Terakreditasi SINTA Peringkat 2
Surat Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti No. 10/E/KPT/2019
masa berlaku mulai Vol. 1 No. 1 tahun 2017 s.d Vol. 5 No. 3 tahun 2021

Terbit online pada laman web jurnal: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 3 No. 3 (2019) 458 - 469 ISSN Media Elektronik: 2580-0760

Pendekatan Machine Learning yang Efisien untuk Prediksi Kanker Payudara

Azminuddin I. S. Azis[1], Irma Surya Kumala Idris[2], Budy Santoso[3], Yasin Aril Mustofa[4]
[1,2,3,4]Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Ichsan Gorontalo
[1]azminuddinazis@unisan.ac.id, [2]irmasuryakumala@unisan.ac.id, [3]budysantoso@unisan.ac.id, [4]yasinaril@unisan.ac.id

Abstract

Breast Cancer is the most common cancer found in women and the death rate is still in second place among other cancers. The high accuracy of the machine learning approach that has been proposed by related studies is often achieved. However, without efficient pre-processing, the model of Breast Cancer prediction that was proposed is still in question. Therefore, this research objective to improve the accuracy of machine learning methods through pre-processing: Missing Value Replacement, Data Transformation, Smoothing Noisy Data, Feature Selection / Attribute Weighting, Data Validation, and Unbalanced Class Reduction which is more efficient for Breast Cancer prediction. The results of this study propose several approaches: C4.5 - Z-Score - Genetic Algorithm for Breast Cancer Dataset with 77,27% accuracy, 7-Nearest Neighbor - Min-Max Normalization - Particle Swarm Optimization for Wisconsin Breast Cancer Dataset - Original with 97,85% accuracy, Artificial Neural Network - Z-Score - Forward Selection for Wisconsin Breast Cancer Dataset - Diagnostics with 98,24% accuracy, and 11-Nearest Neighbor - Min-Max Normalization - Particle Swarm Optimization for Wisconsin Breast Cancer Dataset - Prognostic with 83,33% accuracy. The performance of these approaches is better than standard/normal machine learning methods and the proposed methods by the best of previous related studies.

Keywords: *machine learning, breast cancer prediction, missing value replacement, feature selection, unbalanced class*

Abstrak

Kanker Payudara merupakan kanker yang paling sering ditemukan pada wanita dan tingkat kematiannya masih berada pada posisi dua di antara penyakit kanker laiinya. Akurasi yang tinggi dari pendekatan *machine learning* yang diusulkan oleh penelitian-penelitian terkait sering dicapai. Namun tanpa pra-pengolahan yang efisien, maka model prediksi Kanker Payudara yang diusulkan masih diragukan. Oleh karena itu, penelitian ini bertujuan untuk meningkatkan kinerja akurasi metode-metode *machine learning* melalui pra-pengolahan: *Missing Value Replacement*, *Data Transformation*, *Smoothing Noisy Data*, *Feature Selection* / *Attribute Weighting*, *Data Validation*, dan *Unbalanced Class Reduction* yang lebih efisien untuk prediksi Kanker Payudara. Hasil penelitian ini mengusulkan pendekatan: *C4.5 – Z-Score – Genetic Algorithm* untuk *Breast Cancer Dataset* dengan akurasi 77,27%, *7-Nearest Neighbor – Min-Max Normalization – Particle Swarm Optimization* untuk *Wisconsin Breast Cancer Dataset - Original* dengan akurasi 97,85%, *Artificial Neural Network – Z-Score – Forward Selection* untuk *Wisconsin Breast Cancer Dataset – Diagnostic* dengan akurasi 98,24%, dan *11-Nearest Neighbor – Min-Max Normalization – Particle Swarm Optimization* untuk *Wisconsin Breast Cancer Dataset – Prognostic* dengan akurasi 83,33%. Kinerja pendekatan-pendekatan tersebut lebih baik dari metode-metode *machine learning* standar/normal dan metode yang diusulkan penelitian-penelitian terkait sebelumnya yang terbaik.

Kata kunci: machine learning, prediksi kanker payudara, missing value replacement, feature selection, unbalanced class

1. Pendahuluan

Walaupun tingkat kematian yang disebabkan Kanker Payudara (KP) secara bertahap menurun [1], namun masih berada pada posisi dua besar di antara penyakit kanker lainnya bagi wanita [2]-[4]. Menurut *World Health Organization*, KP merupakan kanker yang paling sering ditemukan pada wanita [5]-[7], termasuk di Indonesia sehingga menjadi salah satu program prioritas pemerintah melalui deteksi dininya [8]. Tingkat kejadiannya masih tinggi, rata-rata 19.3 hingga 96 per 100,000 wanita di berbagai benua [9]-[11]. Prediksinya, 475,000 pasien KP di eropa kemungkinan

Diterima Redaksi : 24-10-2019 | Selesai Revisi : 26-11-2019 | Diterbitkan Online : 10-12-2019

akan mati, sedangkan 32,500 dapat terhindar dari kematian [12].

KP termasuk penyakit kanker yang cukup mudah disembuhkan [2]. Namun banyak kasus KP stadium dini tidak menimbulkan gejala [13]. Pengetahuan masyarakat terhadap KP dan resikonya juga masih minim [8]. Sementara kemungkinan penyembuhannya pada tahap awal lebih cerah [5]. Dengan demikian, masalah ini perlu disikapi dengan peningkatan capaian deteksi dini (prediksi) KP.

Pendakatn *Machine Learning* (ML) memiliki potensi besar dalam domain biomedis komputasi [14], dapat membantu mengatasi masalah prediksi KP [3], [5], [15]-[17], sehingga populer digunakan oleh para peneliti (ditunjukkan pada Tabel 1). *Dataset* publik yang umum digunakan oleh para peneliti dalam riset tentang prediksi KP adalah *Breast Cancer Dataset* (BCD), *Wisconsin Breast Cancer Dataset – Original* (WBCDO), *Wisconsin Breast Cancer Dataset – Diagnostic* (WBCDD), dan *Wisconsin Breast Cancer Dataset – Prognostic* (WBCDP) yang tersedia di *UCI Machine Learning Repository* memiliki masalah *Missing Value* (MV), *Noisy Data* (ND), dan *Unbalanced Class* (UC) yang masih kurang dibahas secara mendalam [15].

Selain itu, metode untuk *Data Validation* (DV) seharusnya dapat menjamin bahwa semua sampel digunakan untuk pelatihan dan pengujian model, begitupun setiap label *class* terwakili secara merata pada pelatihan model. Namun masih banyak peneliltian terkait prediksi KP mengabaikan hal ini (ditunjukkan pada Tabel 1). Akurasi yang tinggi mungkin saja bisa diraih, jika DV dilakukan secara tidak efisien.

Sementara itu, salah satu pendekatan untuk dapat meningkatkan kinerja model klasifikasi adalah *Feature Selection* (FS) maupun *Attribute Weighting* (AW). Namun masih banyak penelitian terkait prediksi KP yang belum menerapkan pendekatan ini untuk meningkatkan akurasi tinggi yang telah mereka capai (ditunjukkan pada tabel 1). Untuk itu, metode-metode untuk FS maupun AW perlu dikaji dengan lengkap.

Begitupun dengan *Data Transformation* (DT), misalnya ketika melakukan *Data Discretization* mesti hati-hati agar dapat meminimalkan hilangnya informasi yang mungkin penting. Bukankah lebih efisien apabila diskretisasi data tidak perlu dilakukan selama kinerja dari metode ML yang diperoleh sama baiknya. Misalnya algoritma *Naïve Bayes* yang dapat menggunakan pendekatan *kernel* atau distibusi *Gaussian* untuk menangani data numerik daripada harus melakukan diskretisasi data.

P. H. Abreu *et al*., 2016 dalam studi *review* yang dilakukannya tentang prediksi KP berbasis metode-metode ML menyatakan bahwa penelitian-penelitian yang ada telah mencapai akurasi yang tinggi, tapi sensitifitasnya masih diragukan, masalah MV dan UC masih jarang ditangani, dan kombinasi metode-metode ML yang berbeda masih dibutuhkan [15].

Tabel 1. State of the Art: Breast Cancer Prediction

Year, Ref	Proposed Method	Dataset	DV	MV	ND	UC	DT	FS	Acc
2006, [18]	BN –WNN – MB	ITTACA	!	☑	☑	⊗	☑	☑	0,845
2007, [6]	LSSVM	WBCDO	☑	⊗	☑	⊗	☑	⊗	98,53
2007, [19]	PSO – MSS	WBCDO	!	⊗	☑	⊗	☑	⊗	100
2009, [2]	SVM – F-Score	WBCDO	!	⊗	☑	⊗	☑	☑	99,51
2010, [20]	ANN – RS	WSCRS	☑	?	?	?	?	?	0,965
2011, [3]	SVM – RS	WBCDO	!	!	☑	⊗	☑	☑	100
2011, [21]	NPBC	NBCD	☑	?	☑	?	☑	☑	38,8
2012, [22]	DT - RS – GA	WBCDD	!	☑	⊗	⊗	☑	☑	95,3
2012, [23]	ANFIS – LR	MMD	☑	!	?	☑	☑	☑	0,928
2012, [24]	SVM	KTTH	!	?	☑	⊗	☑	⊗	84,58
2012, [25]	SVM – RFE	GS	☑	☑	☑	⊗	☑	☑	97%
2013, [26]	DT variants (J48)	Various	☑	?	?	?	☑	?	75,52
2013, [27]	Graph-based SSL	SEER	☑	☑	☑	!	☑	⊗	71
2013, [28]	SSL Co-Training	SEER	☑	☑	☑	!	☑	⊗	76
2014, [29]	Graph-based SSL	GEO	☑	☑	☑	?	☑	⊗	76,7
2014, [30]	k-NN	WBCDP	!	☑	☑	⊗	☑	⊗	90
2014, [31]	FMM–CART–RF	WBCDO	☑	⊗	☑	⊗	☑	⊗	98,84
2015, [32]	NB	WBCDO	!	⊗	☑	⊗	?	⊗	95
2015, [33]	WV (DT, MBL, NB, SVM) – FS	BCD	☑	☑	☑	☑	☑	☑	71,64
		WBCDO	☑	☑	☑	☑	☑	☑	97,42
		WBCDD	☑	☑	☑	☑	☑	☑	95,69
		WBCDP	☑	☑	☑	☑	☑	☑	77,24
2015, [34]	BN – DBN	NKI	☑	!	☑	!	☑	⊗	97
		Metabric	☑	!	☑	!	☑	⊗	92
		Ljubljana	☑	!	☑	!	☑	⊗	74
		WBCDO	☑	!	☑	!	☑	⊗	97
		WBCDD	☑	!	☑	!	☑	⊗	97
2016, [4]	SVM	WBCDO	☑	⊗	☑	⊗	☑	⊗	97,13
2016, [35]	J48	?	☑	☑	?	?	☑	?	86,36
2016, [36]	SMO – Ranker	WBCDP	☑	⊗	☑	⊗	☑	☑	77,27
2017, [37]	HPBCR – PSO	Cohort	☑	?	?	☑	?	☑	85
2017, [38]	NB	BCD	?	⊗	☑	⊗	?	⊗	72,70
2017, [39]	NB	BCD	☑	⊗	☑	⊗	?	?	75,17
2017, [40]	SMO	WBCDO	☑	!	☑	⊗	☑	⊗	96,19
2017, [41]	SVM	WBCDO	☑	!	☑	⊗	☑	⊗	97,07
2018, [42]	WAUCE	SEER	☑	!	☑	⊗	☑	☑	76,42
		WBCDO	☑	!	☑	⊗	☑	☑	97,10
		WBCDD	☑	!	☑	⊗	☑	☑	97,68
2018, [43]	C4.5	OMID	☑	☑	☑	?	☑	?	89,29
2018, [44]	FCLF – CNN	WBCDO	☑	!	☑	⊗	☑	?	98,71
		WBCDD	☑	!	☑	⊗	☑	?	99,57
2018, [45]	FCM	WBCDO	?	?	?	?	?	?	97
2018, [46]	NB	WBCDO	☑	?	☑	⊗	⊗	⊗	97,36
2018, [47]	k-NN	WBCDO	☑	⊗	?	⊗	☑	⊗	97,51
2018, [48]	NB – PSO	WBCDP	!	☑	☑	⊗	?	☑	81,3
2018, [49]	NB – BFS	BCD	☑	?	?	⊗	?	☑	82
2018, [50]	GAOGB	WBCDD	☑	☑	?	⊗	?	⊗	94,28
2018, [51]	ANN – Bagging	WBCDO	☑	!	☑	☑	?	☑	96,5
2018, [52]	WCBA	BCD	?	!	?	⊗	?	⊗	0,709
		WBCDO	?	!	?	⊗	?	⊗	0,968
2018, [53]	k-NN – LDA	WBCDD	☑	☑	?	⊗	☑	☑	97,06
2018, [54]	AB – LR – PCA	WBCDD	☑	☑	☑	?	☑	☑	97,92

Keterangan:
☑ : Tepat/baik ditangani atau tidak perlu ditangani
? : Tidak diketahui atau tidak teridentifikasi
! : Kurang tepat/baik, diabaikan, atau dibuang
⊗ : Tidak tepat/baik atau tidak ditangani

Membuang MV bisa menghilangkan informasi yang mungkin penting, mengakibatkan *bias* [55]. Pendekatan imputasi merupakan salah satu strategi yang efisien digunakan untuk menangani MV [15], [55]. salah satu pendekatan imputasi untuk *Missing Value Replacement* (MVR) yang umum digunakan adalah *mean/mode*.

Adanya *outlier* atau anomali pada data dapat menyebabkan ND [33], [15], berdampak buruk terhadap kinerja model [56]. Cara klasik untuk mereduksi ND atau *Smoothing Noisy Data* (SND) yaitu dengan pendekatan *Data Normalization*, seperti *Min-Max Normalization* (MM) maupun *Z-Score* (ZS) [56]. Pada saat yang sama, metode untuk DT dan MV harus mampu memaksimalkan interdependensi antara nilai atribut dan label *class*, untuk meminimalkan hilangnya informasi karena DT maupun MVR [57].

Sementara itu, UC yang memang sering terjadi pada data klinis dapat menurunkan kinerja model [51]. Pendekatan *ensemble* merupakan salah satu cara yang populer untuk menangani UC [58], [59]. Sejalan dengan itu, klasifikasi berbasis *ensemble* terbukti dapat memberikan keputusan klasifikasi yang lebih akurat dan efisien daripada yang diperoleh metode klasifikasi masing-masing [33], [60]. Beberapa metode *ensemble* yang populer adalah *AdaBoost* (AB), *Bagging* (Ba), dan *Weighted Vote* (WV) dapat digunakan untuk *Unbalanced Class Reduction* (UCR).

Selanjutnya, pendekatan FS maupun AW seringkali dapat meningkatkan kinerja model, terlebih pada *dataset* yang berdimensi besar. FS maupun AW dapat menentukan kekuatan hubungan antar atribut atau memberikan bobot pada atribut-atribut berdasarkan relevansinya, bahkan dapat membuang atribut yang tidak relevan [15], [16], [33]. Oleh karena itu, pendekatan FS dan AW mestinya dikaji lebih mendalam dan lengkap lagi untuk prediksi KP menggunakan metode-metode ML. Ada banyak metode yang dapat digunakan untuk FS dan AW, di antaranya yang populer digunakan adalah *Forward Selection* (FSe), *Backward Elimination* (BEl), *Genetic Algorithm* (GA), *Particle Swarm Optimization* (PSO), *Principal Component Analysis* (PCA), dan *Singular Value Decomposition* (SVD).

Untuk DV, metode yang dapat menjamin bahwa semua sampel digunakan untuk pelatihan dan pengujian model, begitupun setiap label *class* dapat terwakili secara merata pada pelatihan model adalah *Leave One Out Cross Validation* dan *K-Fold Cross Validation* [15]. Ketika ukuran sampel besar, *K-Fold Cross Validation* adalah pilihan terbaik [15], [61].

Dengan demikian, model prediksi KP berbasis ML bisa saja memberikan akurasi yang tinggi, namun tanpa disertai pra pengolahan yang efisien, maka tentu saja kinerja model prediksi KP berbasis ML masih dipertanyakan. Oleh karena itu, penelitian ini bertujuan untuk meningkatkan efisiensi (kinerja akurasi) metode-metode ML (*Artificial Neural Network, Support Vector Machine, C4.5, Naive Bayes*, dan *K-Nearest Neighbor*) yang masuk dalam daftar *top ten data mining algorithms* [62] melalui pra-pengolahan (MVR, DT, SND, FS/AW, DV, dan UCR) yang lebih efisien untuk prediksi KP. Dengan begitu, kedepannya model ini dapat digunakan untuk pengembangan sistem cerdas deteksi dini KP sehingga berdampak dalam mereduksi kematian yang disebabkan KP.

Dataset yang digunakan dalam penelitian ini merupakan *dataset* publik yang umum digunakan dalam penelitian terkait dengan prediksi KP, yaitu BCD, BCWDO, BCWDD, dan BCWDP seperti yang ditunjukkan pada Tabel 1. Keempat *dataset* tersebut dikumpulkan dari *UCI Machine Learning Repository* dengan karakternya masing-masing dintunjukkan pada Tabel 2.

Tabel 2. Karakteristik Dataset

Dataset	Type	Attributes	Instances	MV	UC
BCD	Categorical	9	286	9	R=29.72% N=70.28%
WBCDO	Integer	10	699	16	B=65.50% M=34.50%
WBCDD	Real	32	569	0	B=62.74% M=37.26%
WBCDP	Real	34	198	4	R=23.74% N=76.26%

Keterangan:
R : Recurrence
N : Non-Recurrence
B : Benign
M : Malignant

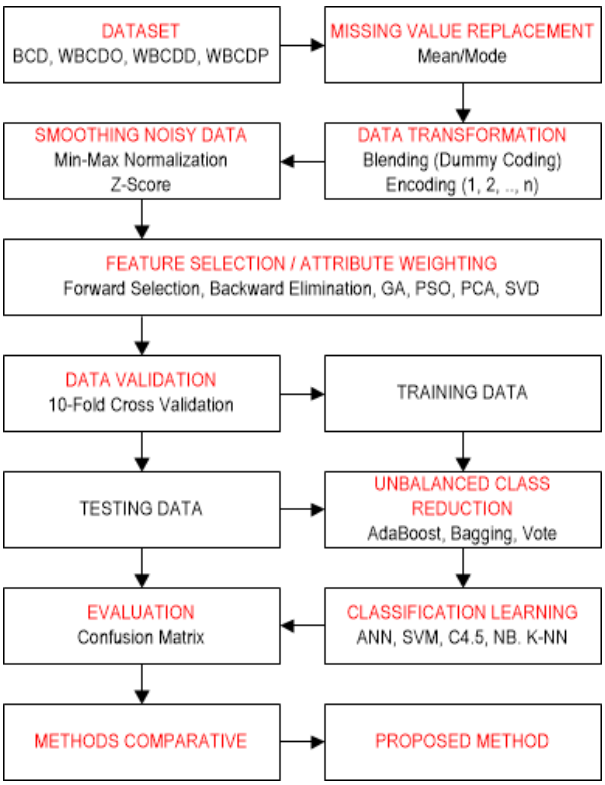
Tipe atribut dari tia-tiap *dataset* menunjukkan bahwa dibutuhkannya DT menyesuaikan karakteristik metode ML yang digunakan. Kecuali pada WBCDD, adanya MV menunjukkan bahwa dibutuhkannya MVR. UC yang terjadi pada BCD dan WBCDP menunjukkan bahwa dibutuhkannya penerapan metode-metode *ensemble*. Begitupun dimensi data yang besar pada WBCDD dan WBCDP menunjukkan bahwa dibutuhkannya FS/AW.

2. Metode Penelitian

Penelitian ini merupakan penelitian eksperimental. Dipandang dari jenis informasi yang diolah, maka penelitian ini merupakan penelitian kuantitatif. Dipandang dari perlakukan terhadap data, maka penelitian ini merupakan penelitian konfirmatori. *Machine Learning* dan klasifikasi dalam *Data Mining* merupakan subjek penelitian ini. Sedangkan objek penelitian ini adalah prediksi KP. Penelitian ini dilaksanakan selama satu tahun lebih, dari Agustus 2018 hingga Oktober 2019.

Prosedur penelitian dimulai dari pengumpulan data menggunakan teknik dokumentasi. Selanjutnya dilakukan MVR menggunakan pendekatan *mean/mode*. Selanjutnya dilakukan DT menggunakan pendekatan

blending (*dummy coding* dan *numeric encoding*) pada *dataset* BCD agar dapat diolah oleh metode-metode ML. Selanjutnya dilakukan SND menggunakan metode MM dan ZS. Selanjutnya dilakukan FS/AW menggunakan metode FSe, BEl, GA, PSO, PCA, dan SVD. Selanjutnya dilakukan DV menggunakan teknik *10-Fold Cross validation*. Selanjutnya penerapan pendekatan *ensemble* untuk menangani UCR menggunakan metode AB, Ba, dan WV. Selanjutnya dilakukan pelatihan metode-metode ML *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), C4.5, *Naive Bayes* (NB), dan *K-Nearest Neighbor* (k-NN). Akhirnya tiap-tiap model/pendekatan yang diuji coba dikomparasi kinerja akurasinya menggunakan teknik *Convusion Matrix*, sehingga diperoleh model yang terbaik (*proposed method*) untuk prediksi KP. Secara keseluruhan, metode yang diusulkan ditunjukkan pada Gambar 1.



Gambar 1. Proposed Method

2.1 Missing Value Replacement (MVR)

Pendekatan yang digunakan untuk MVR, yaitu *mean/mode*. MV dari atribut bertipe numerik diganti dengan nilai *mean* (1). Untuk data pada atribut bertipe integer, nilai mean dibulatkan. Sedangkan MV dari atribut bertipe kategorikal (nominal, binominal, ordinal) diganti dengan nilai *mode*.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (1)$$

2.2 Data Transformation (DT)

DT yang dilakukan adalah mentransformasi data dari atribut bertipe kategorikal (nominal, binominal, dan ordinal) ke tipe numerik. Hal ini dilakukan pada *dataset* BCD agar dapat diolah oleh algoritma ANN dan SVM, serta K-NN walaupun pendekatan *Nominal Measure* yang dimiliki K-NN dapat menangani data kategorikal. Pendekatan yang digunakan untuk DT *categorical to numerical* adalah *Dummy Coding* pada atribut nominal dan *Numeric Encoding* pada atribut ordinal/binominal.

Sedangkan sebaliknya, DT untuk mentransformasi data dari atribut bertipe numerik ke tipe kategorikal tidak dilakukan kerena algoritma ANN, SVM, dan K-NN memang dapat menangani data numerik. Sedangkan algoritma C4.5 memiliki pendekatan *entropy* untuk menangani data numerik. Begitupun algotima NB memiliki pendekatan *kernel* menggunakan metode *G-Reedy*.

2.3 Smoothing Noisy Data (SND)

Pendekatan yang digunakan untuk SND, yaitu MM (2) dan ZS (3).

$$x'_i = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}((new_{max} - new_{min}) + new_{min}) \quad (2)$$

Di mana x_i adalah data atribut ke-*i*, x_{min} adalah data terkecil dari atribut ke-*i*, x_{max} adalah data terbesar dari atribut ke-*i*, new_{min} adalah jangkauan nilai terkecil ke new_{max} adalah jangkauan nilai terbesar untuk hasil normalisasi data.

$$x'_i = \frac{(x_i - \mu)}{\sigma} \quad (3)$$

Di mana x_i adalah data atribut ke-*i*, μ (1) adalah *mean* dari atribut ke-*i*, dan σ adalah *Standard Deviation* (4)

$$\sigma = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right)^{1/2} \quad (4)$$

2.4 Feature Selection (FS) / Attribute Weighting (AW)

Metode-metode yang digunakan untuk FS/AW, yaitu FSe, BEl, GA, PSO, PCA, dan SVD. FSe, BEl, dan GA merupakan metode yang digunakan untuk FS. PSO merupakan metode yang digunakan untuk AW. Sedangkan PCA dan SVD merupakan metode yang digunakan untuk *Dimensionality Reduction*.

Metode FSe dapat digunakan untuk FS dengan tujuan untuk menemukan fitur-fitur *k* dalam set fitur *F* dengan memaksimalkan fungsi *f* (5).

$$x_j = \underset{x_j \in F}{\operatorname{argmax}} f(x_j, y, F\emptyset) \quad (5)$$

Algoritma Forward Selection (FSe) [63]

Input: feature set F, an objective function f, k features to select, and initialize an empty set f∈

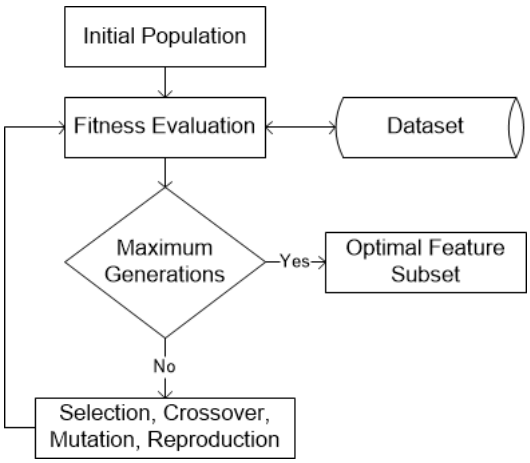
1. Maximize the *objective function* (5);
2. Update relevant feature set such that $F\emptyset \leftarrow F\emptyset \cup x_j$;
3. Remove relevant feature from the original set $F \leftarrow F \backslash x_j$;
4. Repeat until $|F\emptyset| = k$;

Sama seperti FSe, Metode BEl dapat pula digunakan untuk FS. Algoritma BEl adalah sebagai berikut:

Algoritma Backward Elimination (BEl) [64]

Input: $p^0 = \emptyset, p^1 = \{all\ features\}, R = \emptyset, i = 1$
Output: Best Subset p^i
begin
  while $p^i \neq p^{i-1}$ *do*
    for each $v \in p^i$ *do*
      set $p' \leftarrow p^i \backslash v$;
      train the SVM with p' *and*
      get the validation performance $F(p')$;
      if $F(p') \geq F(p^i)$ *then*
        $R \leftarrow R \cup \{v\}$;
      end
    end
    $p^{i+1} \leftarrow p^i \backslash R$;
    $i++$;
    $R = \emptyset$;
  end
  return p^i;
end

Begitupun metode GA, dapat pula digunakan untuk FS. Algoritma GA ditunjukkan pada Gambar 2 [65].



Gambar 2. Genetic Algorithm

Metode PSO memiliki kemampuan untuk mengoptimasi fungsi nonlinier yang dapat digunakan untuk AW.

Algoritma PSO [66]

repeat
  for each particle
    evaluate Objective Function for each particle;
  for each particle
    update best solution;
  update best global solution;
  for each particle
    update the velocity;
    compute the new locations of the articles;
until finished()

Algoritma PCA (6) melakukan *Dimensionality Reduction* dengan memanfaatkan teknik dalam aljabar linier, di mana setiap baris matriks *P* adalah *eigenvector* C_x (8) [67].

$$C_Y = PC_XP^T \tag{6}$$

PCA memerlukan masukan data yang mempunyai sifat *new zero-mean* (7) pada setiap atributnya [67].

$$x_{ij} = x_{ij} - \overline{x_j} \tag{7}$$

Selanjutnya dilakukan perhitungan matriks kovarian (8), di mana X^T adalah matriks transpos dari data *X*. Formula yang digunakan adalah *dot-product* pada setiap atribut [67].

$$C_x = \frac{1}{M}X^TX \tag{8}$$

Nilai *eigenvalue* dan *eigenvector* dari matriks data *X* berturut-turut adalah nilai skala λ dan vektor *u* yang memenuhi Persamaan (9) [67].

$$Xu = \lambda u \tag{9}$$

Dengan mencari matriks ortonormal *P*, di mana *Y* = *PX* dan $C_Y = \frac{1}{M}YY^T$ adalah matriks diagonal, dan kolom dari *P* adalah *principal component* dari *X*, maka Persamaan C_Y bisa dijabarkan sebagai berikut [67].

$$\begin{aligned} C_Y &= \frac{1}{M}YY^T \\ &= \frac{1}{M}(PX)(PX)^T \\ &= \frac{1}{M}PXX^TP^T \\ &= P\left(\frac{1}{M}XX^T\right)P^T \end{aligned} \tag{10}$$

Metode untuk *dimensionality reduction* lainnya yang mirip dengan PCA adalah SVD. Jika PCA menggunakan *eigenvalue* dan *eigenvector* untuk memperoleh solusi, SVD menggunakan dekomposisi nilai tunggal untuk memperoleh solusi, di mana matriks data *X* dapat dibentuk melalui Persamaan (11) [67].

$$A = \sum_{i=1}^{rank(A)} \sigma_i U_i V_i^T \tag{11}$$

Di mana σ_i adalah nilai *singular* ke-*i* dari *A* (nilai ke-*i* pada diagonal Σ), U_i adalah vektor *singular* kiri dari A (kolom ke-*i* dari *U*), dan V_i adalah vektor singular kanan ke-*i* dari A (kolom ke-*i* dari *V*).

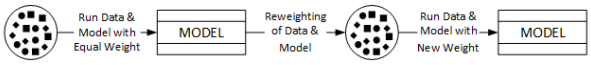
2.5 Data Validation (DV)

Teknik yang digunakan untuk DV, yaitu *K-Fold Cross Validation* dengan nilai K=10. Data dibagi menjadi 10 kelompok yang terwakili secara merata oleh label *class*. Pelatihan dan pengujian/evaluasi model dilakukan disetiap kelompok data (sebanyak 10 iterasi), di mana pada setiap iterasi terdapat 10% data uji dan 90% data latih yang digunakan secara bergantian sehingga

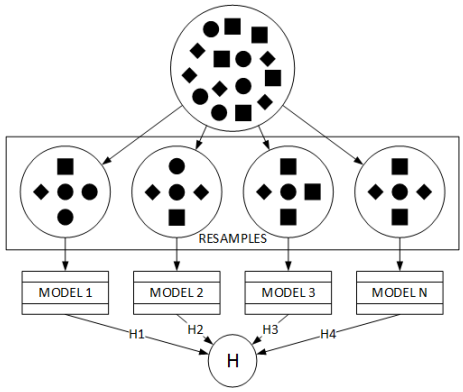
menjamin bahwa setiap sampel telah digunakan pada pelatihan dan pengujian model.

2.6 Unbalanced Class Reduction (UCR)

Pendekatan yang digunakan untuk menangani masalah UC, yaitu pendekatan *ensemble* menggunakan metode AB, Ba, dan WV.



Gambar 3. Boosting Model



Gambar 4. Bagging Model

Algoritma WV adalah sebagai berikut:

1. Hitung akurasi rata-rata (1), maksimum, dan minimum dari *10-Fold Cross Validation* untuk tiap-tiap metode ML yang digunakan.
2. Normalisasikan nilai akurasi rata-rata, maksimum, dan minimum tersebut menggunakan *Min-Max Normalization* (2) dengan *range* [0.1, 1] untuk tiap-tiap metode ML yang digunakan.
3. Hitung selisih antara m_i dengan 1 (12), di mana m_i adalah hasil normalisasi metode ML ke-*i = 1, 2, …, j*.

$$m_i = 1 - m_i \tag{12}$$

4. Hitung *weight* (13) untuk tiap-tiap metode ML yang digunakan.

$$w_i = \frac{m_i}{\sum_{i=1}^{j} m_i} \tag{13}$$

5. Dapatkan keputusan klasifikasi (14), di mana $y(x_d)$ adalah hasil keputusan klasifikasi (label *class*) *instance* ke-*d* dari metode WV, y_c adalah label *class* ke-*c*, y_i adalah label *class* prediksi m_i (metode ke-*i = 1, 2, …, j*).

$$y(x_d) = \underset{y_c \epsilon Y}{\operatorname{argmax}} \sum_{i=1}^{j} w_i (if\ y_i = y_c) \tag{14}$$

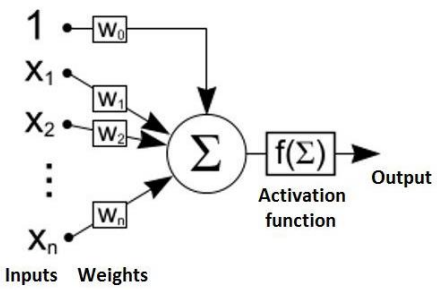
2.7 Machine Learning

Metode-metode ML yang digunakan untuk melakukan klasifikasi (prediksi) KP, yaitu ANN, SVM, C4.5, NB, dan K-NN. Kelima metode tersebut masuk dalam daftar *top ten data mining algorithms* [62].

Metode ANN dapat digunakan untuk estimasi/regresi dan klasifikasi, di mana model matematika *neuron* (15) yang diperkenalkan Mc. Culloch & Pitts didefenisikan sebagai berikut.

$$y = f(\sum_{i=1}^{n} x_i w_i) \tag{15}$$

Di mana signal *x* berupa vektor berdimensi *n* ($x_1, x_2, …, x_n$) akan mengalami penguatan oleh *synapse w* ($w_1, w_2, …, w_n$). Selanjutnya, akumulasi dari penguatan tersebut akan mengalami transformasi oleh fungsi aktifasi *f*. Fungsi *f* ini akan memonitor, bila akumulasi penguatan signal itu telah melebihi batas tertentu. Model matematika neuron (15) tersebut dapat digambarkan dalam bentuk jaringan ANN berikut ini.



Gambar 5. Arsitektur Jaringan ANN [68]

Arsitektur jaringan ANN menunjukkan bahwa suatu jaringan ANN memiliki tiga komponen, yaitu *synapse* ($w_1, w_2, …, w_n$), alat penambah (*adder*), dan fungsi aktifasi (*f*) [68].

Metode SVM dapat digunakan untuk estimasi/regresi dan klasifikasi. Jika *n* merupakan jumlah data yang menjadi *support vector*, dibawah kendala (17), maka formulasi SVM dapat didefenisikan pada Persamaan (16). Pendekatan *kernel* seperti *Anova* dapat diterapkan pada SVM untuk menangani data yang bersifat *nonlinear*.

$$\min_{w,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{16}$$

$$y_i(w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,2,\ldots,n \tag{17}$$

Algoritma C4.5 merupakan varian dari algoritma *Decision Tree* untuk dapat menangani data numerik. *Entropy* diperoleh melalui Persamaan (18), di mana *m* merupakan jumlah nilai yang berbeda pada *class* dan $P(\omega_i|s)$ merupakan proporsi *class* atau nilai fitur ke-*i* yang diproses di *node s*. Sementara *gain* diperoleh melalui Persamaan (19), *spilit info* diperoleh melalui Persamaan (20), dan *gain ratio* diperoleh melalui Persamaan (21).

$$Entropy(s) = -\sum_{i=1}^{m} P(\omega_i|s) Log_2 P(\omega_i|s) \tag{18}$$

$$Gain(s,j) = E(s) - \sum_{i=1}^{n} P(v_i|s)\, E(s_i) \quad (19)$$

$$Split\ Info(s,j) = -\sum_{i=1}^{k} P(v_i|s) Log_2\, P(v_i|s) \quad (20)$$

$$Ratio\ Gain(s,j) = \frac{G(s,j)}{SP(s,j)} \quad (21)$$

Di mana $P(v_i|s)$ merupakan proporsi nilai *v* muncul pada *class* dalam *node*. $E(s_i)$ merupakan *Entropy* komposisi nilai *v* dari *class* ke-*j* dalam data ke-*i* suatu *node*. Selanjutnya *n* merupakan jumlah nilai yang berbeda dalam *node*. Sementara *k* merupakan jumlah pemecahan.

Algoritma NB (22) dapat pula digunakan untuk klasifikasi, di mana $P(C_k)$ adalah probabilitas dari *class ke-k = 1, 2, ..., l*. $P(X_{ij}|C_k)$ adalah probabilitas atribut *ke-j = 1, 2, ..., m*. $C(X_i)$ adalah *class* hasil klasifikasi *instance ke-i = 1, 2, ... n*. Untuk menangani data numerik, maka distribusi *Gaussian* atau pendekatan *kernel* (*G-Reedy*) dapat diterapkan.

$$C(X_i) = \underset{C_k \in Y}{\operatorname{argmax}}\, P(C_k) \prod_{j=1}^{m} P(X_{ij}|C_k) \quad (22)$$

Algoritma k-NN dapat pula digunakan untuk klasifikasi. Formulasi K-NN dapat didefenisikan pada Persamaan (23), di mana perhitungan jarak umumnya menggunakan *Euclidean* (24).

$$y' = \underset{v}{argmax} \sum_{i=1}^{n} x_i y_i \in D_z I(v = y_i) \quad (23)$$

$$d_{ij} = \sqrt{\sum_{j=1}^{m} (X_{ij} - X'_{ij})^2} \quad (24)$$

2.8 Model Evaluation

Pengukuran kinerja akurasi suatu model klasifikasi dapat dilakukan menggunakan *Confussion Matrix*, ditunjukkan pada Tabel 3.

Tabel 3. Confusion Matrix

	Actual +	Actual -	Precision
Predicted +	TP	FP	TP/(TP+FP) *
Predicted -	FN	TN	TN/(TN+FN)
Recall	TP/(TP+FN)	TN/(TN+FP)	
F-Measure	(2*Precision*Sensitivity)/(Precision+Sensitivity)		
Accuracy	(TP+TN) / (TP+TN+FN+FP)		

Keterangan: T (True), F (False), P (Positive); N (Negative)

3. Hasil dan Pembahasan

Tahap pertama yang dilakukan adalah MVR menggunakan pendekatan *mean/mode*, ditunjukkan pada Tabel 4.

Tahap selanjutnya adalah DT yang dilakukan pada *dataset* BCD, ditunjukkan pada Tabel 5. Sedangkan untuk *dataset* lainnya menggunakan pendekatan pada metode ML masing-masing.

Tabel 4. Missing Value Replacement

Dataset	Type	MVR
BCD	Nominal, Binominal, & Ordinal	Mode
WBCDO	Integer	Integer(Mean)
WBCDD	Real	No MV
WBCDP	Real	Mean

Tabel 5. Data Transformation for BCD

Attribute	Type	DT
X1	Ordinal	Encoding [1, 2, 3, 4, 5, 6]
X2	Nominal	Dummy Coding (X2a, X2b, X2c)
X3	Ordinal	Encoding [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]
X4	Ordinal	Encoding [1, 2, 3, 4, 5, 6, 7]
X5	Binominal	Encoding [0, 1]
X6	Ordinal	[1, 2, 3] (tidak perlu)
X7	Binominal	Encoding [0, 1]
X8	Nominal	Dummy Coding (X8a, X8b, X8c, X8d, X8e)
X9	Binominal	Encoding [0, 1]

Hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR dan DT menunjukkan bahwa:

1. Pada *dataset* BCD, C4.5 memberikan kinerja yang terbaik dengan akurasi 73,78%;
2. Pada *dataset* WBCDO, NB memberikan kinerja yang terbaik dengan akurasi 97,42%;
3. Pada *dataset* WBCDD, ANN memberikan kinerja yang terbaik dengan akurasi 97,36%; dan
4. Pada *dataset* WBCDP, ANN memberikan kinerja yang terbaik dengan akurasi 81,82%.

Secara rinci, hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR dan DT ditunjukkan pada Tabel 6.

Tabel 6. Hasil Evaluasi setelah MVR + DT (Normal)

Dataset	Method	DT	Parameters	Accuracy
BCD	ANN	Encoding	hd=1; neurons=15	72,38%
	SVM	Encoding	kernel = anova	73,08%
	C4.5	Entropy	accuracy	73,78%
	NB	Greedy	bandwidth = 0.2	70,98%
	11-NN	Encoding	Correlation	71,68%
WBCDO	ANN	-	hd=1; neurons=9	96,71%
	SVM	-	kernel = anova	97,00%
	C4.5	Entopy	accuracy	94,99%
	NB	Greedy	bandwidth = 0.4	97,42%
	7-NN	-	Euclidean	97,28%
WBCDD	ANN	-	hd=1; neurons=10	97,36%
	SVM	-	kernel = anova	96,66%
	C4.5	Entopy	gain ratio	93,50%
	NB	Greedy	bandwidth = 0.01	94,71%
	7-NN	-	Camberra	95,96%
WBCDP	ANN	-	hd=1; neurons=10	81,82%
	SVM	-	kernel = anova	78,28%
	C4.5	Entopy	accuracy	73,23%
	NB	Greedy	bandwidth = 0.01	77,27%
	11-NN	-	Camberra	77,27%

Setelah MVR dan DT, tahap selanjutnya adalah SND menggunakan pendekatan ZS dan MM. Hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR, DT, dan SND menunjukkan bahwa:

1. Pada *dataset* BCD, SVM – ZS memberikan kinerja yang terbaik dengan akurasi 74,13%, lebih baik 0,35% daripada C4.5 sebagai metode yang terbaik sebelum dilakukan SND;
2. Pada *dataset* WBCDO, NB masih tetap memberikan kinerja yang terbaik dengan akurasi 97,42% sebelum dilakukan SND, walaupun pada metode-metode ML lainnya menunjukkan adanya peningkatan akurasi setelah dilakukan SND, namun tidak lebih baik daripada NB sebelum dilakukan SND;
3. Pada *dataset* WBCDD, ANN masih tetap memberikan kinerja yang terbaik dengan akurasi 97,36%, baik sebelum dilakukan SND maupun sesudahnya; dan
4. Pada *dataset* WBCDP, ANN masih tetap memberikan kinerja yang terbaik dengan akurasi 81,82%, baik sebelum dilakukan SND maupun sesudahnya

Secara rinci, hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR, DT, dan SND ditunjukkan pada Tabel 7 berikut ini.

Tabel 7. Hasil Evaluasi Setelah MVR + DT + SND

Dataset	Method	Normal	ZS	MM
BCD	ANN	72,38%	72,38%	72,38%
	SVM	73,08%	74,13%	73,43%
	C4.5	73,78%	73,78%	73,78%
	NB	70,98%	67,13%	70,63%
	11-NN	71,68%	71,68%	73,78%
WBCDO	ANN	96,71%	96,71%	96,71%
	SVM	97,00%	97,14%	96,28%
	C4.5	94,99%	94,99%	94,99%
	NB	97,42%	96,85%	94,56%
	7-NN	97,28%	96,71%	97,28%
WBCDD	ANN	97,36%	97,36%	97,36%
	SVM	96,66%	96,66%	96,84%
	C4.5	93,50%	93,50%	93,50%
	NB	94,71%	94,19%	94,73%
	7-NN	95,96%	94,20%	95,96%
WBCDP	ANN	81,82%	81,82%	81,82%
	SVM	78,28%	76,77%	77,27%
	C4.5	73,23%	73,23%	73,23%
	NB	77,27%	74,62%	75,25%
	11-NN	77,27%	76,26%	77,27%

Setelah MVR, DT, dan SND, tahap selanjutnya adalah FS/AW menggunakan metode FSe, BEl, GA, PSO, PCA, dan SVD. Hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR, DT, SND, dan FS/AW menunjukkan bahwa:

1. Pada *dataset* BCD, C4.5 – ZS/MM – GA memberikan kinerja yang terbaik dengan akurasi 77,27%, lebih baik 3,14% daripada SVM – ZS sebagai metode yang terbaik sebelum dilakukan FS/AW;
2. Pada *dataset* WBCDO, 7-NN – MM – PSO memberikan kinerja yang terbaik dengan akurasi 97,85%, lebih baik 0,43% daripada NB sebagai metode yang terbaik sebelum dilakukan FS/AW;
3. Pada *dataset* WBCDD, ANN – ZS/MM – FSe memberikan kinerja yang terbaik dengan akurasi 98,24%, lebih baik 0,88% daripada ANN sebagai metode yang terbaik sebelum dilakukan FS/AW; dan
4. Pada *dataset* WBCDP, 11-NN – MM – PSO memberikan kinerja yang terbaik dengan akurasi 83,33%, lebih baik 1,51% daripada ANN sebagai metode yang terbaik sebelum dilakukan FS/AW.

Secara rinci, hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR, DT, SND, dan FS/AW ditunjukkan pada Tabel 8.

Tabel 8. Hasil Evaluasi Setelah MVR + DT + SND + FS/AW

Dataset	Method	Normal	FSe	BEl	GA	PSO	PCA	SVD
BCD	ANN	72,38	75,87	73,78	76,57	74,13	74,13	71,33
	SVM	73,08	75,87	76,22	76,57	76,57	71,33	70,98
	C4.5	73,78	75,87	74,48	77,27	75,87	70,28	68,18
	NB	70,98	75,87	74,83	75,52	76,22	73,43	70,28
	11-NN	71,68	75,17	75,17	76,57	76,22	73,43	69,23
WBCDO	ANN	96,71	96,85	97,00	97,57	97,28	96,57	96,85
	SVM	97,00	96,85	97,28	97,42	97,71	96,57	96,85
	C4.5	94,99	95,14	95,71	96,14	96,14	97,14	97,14
	NB	97,42	96,85	96,85	97,00	96,85	96,14	65,52
	7-NN	97,28	96,85	97,42	97,71	97,85	97,42	94,13
WBCDD	ANN	97,36	98,24	97,89	98,07	97,89	97,19	96,66
	SVM	96,66	97,54	97,54	97,36	97,36	96,49	96,31
	C4.5	93,50	94,38	94,38	96,13	96,13	92,97	91,74
	NB	94,71	94,55	95,43	96,31	96,31	92,27	89,28
	7-NN	95,96	97,36	97,19	97,72	97,72	91,92	81,37
WBCDP	ANN	81,82	82,32	80,81	82,83	82,32	78,28	80,30
	SVM	78,28	80,81	78,28	82,32	80,81	81,82	77,27
	C4.5	73,23	77,78	76,26	78,79	76,77	72,22	67,17
	NB	77,27	77,78	76,26	79,29	77,78	76,77	69,19
	11-NN	77,27	80,81	81,31	82,83	83,33	79,29	78,79

Setelah MVR, DT, SND, dan FS/AW, tahap selanjutnya adalah UC/*ensemble* menggunakan metode WV, Ba, dan AB. Hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR, DT, SND, FS/AW, dan UCR menunjukkan bahwa:

1. Pada *dataset* BCD, 11-NN – MM – GA – WV memberikan kinerja yang terbaik dengan akurasi 77,27%, sama baiknya dengan C4.5 – ZS/MM – GA sebagai metode yang terbaik sebelum dilakukan UCR;
2. Pada *dataset* WBCDO, 7-NN – MM – PSO – AB memerikan kinerja yang terbaik dengan akurasi 97,85%, sama baiknya dengan 7-NN – MM – PSO sebagai metode yang terbaik sebelum dilakukan UCR;
3. Pada *dataset* WBCDD, ANN – ZS/MM – FSe – AB memberikan kinerja yang terbaik dengan akurasi 98,07%, namun tidak lebih baik daripada ANN – ZS/MM – Fse dengan akurasi 98,24% sebagai metode yang terbaik sebelum dilakukan UCR, selisih -0,17%; dan
4. Pada *dataset* WBCDP, 11-NN – MM – PSO – AB memberikan kinerja yang terbaik dengan akurasi 83,33%, sama baiknya dengan 11-NN – MM – PSO sebagai metode yang terbaik sebelum dilakukan UCR.

Dengan demikian, penerapan metode-metode *ensemble* (WV, Ba, AB) untuk UCR ternyata tidak dapat memberikan peningkatan kinerja akurasi pada metode-metode ML, padahal dengan kompleksitas komputasi terhadap waktu yang cukup besar. Hal ini karena UCR dilakukan setelah FS/AW, di mana metode-metode FS/AW mampu mengoptimalkan kinerja metode-metode ML walaupun terjadi UC. Sedangkan UCR dilakukan setelah model dari metode-metode ML sudah optimal karena FS/AW. UCR tidak bisa dilakukan lebih dahulu daripada FS/AW. Seandainya UCR dilakukan tanpa FS/AW, maka hasilnya tidak akan jauh berbeda dengan FS/AW dalam mengoptimalkan kinerja metode-metode ML. Dengan demikian, sampai pada tahap FS/AW, metode-metode ML untuk prediksi KP dianggap sudah optimal kinerjanya. Secara rinci, hasil evaluasi kinerja akurasi metode-metode ML setelah dilakukan MVR, DT, SND, FS/AW, dan UCR ditunjukkan pada Tabel 9 berikut ini.

Tabel 9. Hasil Evaluasi Setelah MVR+DT+SND+FS/AW+UC

Dataset	Method	Normal	WV	Ba	AB
BCD	ANN	72,38%	76,57%	75,87%	75,87%
	SVM	73,08%	75,52%	76,22%	76,92%
	C4.5	73,78%	75,52%	75,87%	75,87%
	NB	70,98%	76,22%	75,52%	75,87%
	11-NN	71,68%	77,27%	76,22%	76,57%
WBCDO	ANN	96,71%	97,14%	97,42%	97,14%
	SVM	97,00%	97,14%	97,28%	97,28%
	C4.5	94,99%	96,85%	97,14%	95,14%
	NB	97,42%	97,14%	96,85%	96,71%
	7-NN	97,28%	97,42%	97,57%	97,85%
WBCDD	ANN	97,36%	97,19%	97,72%	98,07%
	SVM	96,66%	97,72%	97,36%	97,36%
	C4.5	93,50%	97,72%	96,49%	97,72%
	NB	94,71%	97,89%	96,13%	96,49%
	7-NN	95,96%	97,72%	97,54%	97,72%
WBCDP	ANN	81,82%	80,81%	82,83%	82,83%
	SVM	78,28%	80,30%	79,80%	80,81%
	C4.5	73,23%	79,80%	77,27%	81,31%
	NB	77,27%	82,32%	79,29%	79,80%
	11-NN	77,27%	80,81%	81,82%	83,33%

Setelah tiap-tiap model dievaluasi, maka pendekatan terbaik untuk tiap-tiap metode ML dapat diperoleh, ditunjukkan pada Tabel 10 berikut ini.

Tabel 10. Pendekatan Terbaik untuk Setiap Metode ML

Dataset	Method	Normal	SND		FS/AW		UC	
BCD	ANN	72,38	ZS	72,38	GA	76,57	WV	76,57
	SVM	73,08	ZS	74,13	GA	76,57	AB	76,92
	C4.5	73,78	ZS	73,78	GA	77,27	Ba	75,87
	NB	70,98	MM	70,63	PSO	76,22	WV	76,22
	11-NN	71,68	MM	73,78	GA	76,57	WV	77,27
WBCDO	ANN	96,71	ZS	96,71	GA	97,57	Ba	97,42
	SVM	97,00	ZS	97,14	PSO	97,71	Ba	97,28
	C4.5	94,99	ZS	94,99	SVD	97,14	Ba	97,14
	NB	97,42	ZS	96,85	GA	97,00	WV	97,14
	7-NN	97,28	MM	97,28	PSO	97,85	AB	97,85
WBCDD	ANN	97,36	ZS	97,36	FSe	98,24	AB	98,07
	SVM	96,66	MM	96,84	FSe	97,54	WV	97,72
	C4.5	93,50	ZS	93,50	GA	96,13	AB	97,72
	NB	94,71	MM	94,73	GA	96,31	WV	97,89
	7-NN	95,96	MM	95,96	GA	97,72	AB	97,72
WBCDP	ANN	81,82	ZS	81,82	GA	82,83	Ba	82,83
	SVM	78,28	MM	77,27	GA	82,32	AB	80,81
	C4.5	73,23	ZS	73,23	GA	78,79	AB	81,31
	NB	77,27	MM	75,25	GA	79,29	WV	82,32
	11-NN	77,27	MM	77,27	PSO	83,33	AB	83,33

Dengan demikian, pendekatan yang diusulkan untuk prediksi KP berbasis ML, antara lain:

1. Pada *dataset* BCD, C4.5 – ZS/MM – GA memberikan kinerja yang terbaik dan sama baiknya dengan 11-NN – MM – GA – WV dengan akurasi 77,27%, namun karena mempertimbangkan kompleksitas komputasi dari proses UCR, maka C4.5 – ZS/MM – GA dianggap yang terbaik dan lebih efisien untuk prediksi KP karena tidak perlu hingga melakukan UCR;
2. Begitupun pada *dataset* WBCDO, 7-NN – MM – PSO memerikan kinerja yang terbaik dan sama baiknya dengan 7-NN – MM – PSO – AB dengan akurasi 97,85%, sehingga 7-NN – MM – PSO dianggap yang terbaik dan lebih efisien untuk prediksi KP;
3. Pada *dataset* WBCDD, ANN – ZS/MM – FSe dengan akurasi 98,24% sebagai metode yang terbaik tanpa harus melakukan UCR; dan
4. Pada *dataset* WBCDP, 11-NN – MM – PSO memberikan kinerja yang terbaik dan sama baiknya dengan 11-NN – MM – PSO – AB dengan akurasi 83,33%, sehingga 11-NN – MM – PSO dianggap yang terbaik dan lebih efisien untuk prediksi KP.

Secara rinci, pendekatan yang diusulkan untuk prediksi KP berbasis ML ditunjukkan pada Tabel 11 berikut ini.

Tabel 11. Proposed Method

DS	Method	Proposed Approach	Accuracy
BCD	C4.5	Parameter = Accuracy DT = Entropy SND = Z-Score (ZS) FS = GA	77,27%
WBCDO	7-NN	Parameter = Euclidean DT = - SND = Min-Max (MM) AW = PSO	97,85%
WBCDD	ANN	Parameter = 1 hd, 10 neurons hd DT = - SND = Z-Score (ZS) FS = Forward Selection (FSe)	98,24%
WBCDP	11-NN	Parameter = Camberra DT = - SND = Min-Max (MM) AW = PSO.	83,33%

Kinerja yang diperoleh dari metode yang diusulkan pada tiap-tiap *dataset* menunjukkan bahwa adanya peningkatan akurasi dari pada tiap-tiap metode ML standar yang digunakan. Hal ini membuktikan bahwa melalui pra-pengolahan yang lebih efisien, maka kinerja metode-metode ML dapat ditingkatkan. Secara rinci, komparasi antara metode yang diusulkan dengan metode-metode ML standar (normal) dan metode-metode ML dari penelitian terkait sebelumnya ditunjukkan pada Tabel 12, di mana simbol ☑

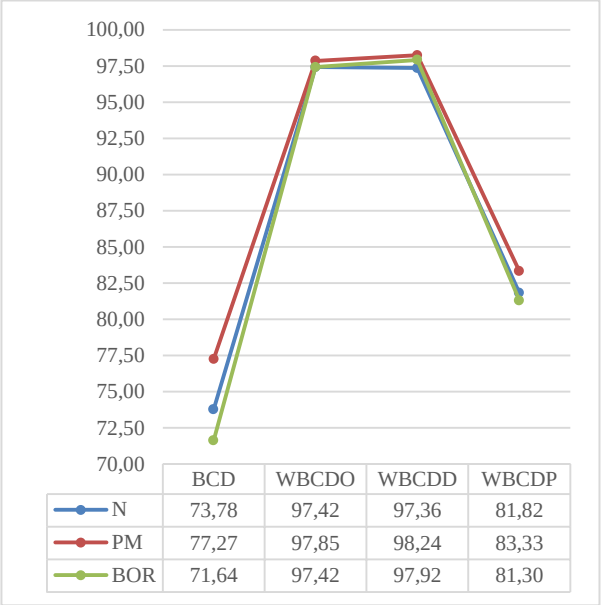
menunjukkan penelitian kami, sedangkan simbol ✔ menunjukkan penelitian terkait sebelumnya yang kami anggap terbaik.

Merujuk penelitian-penelitian terkait pada Tabel 1, kami tidak dapat membandingkan hasil penelitian ini dengan penelitian terkait sebelumnya yang dianggap tidak melakukan pra-pengolaahan dengan efisien, misalnya mengabaikan atau membuang adanya MV bisa jadi memperoleh akurasi yang lebih tinggi karena jumlah data yang berkurang, tapi bisa jadi pula mengabaikan informasi yang mungkin penting, mengakibatkan bias. Begitupun apabila DT, SND, FS/AW, bahkan DV kurang diperhatikan. Penelitian seperti itu tentu saja tidak dapat dianggap yang terbaik walaupun menunjukkan akurasi yang tertinggi. Tetapi kami tetap memasukkannya dalam daftar perbandingan. Seperti yang dikemukakan oleh Abreu *et al*., 2016 dalam *review* yang mereka lakukan terhadap prediksi KP berbasis ML, bahwa penelitian-penelitian terkait prediksi KP berbasis ML memang telah memperoleh akurasi yang tinggi, namun secara keseluruhan, pra-pengolahan belum dilakukan secara optimal/efisien, efisiensi maupun sensitifitasnya masih diragukan [15].

Tabel 12. Komparasi Metode

Dataset	Year, Ref	Proposed Method	Accuracy	Des
BCD	2019, Our	Best of ML Normal: C4.5	73,78%	☑
		Proposed: C4.5-ZS-GA	77,27%	☑
	2015, [33]	WV (DT, MBL, NB, SVM) - FS	71,64%	✔
	2017, [38]	NB	72,70%	
	2017, [39]	NB	75,17%	
	2018, [49]	NB – BFS	82,00%	
	2018, [52]	WCBA	70,93%	
WBCDO	2019, Our	Best of ML Normal: NB	97,42%	☑
		Proposed: 7NN-MM-PSO	97,85%	☑
	2007, [6]	LSSVM	98,53%	
	2007, [19]	PSO – MSS	100%	
	2009, [2]	SVM – F-Score	99,51%	
	2011, [3]	SVM – RS	100%	
	2014, [31]	FMM–CART–RF	98,84%	
	2015, [32]	NB	95,00%	
	2015, [33]	WV (DT, MBL, NB, SVM) - FS	97,42%	✔
	2015, [34]	BN – DBN	97,00%	
	2016, [4]	SVM	97,13%	
	2017, [40]	SMO	96,19%	
	2017, [41]	SVM	97,07%	
	2018, [42]	WAUCE	97,10%	
	2018, [50]	FCLF – CNN	98,71%	
	2018, [45]	FCM	97,00%	
	2018, [46]	NB	97,36%	
	2018, [47]	K-NN	97,51%	
	2018, [51]	ANN – Bagging	96,50%	
	2018, [52]	WCBA	96,80%	
WBCDD	2019, Our	Best of ML Normal: ANN	97,36%	☑
		Proposed: ANN-ZS-FSe	98,24%	☑
	2012, [22]	DT - RS – GA	95,30%	
	2015, [33]	WV (DT, MBL, NB, SVM) - FS	95,69%	
	2015, [34]	BN – DBN	97,00%	
	2018, [42]	WAUCE	97,68%	
	2018, [44]	FCLF – CNN	99,57%	
	2018, [50]	GAOGB	94,28%	
	2018, [53]	K-NN – LDA	97,06%	
	2018, [54]	AB – LR – PCA	97,92%	✔
WBCDP	2019, Our	Best of ML Normal: ANN	81,82%	☑
		Proposed: 11NN-MM-PSO	83,33%	☑
	2014, [30]	K-NN	90,00%	
	2015, [33]	WV (DT, MBL, NB, SVM) - FS	77,24%	
	2016, [36]	SMO – Ranker	77,27%	
	2018, [48]	NB – PSO	81,30%	✔

Dengan demikian, komparasi antara metode yang diusulkan dengan metode-metode ML standar (normal) dan metode ML dari penelitian terkait sebelumnya yang terbaik, dapat ditunjukkan pada Gambar 6.



Keterangan:
N : *Best of ML Normal (MVR + DT)*
PM : *Proposed Method*
BOR : *Best of Related Research*

Gambar 6. ML Normal vs Proposed Method vs Other Research

4. Kesimpulan

Berdasarkan hasil penelitian yang diperoleh, dapat disimpulkan bahwa melalui pra-pengolahan yang efisien (*Missing Value Replacement*, *Data Transformation*, *Smoothing Noisy Data*, *Feature Selection* atau *Attribute Weighting*, *Data Validation*, dan *Unbalanced Class Reduction*) untuk prediksi Kanker Payudara berbasis *machine learning*, maka kinerja akurasi dari metode-metode *machine learning* dapat ditingkatkan, sehingga diperoleh pendekatan-pendekatan *machine learning* yang efisien untuk prediksi Kanker Payudara, yaitu:

1. C4.5 – ZS – GA untuk *dataset* BCD dengan akurasi 77,27%, lebih baik 3,49% dari C4.5 sebagai *best of ML normal*, dan lebih baik 5,63% dari *Weighted Vote* (*Decision Tree*, *Memory Based Learner*, NB, SVM) – *Fisher Score* sebagai *best of related research*;
2. 7NN – MM – PSO untuk *dataset* WBCDO dengan akurasi 97,85%, lebih baik 0,43% dari NB sebagai *best of ML normal* dan dari *Weighted Vote* (*Decision Tree*, *Memory Based Learner*, NB, SVM) – *Fisher Score* sebagai *best of related research*;

3. ANN – ZS – FSe untuk *dataset* WBCDD dengan akurasi 98,24%, lebih baik 0,88% dari ANN sebagai *best of ML normal*, dan lebih baik 0,32% dari AB – *Logistic Regression* – PCA sebagai *best of related research*; dan
4. 11NN – MM – PSO untuk *dataset* WBCDP dengan akurasi 83,33%, lebih baik 1,51% dari ANN sebagai *best of ML normal*, dan lebih baik 2,03% dari NB – PSO sebagai *best of related research*.

Dengan demikian, pendekatan-pendekatan tersebut menunjukkan kinerja akurasi yang lebih baik dari pada standar/normal ML bahkan lebih baik dari pada penelitian-penelitian terkait, sehingga pendekatan-pendekatan tersebut dapat digunakan untuk pengembangan alat bantu deteksi dini Kanker Payudara berbasis *machine learning* yang diharapkan bisa berdampak terhadap upaya mereduksi tingkat kematian yang disebabkan Kanker Payudara. Namun tentunya melalui riset-riset selanjutnya yang lebih mendalam lagi.

Ucapan Terima Kasih

Penelitian ini didukung dan didanai oleh: (1) Direktorat Riset dan Pengabdian Masyarakat; (2) Kementrian Riset dan Pendidikan Tinggi Republik Indonesia.

Daftar Rujukan

[1] National Breast Cancer Coalition, "BreastCancerDeadline2020," 2017. [Online]. Available: http://www.breastcancerdeadline2020.org/. [Accessed: 18-Sep-2018].

[2] M. F. Akay, "Support Vector Machines Combined with Feature Selection for Breast Cancer Diagnosis," *Expert Syst. Appl.*, vol. 36, pp. 3240–3247, 2009.

[3] H. Chen, B. Yang, J. Liu, and D. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 9014–9022, 2011.

[4] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," in *Procedia Computer Science*, 2016, vol. 83, pp. 1064–1069.

[5] A. K. Dubey, U. Gupta, and S. Jain, "Breast Cancer Statistics and Prediction Methodology: A Systematic Review and Analysis," *Asian Pasific J. Cancer Prev.*, vol. 16, no. 10, pp. 4237–4245, 2015.

[6] K. Polat and S. Guneş, "Breast Cancer Diagnosis Using Least Square Support Vector Machine," *Digit. Signal Process.*, vol. 17, no. 4, pp. 694–701, 2007.

[7] S. Shah, "BreastCancerIndia.net," 2014. [Online]. Available: http://www.breastcancerindia.net/. [Accessed: 18-Sep-2018].

[8] depkes.go.id, "Pemerintah Targetkan 80% Perempuan dapat Deteksi Dini Kanker Payudara dan Kanker Serviks," 2013. [Online]. Available: http://www.depkes.go.id/development/site/jkn/index.php?cid=13100003&id=pemerintah-targetkan-80%25-perempuan-dapat-deteksi-dini-kanker-payudara-dan-kanker-serviks.html. [Accessed: 18-Sep-2018].

[9] International Agency for Research of Cancer, "Global Cancer Observatory," 2018. [Online]. Available: http://gco.iarc.fr/. [Accessed: 18-Sep-2018].

[10] Y. Zhu, L. Zhou, S. Jiao, and L. Xu, "Relationship Between Soy Food Intake and Breast Cancer in China," *Asian Pacific J. Cancer Prev.*, vol. 12, no. 11, pp. 2837–2840, 2011.

[11] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global Cancer Statistics," *CA. Cancer J. Clin.*, vol. 61, no. 2, pp. 69–90, 2011.

[12] G. Carioli, M. Malvezzi, T. Rodriguez, P. Bertuccio, E. Negri, and C. La Vecchia, "Trends and predictions to 2020 in breast cancer mortality in Europe," *The Breast*, vol. 36, pp. 89–95, 2017.

[13] A. J. Vickers, "Prediction Models in Cancer Care," *CA. Cancer J. Clin.*, vol. 61, no. 5, pp. 315–326, 2011.

[14] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: big data for personalized healthcare," in *IEEE Journal of Biomedical and Health Informatics*, 2015, vol. 19, no. 4, pp. 1209–1215.

[15] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting Breast Cancer Recurrence Using Machine Learning Tehniques: A Systematic Review," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 52:1-52:40, 2016.

[16] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[17] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89–109, 2001.

[18] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. 184–190, 2006.

[19] G. S. Tewolde and D. M. Hanna, "Particle swarm optimization for classification of breast cancer data using single and multisurface methods of data separation," in *2007 IEEE International Conference on Electro/Information Technology*, 2007, pp. 443–446.

[20] T. Ayer, O. Alagoz, J. Chhatwal, J. W. Shavlik, C. E. Kahn, and E. S. Burnside, "Breast Cancer Risk Estimation With Artificial Neural Networks Revisited," *Cancer*, pp. 3310–3321, 2010.

[21] D. Soria, J. M. Garibaldi, F. Ambrogi, E. M. Biganzoli, and I. O. Ellis, "A non-parametric version of the naive Bayes classifier," *Knowledge-Based Syst.*, vol. 24, no. 6, pp. 775–784, 2011.

[22] H. I. Elshazly, N. I. Ghali, A. M. El Korany, and A. E. Hassanien, "Rough Sets and Genetic Algorithms: A hybrid approach to breast cancer classification," in *2012 World Congress on Information and Communication Technologies*, 2012, pp. 260–265.

[23] M. Huang, Y. Hung, W. Lee, R. K. Li, and T. Wang, "Usage of Case-Based Reasoning, Neural Network and Adaptive Neuro-Fuzzy Inference System Classification Techniques in Breast Cancer Dataset Classification Diagnosis," *J. Med. Syst.*, vol. 36, no. 2, pp. 407–414, 2012.

[24] W. Kim *et al.*, "Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine," *J. Breast Cancer*, vol. 15, no. 2, pp. 230–238, 2012.

[25] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A Gene Signature for Breast Cancer Prognosis Using Support Vector Machine," in *2012 5th International Conference on BioMedical Engineering and Informatics*, 2012, pp. 928–931.

[26] H. Palivela, K. Patil, Y. H K, and V. S, "Survey On Mining Techniques For Breast Cancer Related Data," in *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, 2013.

[27] K. Park, A. Ali, D. Kim, Y. An, M. Kim, and H. Shin, "Robust predictive model for evaluating breast cancer survivability," *Eng. Appl. Artif. Intell.*, vol. 26, no. 9, pp. 2194–2205, 2013.

[28] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 4, pp. 613–618, 2013.

[29] C. Park, J. Ahn, H. Kim, and S. Park, "Integrative Gene Network Construction to Analyze Cancer Recurrence Using Semi-Supervised Learning," *PLoS One*, vol. 9, no. 1, pp. 1–9, 2014.

[30] A. P. Pawlovsky and M. Nagahashi, "A Method to Select a Good Setting for the kNN Algorithm when Using it for Breast

Cancer Prognosis," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014, pp. 189–192.
[31] M. Seera and C. P. Lim, "A Hybrid Intelligent System for Medical Data Classification," *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, 2014.
[32] G. D. Rashmi, A. Lekha, and N. Bawane, "Analysis of Efficiency of Classification and Prediction Algorithms (Naive Bayes) for Breast Cancer Dataset," in *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT)*, 2015, pp. 108–113.
[33] S. Bashir, U. Qamar, and F. H. Khan, "Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble," *Qual. Quant.*, vol. 49, no. 5, pp. 2061–2076, 2015.
[34] M. Khademi and N. S. Nedialkov, "Probabilistic Graphical Models and Deep Belief Networks for Prognosis of Breast Cancer," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 727–732.
[35] S. Muthuselvan, K. S. Sundaram, and Prabasheela, "Prediction of Breast Cancer Using Classification Rule Mining Techniques in Blood Test Datasets," in *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, 2016.
[36] A. I. Pritom, A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting Breast Cancer Recurrence using Effective Classification and Feature Selection Technique," in *2016 19th International Conference on Computer and Information Technology (ICCIT)*, 2016, pp. 310–314.
[37] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mananas, and F. Mokarian, "A Hybrid Computer-aided-diagnosis System for Prediction of Breast Cancer Recurrence (HPBCR) Using Optimized Ensemble Learning," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 75–85, 2017.
[38] D. Verma and N. Mishra, "Comparative Analysis of Breast Cancer and Hypothyroid Dataset using Data Mining Classification Techniques," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2017, pp. 1624–1626.
[39] D. Verma and N. Mishra, "Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 533–538.
[40] V. Chaurasia and S. Pal, "A Novel Approach for Breast Cancer Detection using Data Mining Techniques," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 2, no. 1, 2017.
[41] D. E. Gbenga, N. Christopher, and D. C. Yetunde, "Performance Comparison of Machine Learning Techniques for Breast Cancer Detection," *Nov. J. Eng. Appl. Sci.*, vol. 6, no. 1, pp. 1–8, 2017.
[42] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis," *Eur. J. Oper. Res.*, vol. 267, no. 2, pp. 687–699, 2018.
[43] M. Vazifehdan, M. H. Moattar, and M. Jalali, "A hybrid Bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction," *J. King Saud Univ. - Comput. Inf. Sci.*, 2018.
[44] K. Liu, G. Kang, N. Zhang, and B. Hou, "Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks," *IEEE Access*, vol. 6, pp. 23722–23732, 2018.
[45] A. K. Dubey, U. Gupta, and S. Jain, "Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 1, p. 18, 2018.
[46] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of Benign and Malignant Breast Cancer using Data Mining Techniques," *J. Algorithm. Comput. Technol.*, vol. 12, no. 2, pp. 119–126, 2018.
[47] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast Cancer Classification Using Machine Learnin," in *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018.
[48] S. Sakri, N. A. Rashid, and Z. M. Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018.
[49] B. Tamilvanan and V. M. Bhaskaran, "An Efficient Classifications Model for Breast Cancer Prediction Based on Dimensionality Reduction Techniques," *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 1, pp. 448–456, 2018.
[50] H. Lu, H. Wang, and S. W. Yoon, "A Dynamic Gradient Boosting Machine Using Genetic Optimizer for Practical Breast Cancer Prognosis," *Expert Syst. Appl.*, 2018.
[51] I. Fakhruzi, "An Artificial Neural Network with Bagging to Address Imbalance Datasets on Clinical Prediction," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, no. 1, pp. 895–898.
[52] J. Alwidian, B. H. Hammo, and N. Obeid, "WCBA : Weighted classification based on association rules algorithm for breast cancer disease," *Appl. Soft Comput. J.*, vol. 62, pp. 536–549, 2018.
[53] A. Joshi and A. Mehta, "Analysis of K-Nearest Neighbor Technique for Breast Cancer Disease Classification," *Int. J. Recent Sci. Res.*, vol. 9, no. I, pp. 26126–26130, 2018.
[54] K. Goyal, P. Sodhi, P. Aggarwal, and M. Kumar, "Comparative Analysis of Machine Learning Algorithms for Breast Cancer Prognosis," in *Proceedings of 2nd International Conference on Communication, Computing and Networking*, 2018, p. 727.734.
[55] G. E. A. P. A. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, 2003.
[56] M. M. Suarez-Alvarez, D.-T. Pham, M. Y. Prostov, and Y. I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," in *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2012, vol. 468, no. 2145, pp. 2630–2651.
[57] S. Kotsiantis and D. Kanellopoulos, "Discretization Techniques : A Recent Survey," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 32, no. 1, pp. 47–58, 2006.
[58] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A Review," *Int. J. Adv. Soft Comput. its Appl.*, vol. 7, no. 3, pp. 176–204, 2015.
[59] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 4, pp. 463–484, 2012.
[60] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, no. 1–2, pp. 1–39, 2010.
[61] M. S. Santos, P. H. Abreu, P. J. Garcia-Laencina, A. Simao, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *J. Biomed. Inform.*, vol. 58, pp. 49–59, 2015.
[62] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, 2008.
[63] J.-L. Bouchota, W. L. Trimble, G. Ditzler, Y. Lan, S. Essinger, and G. Rosen, "Advances in Machine Learning for Processing and Comparison of Metagenomic Data," in *Computational Systems Biology (Second Edition)*, Science Direct, 2014.
[64] L. Xie, Z. Fu, W. Feng, and Y. Luo, "Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news," *Multimed. Syst.*, vol. 17, no. 2, pp. 101–112, 2011.
[65] H. Le and L. Tran, "Automatic feature selection for named entity recognition using genetic algorithm," in *Proceedings of the Fourth Symposium on Information and Communication Technology*, 2013.
[66] Kennedy, Eberhart, and Shi, *Swarm Intelligence*. Morgan Kaufmann division of Academic Press, 2001.
[67] E. Prasetyo, *Data Mining: Konsep dan Aplikasinya Menggunakan Matlab*. Yogyakarta, Indonesia: Andi Offset, 2012.
[68] T. Sutojo, E. Mulyanto, and V. Suhartono, *Kecerdasan Buatan*. Yogyakarta, Indonesia: Andi Offset, 2011.