.M., Biswas, N., Rikta, S.T. & Dey, S.K. 2023. Machine learning-based diagnosis of breast cancer utilizing feature optimization technique. *Computer Methods and Programs in Biomedicine Update 3*(February): 100098. https://doi.org/10.1016/j.cmpbup.2023.100098.

Umer, M.J., Sharif, M., Alhaisoni, M., Tariq, U., Kim, Y.J. & Chang, B. 2023. A framework of deep learning and selection-based breast cancer detection from histopathology images. *Computer Systems Science and Engineering 45*(2): 1001–1016.

Van der Velden, B.H.M., Kuijf, H.J., Gilhuijs, K.G.A. & Viergever, M.A. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis 79*

Veeramuthu, A. & Meenakshi, S. 2014. An efficient approach for medical image classification using association rules. *International Journal of Computer Applications 90*(3): 1–6.

Venkatesh, B. & Anuradha, J. 2019. A review of feature selection and its methods. *Cybernetics and Information Technologies 19*(1): 3–26.

Veroneze, R., Bertrand Van Ouytsel, C.H., Dam, K.H.T. & Legay, A. 2024. Feature selection for packer classification based on association rule mining. *Engineering Applications of Artificial Intelligence*

Vijayasarveswari, V., Andrew, A.M., Jusoh, M., Sabapathy, T., Raof, R.A.A., Yasin, M.N.M., Ahmad, R.B., Khatun, S. & Rahim, H.A. 2020. Multi-stage feature selection (MSFS) algorithm for UWB-based early breast cancer size prediction. *PLoS ONE 15*(8 August): 1–21.

Wan, J., Chen, H., Li, T., Yang, X. & Sang, B. 2021. Dynamic interaction feature selection based on fuzzy rough set. *Information Sciences 581*: 891–911.

Wang, Gang, Lochovsky, F.H. & Yang, Q. 2004. Feature selection with conditional mutual information maximin in text categorization. *International Conference on Information and Knowledge Management, Proceedings* (November 2004): 342–349.

Wang, Guangtao & Song, Q. 2012. Selecting feature subset via constraint association rules. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7301 LNAI*(PART 2): 304–321.

Wang, Guangtao & Song, Q. 2013. A novel feature subset selection algorithm based on association rule mining. *Intelligent Data Analysis 17*(5): 803–835.

Wang, Haitian, Lo, S.H., Zheng, T. & Hu, I. 2012. Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics 28*(21): 2834–2842.

Wang, Hongnian, Zhang, M., Mai, L., Li, X., Bellou, A. & Wu, L. 2025. An effective multi-step feature selection framework for clinical outcome prediction using electronic medical records. *BMC Medical Informatics and Decision Making 25*(1)

Wang, Lianxi, Jiang, S. & Jiang, S. 2021. A feature selection method via analysis of relevance, redundancy, and interaction. *Expert Systems with Applications 183*(June): 115365. https://doi.org/10.1016/j.eswa.2021.115365.

Wang, Ling, Ma, Q. & Meng, J. 2019. Incremental fuzzy association rule mining for classification and regression. *IEEE Access 7*: 121095–121110.

Wang, Lulu & Liandong, Y. 2019. *Introductory Chapter: Computer-Aided Diagnosis for Biomedical Applications*. *Computer Architecture in Industrial, Biomechanical and Biomedical Engineering*. Intechopen.

Wang, S., Dai, Y., Shen, J. & Xuan, J. 2021. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports 11*(1): 1–11. https://doi.org/10.1038/s41598-021-03430-5.

Wang, Wenjing, Guo, Min, Han, Tongtong & Ning, Shiyong. 2023. A novel feature selection method considering feature interaction in neighborhood rough set. *Intelligent Data Analysis 27*(2): 345–359.

Whitney, H.M., Li, H., Ji, Y., Liu, P. & Giger, M.L. 2021. Multi-stage harmonization for robust AI across breast MR databases. *Cancers 13*(19)

Winsberg, F., Elkin, M., Macy, J., Bordaz, V. & Weymouth, W. 1967. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology 89*(2): 211–215.

Wolberg, W., Mangasarian, O., Street, N. & Street, W. 1995. Breast Cancer Wisconsin (Diagnostic). *UCI Machine Learning Repository*

Wolberg, W., Street, W. & Mangasarian, Olvi. 1995. Breast Cancer Wisconsin (Prognostic). *UCI Machine Learning Repository*

World Health Organization. 2024. Breast Cancer https://www.who.int/news-room/fact-sheets/detail/breast-cancer [3 October 2024].

Wu, S., Yu, S., Yang, Y. & Xie, Y. 2013. Feature and contrast enhancement of mammographic image based on multiscale analysis and morphology. *Computational and Mathematical Methods in Medicine 2013*

Xie, J., Wu, J. & Qian, Q. 2009. feature selection algorithm based on association rules mining method. *Proceedings of the 2009 8th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2009* 357–362.

Yang, H.H. & Moody, J. 1999. Feature selection based on joint mutual information. *Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis* 22–25.

Yu, L. & Liu, H. 2003. Feature selection for high-dimensional data: a fast correlation-based filter solution. *Proceedings, Twentieth International Conference on Machine Learning 2*(January 2003): 856–863.

Zadeh, L.A. 1965. Fuzzy sets. *Information and Control 8*: 338–353.

Zeid, M.A.E., El-Bahnasy, K. & Abu-Youssef, S.E. 2022. An efficient optimized framework for analyzing the performance of breast cancer using machine learning algorithms. *Journal of Theoretical and Applied Information Technology 100*(14): 5165–5178.

Zeng, Z., Zhang, H., Zhang, R. & Yin, C. 2015. A novel feature selection method considering feature interaction. *Pattern Recognition 48*(8): 2656–2666.

Zeng, Z., Zhang, H., Zhang, R. & Zhang, Y. 2015. A mixed feature selection method considering interaction. *Mathematical Problems in Engineering 2015*

Zhang, L. 2021. A Feature Selection Algorithm integrating maximum classification information and minimum interaction feature dependency information. *Computational Intelligence and Neuroscience 2021*

Zhang, S.J. & Zhou, Q. 2011. new feature subset selection algorithm using class association rules mining. *Key Engineering Materials 474–476*: 622–625.

Zhao, Z. & Liu, H. 2007. Searching for interacting features. *IJCAI International Joint Conference on Artificial Intelligence* (3): 1156–1161.

Zhao, Z. & Liu, H. 2009. Searching for interacting features in subset selection. *Intelligent Data Analysis 13*(2): 207–228.

Zhou, X., Wang, B., Liang, L., Yan, J. & Pan, D. 2019. An operational parameter optimization method based on association rules mining for chiller plant. *Journal of Building Engineering 26*(July): 100870. https://doi.org/10.1016/j.jobe.2019.100870.

Žlahtič, B., Završnik, J., Blažun Vošner, H. & Kokol, P. 2024. Transferring black-box decision making to a white-box model. *Electronics (Switzerland) 13*(10)

LAMPIRAN A

HASIL PRESTASI KAEDAH FD-CARI BAGI SETIAP MODEL PENGELASAN

a) Prestasi kaedah FD-CARI bagi setiap model pengelasan pada 10 percubaan bagi set data WDBC

Model Pengelasan	Percubaan	AUC	ACC	SEN	SPE	F1
DT	1	0.9676	0.9474	0.9155	0.9664	0.9286
	2	0.9573	0.9421	0.8873	0.9748	0.9197
	3	0.9502	0.9000	0.8732	0.9160	0.8671
	4	0.9817	0.9316	0.8592	0.9748	0.9037
	5	0.9699	0.8947	0.8451	0.9244	0.8571
	6	0.9675	0.9158	0.8873	0.9328	0.8873
	7	0.9699	0.9263	0.9577	0.9076	0.9067
	8	0.9551	0.9105	0.9296	0.8992	0.8859
	9	0.9844	0.9474	0.8592	1.0000	0.9242
	10	0.9793	0.9263	0.8873	0.9496	0.9000
RF	1	0.9879	0.9421	0.9437	0.9412	0.9241
	2	0.9822	0.9526	0.9155	0.9748	0.9353
	3	0.9833	0.9474	0.9296	0.9580	0.9296
	4	0.9929	0.9474	0.9437	0.9496	0.9306
	5	0.9821	0.9263	0.9296	0.9244	0.9041
	6	0.9760	0.9105	0.9014	0.9160	0.8828
	7	0.9831	0.9526	0.9296	0.9664	0.9362
	8	0.9815	0.9211	0.9014	0.9328	0.8951
	9	0.9933	0.9789	0.9437	1.0000	0.9710
	10	0.9911	0.9526	0.9577	0.9496	0.9379
NB	1	0.9966	0.9632	0.9014	1.0000	0.9481
	2	0.9890	0.9368	0.8310	1.0000	0.9077
	3	0.9847	0.9421	0.8592	0.9916	0.9173
	4	0.9975	0.9526	0.8732	1.0000	0.9323
	5	0.9917	0.9263	0.8028	1.0000	0.8906
	6	0.9866	0.9316	0.8169	1.0000	0.8992
	7	0.9849	0.9632	0.9155	0.9916	0.9489
	8	0.9896	0.9474	0.8592	1.0000	0.9242
	9	0.9934	0.9579	0.8873	1.0000	0.9403
	10	0.9969	0.9579	0.9014	0.9916	0.9412
LR	1	0.9974	0.9737	0.9577	0.9832	0.9645
	2	0.9908	0.9632	0.9296	0.9832	0.9496
	3	0.9869	0.9474	0.9437	0.9496	0.9306
	4	0.9995	0.9842	0.9718	0.9916	0.9787
	5	0.9929	0.9474	0.9296	0.9580	0.9296

bersambung...

…sambungan

	6	0.9897	0.9579	0.9577	0.9580	0.9444
	7	0.9847	0.9632	0.9718	0.9580	0.9517
	8	0.9886	0.9526	0.9437	0.9580	0.9371
	9	0.9915	0.9684	0.9437	0.9832	0.9571
	10	0.9973	0.9789	0.9859	0.9748	0.9722
SVM	1	0.9981	0.9737	0.9718	0.9748	0.9650
	2	0.9902	0.9684	0.9437	0.9832	0.9571
	3	0.9856	0.9421	0.9155	0.9580	0.9220
	4	0.9988	0.9789	0.9577	0.9916	0.9714
	5	0.9928	0.9526	0.9296	0.9664	0.9362
	6	0.9895	0.9632	0.9577	0.9664	0.9510
	7	0.9847	0.9632	0.9577	0.9664	0.9510
	8	0.9898	0.9579	0.9437	0.9664	0.9437
	9	0.9910	0.9737	0.9577	0.9832	0.9645
	10	0.9969	0.9737	0.9718	0.9748	0.9650

b) Prestasi kaedah FD-CARI bagi setiap model pengelasan pada 10 percubaan bagi set data WPBC

Model Pengelasan	**Percubaan**	**AUC**	**ACC**	**SEN**	**SPE**	**F1**
DT	1	0.7040	0.6462	0.6667	0.6400	0.4651
	2	0.7207	0.6308	0.7333	0.6000	0.4783
	3	0.7627	0.6769	0.8000	0.6400	0.5333
	4	0.6373	0.6462	0.7333	0.6200	0.4889
	5	0.6493	0.6154	0.6000	0.6200	0.4186
	6	0.7547	0.6615	0.6667	0.6600	0.4762
	7	0.6273	0.6923	0.4000	0.7800	0.3750
	8	0.7100	0.6154	0.7333	0.5800	0.4681
	9	0.7127	0.6923	0.6000	0.7200	0.4737
	10	0.5867	0.5538	0.4000	0.6000	0.2927
RF	1	0.7267	0.6462	0.6000	0.6600	0.4390
	2	0.6307	0.6308	0.4667	0.6800	0.3684
	3	0.7720	0.8000	0.6000	0.8600	0.5806
	4	0.6060	0.6615	0.4667	0.7200	0.3889
	5	0.5680	0.5538	0.4000	0.6000	0.2927
	6	0.7653	0.7231	0.6000	0.7600	0.5000
	7	0.5787	0.6154	0.4667	0.6600	0.3590
	8	0.7400	0.7385	0.5333	0.8000	0.4848
	9	0.6880	0.6923	0.6667	0.7000	0.5000
	10	0.6093	0.6615	0.2667	0.7800	0.2667
NB	1	0.7853	0.6000	0.8000	0.5400	0.4800

bersambung...

…sambungan

	2	0.5387	0.5231	0.6000	0.5000	0.3673
	3	0.7293	0.6769	0.6667	0.6800	0.4878
	4	0.6787	0.6462	0.7333	0.6200	0.4889
	5	0.6880	0.5538	0.6667	0.5200	0.4082
	6	0.7933	0.7385	0.8000	0.7200	0.5854
	7	0.7227	0.6000	0.7333	0.5600	0.4583
	8	0.6627	0.6000	0.6667	0.5800	0.4348
	9	0.6107	0.5077	0.6000	0.4800	0.3600
	10	0.5800	0.5385	0.6000	0.5200	0.3750
LR	1	0.7253	0.6308	0.9333	0.5400	0.5385
	2	0.7080	0.5846	0.7333	0.5400	0.4490
	3	0.7733	0.6308	0.8000	0.5800	0.5000
	4	0.6867	0.5692	0.8667	0.4800	0.4815
	5	0.7273	0.5385	0.6667	0.5000	0.4000
	6	0.8587	0.6462	0.9333	0.5600	0.5490
	7	0.7180	0.5692	0.7333	0.5200	0.4400
	8	0.7467	0.5538	0.8000	0.4800	0.4528
	9	0.7733	0.5692	0.9333	0.4600	0.5000
	10	0.6653	0.4769	0.8000	0.3800	0.4138
SVM	1	0.8080	0.6154	0.8000	0.5600	0.4898
	2	0.6280	0.5846	0.6667	0.5600	0.4255
	3	0.7653	0.6769	0.6667	0.6800	0.4878
	4	0.6933	0.6154	0.8000	0.5600	0.4898
	5	0.6973	0.5846	0.6667	0.5600	0.4255
	6	0.8587	0.7077	0.8000	0.6800	0.5581
	7	0.7400	0.5692	0.7333	0.5200	0.4400
	8	0.7600	0.6462	0.8000	0.6000	0.5106
	9	0.7720	0.6154	0.8667	0.5400	0.5098
	10	0.6293	0.5231	0.6000	0.5000	0.3673

c) Prestasi kaedah FD-CARI bagi setiap model pengelasan pada 10 percubaan bagi set data BCC

Model Pengelasan	**Percubaan**	**AUC**	**ACC**	**SEN**	**SPE**	**F1**
DT	1	0.8209	0.7179	0.6818	0.7647	0.7317
	2	0.7193	0.7179	0.8636	0.5294	0.7755
	3	0.6644	0.6667	0.7727	0.5294	0.7234
	4	0.8770	0.7949	0.6818	0.9412	0.7895
	5	0.7567	0.7436	0.7727	0.7059	0.7727
	6	0.7045	0.6923	0.7273	0.6471	0.7273

bersambung...

…sambungan

	7	0.8035	0.7436	0.6364	0.8824	0.7368
	8	0.7487	0.7692	0.8636	0.6471	0.8085
	9	0.7166	0.5897	0.4545	0.7647	0.5556
	10	0.8650	0.8205	0.7727	0.8824	0.8293
RF	1	0.9144	0.8462	0.9091	0.7647	0.8696
	2	0.8249	0.7436	0.7727	0.7059	0.7727
	3	0.7754	0.7692	0.8182	0.7059	0.8000
	4	0.8543	0.7692	0.8182	0.7059	0.8000
	5	0.8061	0.7949	0.8182	0.7647	0.8182
	6	0.7447	0.6667	0.6364	0.7059	0.6829
	7	0.8743	0.7692	0.6818	0.8824	0.7692
	8	0.8543	0.7949	0.9091	0.6471	0.8333
	9	0.8209	0.7436	0.7273	0.7647	0.7619
	10	0.9425	0.8205	0.9091	0.7059	0.8511
NB	1	0.9412	0.7949	0.6818	0.9412	0.7895
	2	0.7914	0.5128	0.2273	0.8824	0.3448
	3	0.7193	0.5385	0.3182	0.8235	0.4375
	4	0.7380	0.6410	0.4091	0.9412	0.5625
	5	0.7460	0.6154	0.4545	0.8235	0.5714
	6	0.6952	0.4872	0.1818	0.8824	0.2857
	7	0.8396	0.5897	0.3182	0.9412	0.4667
	8	0.7139	0.6154	0.5000	0.7647	0.5946
	9	0.6979	0.6410	0.4091	0.9412	0.5625
	10	0.8904	0.6667	0.4091	1.0000	0.5806
LR	1	0.9920	0.8718	0.7727	1.0000	0.8718
	2	0.7246	0.6923	0.7273	0.6471	0.7273
	3	0.7005	0.6410	0.5909	0.7059	0.6500
	4	0.8102	0.7436	0.6818	0.8235	0.7500
	5	0.7807	0.6667	0.5455	0.8235	0.6486
	6	0.7433	0.6154	0.5000	0.7647	0.5946
	7	0.8182	0.6923	0.5909	0.8235	0.6842
	8	0.7941	0.7436	0.7727	0.7059	0.7727
	9	0.6925	0.6667	0.7273	0.5882	0.7111
	10	0.8663	0.7692	0.7273	0.8235	0.7805
SVM	1	0.9813	0.9231	0.9545	0.8824	0.9333
	2	0.8128	0.7436	0.6818	0.8235	0.7500
	3	0.7701	0.7179	0.6818	0.7647	0.7317
	4	0.8610	0.8462	0.8636	0.8235	0.8636
	5	0.7995	0.6923	0.6364	0.7647	0.7000
	6	0.7754	0.6154	0.5000	0.7647	0.5946

bersambung...

…sambungan

7	0.8690	0.7179	0.5909	0.8824	0.7027
8	0.8904	0.7949	0.7727	0.8235	0.8095
9	0.7647	0.7436	0.7273	0.7647	0.7619
10	0.9144	0.8718	0.8636	0.8824	0.8837

LAMPIRAN B

HASIL PRESTASI KAEDAH FD-CARI+SA BAGI SETIAP MODEL PENGELASAN DAN SUBSET FITUR AKHIR BAGI 20 LELARAN

1a) Prestasi kaedah FD-CARI+SA bagi setiap model pengelasan pada 20 lelaran bagi set data WDBC

Lelaran	Model Pengelasan					Purata Skor AUC
	DT	RF	NB	LR	SVM	
1	0.9630 ± 0.0182	0.9854 ± 0.0054	0.9800 ± 0.0057	0.9794 ± 0.0087	0.9846 ± 0.0054	0.9785
2	0.9653 ± 0.0097	0.9914 ± 0.0054	0.9861 ± 0.0048	0.9869 ± 0.0058	0.9918 ± 0.0050	0.9843
3	0.9713 ± 0.0105	0.9854 ± 0.0056	0.9774 ± 0.0049	0.9792 ± 0.0067	0.9860 ± 0.0052	0.9798
4	0.9533 ± 0.0121	0.9944 ± 0.0033	0.9875 ± 0.0040	0.9867 ± 0.0053	0.9933 ± 0.0030	0.9831
5	0.9653 ± 0.0097	0.9914 ± 0.0054	0.9861 ± 0.0048	0.9869 ± 0.0058	0.9918 ± 0.0050	0.9843
6	0.9645 ± 0.0121	0.9853 ± 0.0053	0.9831 ± 0.0051	0.9803 ± 0.0073	0.9848 ± 0.0053	0.9796
7	0.9677 ± 0.0143	0.9853 ± 0.0053	0.9748 ± 0.0050	0.9800 ± 0.0068	0.9847 ± 0.0060	0.9785
8	0.9641 ± 0.0120	0.9943 ± 0.0033	0.9827 ± 0.0055	0.9884 ± 0.0057	0.9919 ± 0.0048	0.9843
9	0.9645 ± 0.0121	0.9853 ± 0.0053	0.9831 ± 0.0051	0.9803 ± 0.0073	0.9849 ± 0.0053	0.9796
10	0.9675 ± 0.0155	0.9915 ± 0.0048	0.9873 ± 0.0054	0.9867 ± 0.0063	0.9919 ± 0.0047	0.9850
11	0.9687 ± 0.0071	0.9916 ± 0.0048	0.9895 ± 0.0047	0.9850 ± 0.0051	0.9893 ± 0.0051	0.9848
12	0.9682 ± 0.0127	0.9853 ± 0.0054	0.9740 ± 0.0044	0.9834 ± 0.0063	0.9867 ± 0.0053	0.9795
13	0.9658 ± 0.0108	0.9944 ± 0.0033	0.9844 ± 0.0051	0.9870 ± 0.0064	0.9927 ± 0.0045	0.9849
14	0.9657 ± 0.0108	0.9915 ± 0.0049	0.9832 ± 0.0048	0.9848 ± 0.0070	0.9914 ± 0.0053	0.9833
15	**0.9681 ± 0.0132**	**0.9920 ± 0.0048**	**0.9916 ± 0.0054**	**0.9870 ± 0.0067**	**0.9913 ± 0.0048**	**0.9860**
16	0.9450 ± 0.0141	0.9939 ± 0.0032	0.9835 ± 0.0046	0.9874 ± 0.0057	0.9927 ± 0.0030	0.9805
17	0.9618 ± 0.0114	0.9946 ± 0.0031	0.9891 ± 0.0039	0.9874 ± 0.0061	0.9927 ± 0.0033	0.9851
18	0.9705 ± 0.0116	0.9920 ± 0.0047	0.9828 ± 0.0040	0.9876 ± 0.0051	0.9856 ± 0.0060	0.9837
19	0.9660 ± 0.0110	0.9932 ± 0.0036	0.9861 ± 0.0043	0.9864 ± 0.0048	0.9935 ± 0.0032	0.9850
20	0.9641 ± 0.0120	0.9943 ± 0.0033	0.9827 ± 0.0055	0.9884 ± 0.0057	0.9919 ± 0.0048	0.9843

1b) Subset fitur akhir selepas permurnian subset lelaran bagi set data WDBC

Lelaran	Bilangan Subset Fitur Awal	Subset Fitur Akhir Selepas Pemurnian Subset Fitur	
		Bilangan Fitur	Senarai Fitur
1	22	4	['compactness_mean', 'concave points_worst', 'perimeter_mean', 'radius_worst']
2	22	6	['compactness_mean', 'concave points_worst', 'perimeter_mean', 'radius_worst', 'symmetry_worst', 'texture_worst']
3	18	3	['compactness_mean', 'concave points_worst', 'radius_worst']
4	23	8	['compactness_mean', 'concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'symmetry_worst', 'texture_worst']
5	22	6	['compactness_mean', 'concave points_worst', 'perimeter_mean', 'radius_worst', 'symmetry_worst', 'texture_worst']
6	24	4	['concave points_worst', 'fractal_dimension_worst', 'perimeter_mean, 'radius_worst']
7	25	4	['concave points_worst', 'fractal_dimension_worst', 'fractal_dimension_worst', 'radius_worst']
8	24	11	['compactness_mean', 'compactness_se', 'concave points_se', 'concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'smoothness_mean', 'symmetry_worst', 'texture_worst']
9	23	4	['concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_worst']
10	22	5	['compactness_mean', 'concave points_worst', 'perimeter_mean', 'radius_worst', 'texture_worst']
11	23	5	['concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_worst', 'texture_worst']
12	26	5	['concave points_se', 'concave points_worst', 'fractal_dimension_worst', 'fractal_dimension_worst', 'radius_worst']
13	23	10	['compactness_mean', 'compactness_se', 'concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'smoothness_mean', 'symmetry_worst', 'texture_worst']
14	23	6	['compactness_mean', 'concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_worst', 'texture_worst']
15	**20**	**3**	**['concave points_worst', 'radius_worst', 'texture_worst']**
16	25	10	['compactness_mean', 'concave points_se', 'concave points_worst', 'fractal_dimension_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'symmetry_worst', 'texture_worst']
17	21	8	['compactness_mean', 'concave points_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'smoothness_mean', 'symmetry_worst', 'texture_worst']

bersambung…

sambungan...

18	26	8	['concave points_se', 'concave points_se', 'concave points_worst', 'fractal_dimension_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_worst', 'texture_worst']
19	24	9	['compactness_mean', 'concave points_se', 'concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'symmetry_worst', 'texture_worst']
20	24	11	['compactness_mean', 'compactness_se', 'concave points_se', 'concave points_worst', 'fractal_dimension_worst', 'perimeter_mean', 'radius_se', 'radius_worst', 'smoothness_mean', 'symmetry_worst', 'texture_worst']

2a) Prestasi kaedah FD-CARI+SA bagi setiap model pengelasan pada 20 lelaran bagi set data WPBC

Lelaran	Model Pengelasan					Purata Skor AUC
	DT	RF	NB	LR	SVM	
1	0.5102 ± 0.0610	0.5114 ± 0.0733	0.6051 ± 0.0445	0.6500 ± 0.0761	0.6448 ± 0.0716	0.5843
2	0.5217 ± 0.0556	0.5785 ± 0.0512	0.6191 ± 0.0714	0.6331 ± 0.0649	0.6329 ± 0.0745	0.5971
3	0.5451 ± 0.0945	0.5741 ± 0.0701	0.5967 ± 0.0448	0.6385 ± 0.0592	0.5922 ± 0.1409	0.5893
4	0.5243 ± 0.0439	0.5126 ± 0.0509	0.6352 ± 0.0573	0.6489 ± 0.0652	0.5680 ± 0.1538	0.5778
5	0.5243 ± 0.0439	0.5126 ± 0.0509	0.6352 ± 0.0573	0.6489 ± 0.0652	0.5817 ± 0.1220	0.5805
6	**0.6214 ± 0.0872**	**0.7069 ± 0.0890**	**0.7061 ± 0.0524**	**0.7400 ± 0.0552**	**0.7429 ± 0.0568**	**0.7035**
7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	0.4896 ± 0.0462	0.5167 ± 0.0644	0.6576 ± 0.0891	0.6322 ± 0.0686	0.6498 ± 0.0834	0.5892
9	0.5240 ± 0.0862	0.5521 ± 0.0771	0.6080 ± 0.0462	0.6540 ± 0.0683	0.6497 ± 0.0674	0.5975
10	0.4938 ± 0.0633	0.5035 ± 0.0518	0.6203 ± 0.0715	0.6403 ± 0.0660	0.6428 ± 0.0791	0.5801
11	0.5243 ± 0.0439	0.5126 ± 0.0509	0.6352 ± 0.0573	0.6489 ± 0.0652	0.5205 ± 0.1685	0.5683
12	0.5104 ± 0.0984	0.4961 ± 0.0768	0.6463 ± 0.0580	0.6525 ± 0.0790	0.6695 ± 0.0658	0.5950
13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
14	0.4872 ± 0.0808	0.5229 ± 0.0597	0.6165 ± 0.0719	0.6353 ± 0.0620	0.6129 ± 0.0908	0.5750
15	0.6171 ± 0.0924	0.6891 ± 0.0902	0.7031 ± 0.0710	0.7457 ± 0.0506	0.7288 ± 0.0644	0.6967
16	0.5243 ± 0.0439	0.5126 ± 0.0509	0.6352 ± 0.0573	0.6489 ± 0.0652	0.5765 ± 0.1456	0.5795
17	0.5243 ± 0.0439	0.5126 ± 0.0509	0.6352 ± 0.0573	0.6489 ± 0.0652	0.5297 ± 0.1676	0.5701
18	**0.6214 ± 0.0872**	**0.7069 ± 0.0890**	**0.7061 ± 0.0524**	**0.7400 ± 0.0552**	**0.7429 ± 0.0568**	**0.7035**
19	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
20	0.4896 ± 0.0462	0.5167 ± 0.0644	0.6576 ± 0.0891	0.6322 ± 0.0686	0.6497 ± 0.0820	0.5892

2b) Subset fitur akhir selepas permurnian subset lelaran bagi set data WPBC

Lelaran	Bilangan Subset Fitur Awal	Subset Fitur Akhir Selepas Pemurnian Subset Fitur	
		Bilangan Fitur	Senarai Fitur
1	8	3	['area2', 'perimeter3', 'tumor_size']
2	14	4	['compactness3', 'radius2', 'radius3', 'tumor_size']
3	14	3	['radius2', 'radius3', 'tumor_size']
4	3	1	['tumor_size']
5	4	1	['tumor_size']
6	**18**	**4**	**['Time', 'radius2', 'radius3', 'tumor_size']**
7	0	0	[]
8	7	3	['compactness3', 'perimeter3', 'tumor_size']
9	11	3	['area2', 'radius3', 'tumor_size']
10	10	4	['area2', 'compactness', 'perimeter3, 'tumor_size']
11	3	1	['tumor_size']
12	7	2	['perimeter3', 'tumor_size']
13	0	0	[]
14	11	4	['area2', 'compactness3', 'radius3', 'tumor_size']
15	15	5	['Time', 'compactness3', 'radius2', 'radius3', 'tumor_size']
16	3	1	['tumor_size']
17	3	1	['tumor_size']
18	**18**	**4**	**['Time, 'radius2', 'radius3', 'tumor_size']**
19	0	0	[]
20	7	3	['compactness3', 'perimeter3', 'tumor_size']

3a) Prestasi kaedah FD-CARI+SA bagi setiap model pengelasan pada 20 lelaran bagi set data BCC

Lelaran	Model Pengelasan					Purata Skor AUC
	DT	RF	NB	LR	SVM	
1	0.6943 ± 0.0694	0.8048 ± 0.0827	0.7695 ± 0.0460	0.7719 ± 0.0796	0.8270 ± 0.0558	0.7735
2	0.6912 ± 0.0712	0.8275 ± 0.0589	0.7647 ± 0.0786	0.8003 ± 0.0601	0.2333 ± 0.0905	0.6634
3	**0.7620 ± 0.0775**	**0.8412 ± 0.0602**	**0.7773 ± 0.0861**	**0.7922 ± 0.0894**	**0.8439 ± 0.0717**	**0.8044**
4	0.6778 ± 0.0812	0.7436 ± 0.0461	0.6546 ± 0.1204	0.6481 ± 0.0638	0.5987 ± 0.1059	0.6645
5	0.6434 ± 0.0740	0.8000 ± 0.0887	0.7008 ± 0.0602	0.7791 ± 0.0808	0.7864 ± 0.0472	0.7420
6	0.6572 ± 0.0488	0.8135 ± 0.0579	0.7821 ± 0.0534	0.7941 ± 0.0587	0.8275 ± 0.0576	0.7749
7	0.7326 ± 0.0794	0.8183 ± 0.0715	0.7831 ± 0.0461	0.7922 ± 0.0623	0.8273 ± 0.0516	0.7907
8	0.7035 ± 0.0792	0.8222 ± 0.0781	0.7658 ± 0.0611	0.7872 ± 0.0785	0.8195 ± 0.0609	0.7796
9	0.7035 ± 0.0792	0.8222 ± 0.0781	0.7658 ± 0.0611	0.7872 ± 0.0785	0.8189 ± 0.0597	0.7795
10	0.6434 ± 0.0740	0.8000 ± 0.0887	0.7008 ± 0.0602	0.7791 ± 0.0808	0.7866 ± 0.0470	0.7420
11	0.6572 ± 0.0488	0.8135 ± 0.0579	0.7821 ± 0.0534	0.7941 ± 0.0587	0.8277 ± 0.0573	0.7749
12	**0.7620 ± 0.0775**	**0.8412 ± 0.0602**	**0.7773 ± 0.0861**	**0.7922 ± 0.0894**	**0.8439 ± 0.0717**	**0.8044**
13	0.7020 ± 0.0595	0.8201 ± 0.0779	0.7941 ± 0.0434	0.7947 ± 0.0723	0.8107 ± 0.0578	0.7843
14	0.6572 ± 0.0488	0.8135 ± 0.0579	0.7821 ± 0.0534	0.7941 ± 0.0587	0.8275 ± 0.0576	0.7749
15	0.6778 ± 0.0812	0.7436 ± 0.0461	0.6546 ± 0.1204	0.6481 ± 0.0638	0.6270 ± 0.0837	0.6702
16	0.7253 ± 0.0863	0.8260 ± 0.0737	0.7936 ± 0.0630	0.8099 ± 0.0585	0.2313 ± 0.1242	0.6772
17	0.6786 ± 0.0476	0.7401 ± 0.0594	0.7317 ± 0.1032	0.7128 ± 0.0526	0.6351 ± 0.1177	0.6997
18	**0.7620 ± 0.0775**	**0.8412 ± 0.0602**	**0.7773 ± 0.0861**	**0.7922 ± 0.0894**	**0.8439 ± 0.0717**	**0.8044**
19	0.6700 ± 0.0659	0.8198 ± 0.0727	0.7831 ± 0.0461	0.7922 ± 0.0623	0.8278 ± 0.0521	0.7786
20	**0.7620 ± 0.0775**	**0.8412 ± 0.0602**	**0.7773 ± 0.0861**	**0.7922 ± 0.0894**	**0.8439 ± 0.0717**	**0.8044**

3b) Subset fitur akhir selepas permurnian subset lelaran bagi set data BCC

Lelaran	Bilangan Subset Fitur Awal	Subset Fitur Akhir Selepas Pemurnian Subset Fitur	
		Bilangan Fitur	Senarai Fitur
1	13	7	['Adiponectin', 'Age', 'Glucose', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
2	17	6	['Adiponectin', 'Age', 'BMI', 'Glucose', 'MCP.1', 'Resistin']
3	**11**	**4**	**['Age', 'Glucose', 'Leptin', 'Resistin']**
4	7	3	['Age', 'Leptin', 'Resistin']
5	13	6	['Adiponectin', 'Age', 'Glucose', 'Leptin', 'MCP.1', 'Resistin']
6	16	8	['Adiponectin', 'Age', 'BMI', 'Glucose', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
7	21	9	['Adiponectin', 'Age', 'BMI', 'Glucose', 'HOMA', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
8	11	5	['Age', 'Glucose', 'Leptin', 'MCP.1', 'Resistin']
9	11	5	['Age', 'Glucose', 'Leptin', 'MCP.1', 'Resistin']
10	13	6	['Adiponectin', 'Age', 'Glucose', 'Leptin', 'MCP.1', 'Resistin']
11	17	8	['Adiponectin', 'Age', 'BMI', 'Glucose', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
12	**9**	**4**	**['Age', 'Glucose', 'Leptin', 'Resistin']**
13	11	6	['Age', 'Glucose', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
14	20	8	['Adiponectin', 'Age', 'BMI', 'Glucose', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
15	7	3	['Age', 'Leptin', 'Resistin']
16	18	7	['Adiponectin', 'Age', 'BMI', 'Glucose', 'Insulin', 'MCP.1', 'Resistin']
17	7	4	['Age', 'Insulin', 'Leptin', 'Resistin']
18	**9**	**4**	**['Age', 'Glucose', 'Leptin', 'Resistin']**
19	19	9	['Adiponectin', 'Age', 'BMI', 'Glucose', 'HOMA', 'Insulin', 'Leptin', 'MCP.1', 'Resistin']
20	**9**	**4**	**['Age', 'Glucose', 'Leptin', 'Resistin']**

LAMPIRAN C

SENARAI PENERBITAN

1. A Karim, S.A., Mohamad, U.H., Nohuddin, P.N.E. (2025). Discovery of Interpretable Patterns of Breast Cancer Diagnosis via Class Association Rule Mining (CARM) with SHAP-based Explainable AI (XAI). Malaysian Journal of Fundamental and Applied Sciences (MJFAS).

2. A Karim, S.A., Mohamad, U.H., Nohuddin, P.N.E. (2025). Interaction-Based Feature Selection Technique Using Fuzzy Discretization and Class Association Rule Mining for Breast Cancer Classification. Malaysian Journal of Fundamental and Applied Sciences (MJFAS).

3. Karim, S.A., Mohamad, U.H., Nohuddin, P.N.E. (2023). Feature Selection Techniques on Breast Cancer Classification Using Fine Needle Aspiration Features: A Comparative Study. In: Badioze Zaman, H., et al. Advances in Visual Informatics. IVIC 2023.

4. Shahiratul A Abd Karim and Puteri N E Nohuddin. (2021). Bibliometric Analysis of Data Mining on Medical Imaging. Asian Conference on Intelligent Computing and Data Sciences (ACIDS) 2021. Journal of Physics: Conference Series.