Jurnal Pendidikan Teknologi dan Kejuruan
Vol. 18, No. 1, Januari 2021
P-ISSN : 0216-3241 E-ISSN : 2541-0652

MENINGKATKAN KINERJA K-NN UNTUK KLASIFIKASI KANKER PAYUDARA DENGAN SELEKSI FITUR

H. Harafani[1)], H. Aji Al-Kautsar[2)]

[1] Program Studi Teknik Informatika, STMIK Nusa Mandiri
[2]Program Studi Teknologi Komputer, Universitas Bina Sarana Informatika
Email : hani.hhf@nusamandiri.ac.id, hanggoro.hgr@bsi.ac.id

ABSTRAK

Kanker payudara adalah kanker paling umum yang menyerang wanita di seluruh dunia. Machine Learning telah banyak digunakan untuk membantu dalam mendukung keputusan para ahli kesehatan dalam memprediksi penyakit kanker payudara . Algoritma K-NN digunakan pada penelitian ini untuk mengklasifikasi dataset kanker payudara Coimbra dan forward selection diimplementasikan untuk menghindari sensitivitas K-NN terhadap attribute yang tidak relevan dan berkorelasi, sehingga kinerja K-NN dapat lebih maksimal. Tujuan penelitian ini adalah untuk membandingkan kinerja dari kombinasi K-NN dan forward selection dengan algoritma machine learning lainnya, sekaligus meningkatkan kinerja K-NN dalam mengklasifikasi dataset kanker payudara. Hasil menunjukan akurasi tertinggi diperoleh dari kombinasi KNN+FS dan Kombinasi K-NN+OS sebesar 91,43% dengan proporsi data 70/30% dan nilai K sebesar 1. Sedangkan Kombinasi K-NN+BE mengalami overfitting. Attribute MCP.1 tidak mempunyai korelasi hampir pada seluruh eksperimen.Implementasi berbagai metode fitur seleksi pada K-NN sangat membantu K-NN dalam meningkatkan performansinya.

Kata kunci: K-NN, Seleksi, Fitur, Klasifikasi, Payudara

ABSTRACT

Breast cancer is the most common cancer affecting women worldwide. Machine Learning has been widely used to support the decisions of health experts in predicting breast cancer. The K-NN algorithm is used in this study to classify the Coimbra breast cancer dataset and forward selection is implemented to avoid K-NN sensitivity to irrelevant and correlated attributes, so that K-NN performance can be maximized. The purpose of this study was to compare the performance of the combination of K-NN and forward selection with other machine learning algorithms, as well as to improve the performance of K-NN in classifying the breast cancer dataset. The results achieved were obtained from the combination of KNN + FS and the combination of K-NN + OS of 91.43% with a data proportion of 70/30% and a K value of 1. While the K-NN+BE combination is experiencing over fitting. The MCP.1 attribute does not have any correlation in almost all experiments. The implementation of various selection feature methods on K-NN really helps K-NN in improving its performance

Keywords : *K-NN, Selection, Feature, Classification, Breast*

1. PENDAHULUAN

Kanker payudara adalah kanker paling umum yang menyerang wanita di seluruh dunia[1]. Kanker payudara merupakan masalah kesehatan yang serius , karena menyumbang lebih dari 1,6% dari total kematian pada wanita di seluruh dunia[2]. Menurut Holford selama beberapa decade terakhir tren yang baik dalam kematian akibat kanker payudara disebabkan oleh kemajuan dalam diagnosis dan pengobatan yang terus bertambah [3]. Oleh karena itu deteksi dini kanker payudara memainkan peran penting dalam perencanaan dan hasil pengobatan terkait[2], dan diagnosa awal dari kanker payudara dan metastasis pada pasien berdasarkan system yang akurat dapat meningkatkan kelangsungan hidup pasien sampai lebih dari 86%[4].

Industri perawatan kesehatan sangat intensif terhadap data dan membutuhkan platform data besar yang interaktif dan dinamis dengan teknologi dan alat inovatif untuk memajukan perawatan dan

layanan pasien [5], sehingga model prediksi data untuk tingkat kelangsungan hidup pasien kanker dapat membantu dalam prognosis dan pengelolaan kanker[1]. Machine Learning telah banyak digunakan untuk membantu dalam mendukung keputusan para ahli kesehatan dalam memprediksi penyakit kanker payudara seperti yang dilakukan oleh [4], [6], [7]. Beberapa metode banyak digunakan oleh para peneliti dunia untuk mengklasifikasikan dataset kanker payudara diantaranya Neural Network[8] dengan accuracy 84,55%, Support Vector Machine[9] dengan accuracy 85,38%, dan K-NN[2] dengan accuracy 92,10%.

Pada penelitian ini Dataset kanker payudara Coimbra digunakan sejak dataset ini menjadi semakin populer untuk diteliti[10]–[12] sekaligus melanjutkan penelitian sebelumnya[9]. Dataset ini digunakan sebagai data latih dan data uji. K-NN akan diadaptasi untuk mengklasifikasi dataset kanker payudara Coimbra karena K-NN dapat menghasilkan akurasi yang baik pada data yang berjumlah banyak[13]. Walaupun K-NN merupakan algoritma yang lugas dan kuat[14], namun K-NN memiliki beberapa kekurangan, salah satu diantaranya adalah sensitive terhadap attribute yang tidak relevan dan berkorelasi[15]. Terlebih lagi data kanker payudara mempunyai banyak fitur yang berisi informasi mengenai kanker payudara, tetapi tidak semua fitur merupakan fitur yang relevan[16].

Kehadiran seleksi fitur turut membantu machine learning dalam meningkatkan performanya [2], algoritma seleksi fitur yang banyak digunakan untuk kasus klasifikasi diantaranya forward selection[17], [18], [19], Backward Selection[20],[21], dan Optimize Selection[22]. Pada penelitian ini beberapa metode seleksi fitur akan diterapkan diantaranya forward selection , backward elimination, dan Optimize selection akan diimplementasikan untuk memilih fitur yang relevan, sehingga kinerja K-NN dapat lebih maksimal. Tujuan penelitian ini adalah untuk membandingkan kinerja K-NN dari kombinasi K-NN dengan berbagai metode seleksi fitur. sekaligus meningkatkan kinerja K-NN dalam mengklasifikasi dataset kanker payudara.

Penggunaan seleksi fitur untuk meningkatkan kinerja klasifikasi dari machine learning telah banyak digunakan oleh para peneliti, seperti yang dilakukan oleh Wahyuni [7] yang menerapkan metode seleksi fitur untuk meningkatkan hasil diagnosis kanker payudara. Dataset kanker payudara yang digunakan adalah dataset *Wisconsin Breast Cancer Database* (WBCD). Metode seleksi fitur yang digunakan adalah F-Score, dan algoritma klasifikasi yang digunakan yaitu SMO, Naïve Bayes, Multilayer Perceptron, dan C4.5. Metode evaluasi yang digunakan adalah 10 *fold cross validation.* Hasil penelitian menunjukkan akurasi tertinggi didapat oleh kombinasi F-score dengan Naïve Bayes yaitu 97,65% dengan jumlah attribute yang digunakan 6 dari 9 attribut. Sedangkan tes ujibeda juga menunjukkan penggunaan F-Score pada daive bayes memiliki perbedaan yang signifikan.

Astuti [17] juga menggunakan forward selection pada algoritma naïve bayes untuk mengklasifikasi benih gandum. Dataset yang digunakan adalah data benih gandum publik yang diambil dari UCI Machine Learning Repository. Metode evaluasi menggunakan 10 *fold cross validation*. Perbandingan hasil akurasi menunjukkan bahwa akurasi forward slection dan naïve bayes lebih tinggi dari pada akurasi naïve bayes tanpa forward selection yaitu 93,81%. Pada penelitian sebelumnya juga Harafani [9] menggunakan forward selection pada SVM untuk memprediksi kanker payudara.

Dataset yang digunakan pada penelitian sebelumnya [9] sama dengan dataset yang digunakan pada penelitian ini yaitu dataset kanker payudara Coimbra yang di dapat dari UCI Machine Learning Repository. Metode evaluasi menggunakan *10-fold cross validation*. Hasil menunjukkan kombinasi Forward selection dengan SVM (Kernel RBF) memiliki akurasi yang paling tinggi yaitu sebesar 85,38%. Perbandingan metode, metode evaluasi dan akurasi penelitian terkait dapat dilihat pada Tabel 1.

Tabel 1. Perbandingan Metode, Akurasi, dan Evaluasi penelitian terkait

Peneliti	Tahun	Dataset	Masalah penelitian	Metode Seleksi Fitur	Machine Learning	Performance
Wahyuni	2016	Wisconsin Breast Cancer	Data medis berdimensi Tinggi, fitur perlu direduksi Untuk menghindari Over-fitting	F-Score	-SMO - NB -MLP -C.45	Accuracy NB + FS = 97,65%
Astuti	2018	Seeds	Klasifikasi benih gandum Belum pernah mengguna-Kan seleksi fitur	Forward Selection	-NB	Accuracy NB+FS = 93,81%

Peneliti	Tahun	Dataset	Masalah penelitian	Metode Seleksi Fitur	Machine Learning	Performance
Harafani	2019	Coimbra Breast Cancer	SVM memiliki masalah Dalam memilih fitur untuk Input yang optimal	Forward Selection	-SVM	Accuracy SVM(RBF)+ FS = 85,38%
Harafani	2020	Coimbra Breast Cancer	K-NN sensitive terhadap Attribute yang tidak relevan Dan berkorelasi	Forward Selection, Backward Elimination, Optimize Selection	-KNN	?

2. METODE

A. DATASET

Penelitian ini menggunakan dataset kanker payudara Coimbra yang didapat dari UCI Machine Learning Repository. Dataset ini terdiri dari 10 attribut diantaranya Age, BMI(Body Mass Index), Glucose, Insulin, HOMA(Homoestatis Model Assesment), Leptin, Adiponectin, Resitin, MCP-1(Chemokine Monocyte), dan terakhir adalah label yang terdiri dari Healthy dan Patients. Data ini didapat dari relawan pasien rumah sakit pusat di Coimbra Portugal antara Tahun 2003-2013[23]. Total record daaset berjumlah 116. Sampel terdiri dari 64 wanita yang mengidap kanker payudara, 52 wanita sehat. Dataset kanker payudara Coimbra dapat dilihat pada Tabel 2.

Tabel 2. Dataset Kanker Payudara Coimbra

Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP-1	Classification
85	26,6	96	4,462	1,056602	7,85	7,9317	9,6135	232,006	Healthy
76	27,1	110	26,211	7,111918	21,778	4,935635	8,49395	45,843	Healthy
77	25,9	85	4,58	0,960273	13,74	9,75326	11,774	488,829	Healthy
45	21,30395	102	13,852	3,485163	7,6476	21,05663	23,03408	552,444	Patient
45	20,83	74	4,56	0,832352	7,7529	8,237405	28,0323	382,955	Patient
49	20,95661	94	12,305	2,853119	11,2406	8,412175	23,1177	573,63	Patient
34	24,24242	92	21,699	4,924226	16,7353	21,82375	12,06534	481,949	Patient
42	21,35991	93	2,999	0,687971	19,0826	8,462915	17,37615	321,919	Patient
68	21,08281	102	6,2	1,55992	9,6994	8,574655	13,74244	448,799	Patient
51	19,13265	93	4,364	1,001102	11,0816	5,80762	5,57055	90,6	Patient

B. METODE PENELITIAN

Pada tahap pertama pada penelitian ini, dataset kanker payudara Coimbra akan dibagi terlebih dahulu menjadi data training dan data testing, perbandingan data training dengan data testing yang dipercobakan pada penelitian ini adalah 90/10%, 80/10%, dan 70/10%, kemudian data training yang sudah dipisah dengan data testing akan diseleksi terlebih dahulu attributnya dengan cara pembobotan oleh metode Forward selection.

Berbagai metode seleksi fitur diterapkan untuk menyeleksi fitur yang tidak terpakai seperti Forward selection, Backward Elimination, dan Optimize Selection.

Forward selection digunakan untuk menyeleksi setiap fitur yang tidak terpakai saat memulai iterasi fitur, kemudian fitur-fitur tersebut akan ditambahkan pada subset fitur yang dipilih sebelumnya[24]. Jadi mulanya forward selection akan memulai iterasi dengan tanpa variable (empty model)[18], kemudian proses data yang di training dikerjakan secara step-by-step mulai dari 1 variable sampai dengan jumlah variable yang menghasilkan performa akurasi paling baik[25].

Backward Elimination merupakan suatu metode yang memiliki fungsi untuk mengoptimalkan kinerja suatu model dengan sistem kerja pemilihan mundur. Pemilihan variabel dilakukan dengan pemilihan kedepan yaitu menguji semua variable kemudian menghapus/ mengeliminasi variable yang dianggap tidak signifikan[26]. Backward Elimination digunakan untuk mengukur dampak dari eliminasi sekumpulan fitur pada kinerja klasifikasi yang mana mirip dengan seleksi mundur secara berurutan[21].

Optimize selection adalah satu optimasi seleksi fitur yang dapat memilih atribut-atribut terbaik sehingga dapat mampu meningkatkan tingkat akurasi pada subset yang diujikan[22]. Optimize selection memiliki prinsip mencari dan memilih fitur-fitur dari yang memiliki nilai terbaik dari seluruh subset.

Setelah attribut selesai diseleksi, dataset baru dengan attribut terpilih akan menjadi inputan bagi metode K-NN, kemudian model klasifikasi yang dihasilkan oleh K-NN akan dipercobakan pada data testing. Menurut Widhiasih dalam [27] menjelaskan bahwa prinsip kerja dari K-Nearest Neighbor (KNN) adalah mencari jarak terdekat antara data akan dievaluasi dengan K tetangga (neighbor) terdekatnya dalam data pelatihan. Data training diproyeksikan ke ruang berdimensi banyak, yang mana masing-masing dimensi menjelaskan fitur dari data. Algoritma KNN merupakan salah satu dati algoritma artificial learning yang paling simpel[15].ide dasarnya adalah membuat instance tetangga terdekat. Klas dari instance yang baru kemudian ditentukan sesuai dengan kelas yang paling sering diantara k tetangga terdekat. Pilihan nilai k harus dipilih secara apriori. Berbagai teknik telah diusulkan untuk memilih nilai k . nilai k ini tidak boleh kelipatan dari jumlah kelas untuk menghindari seri untuk data yang genap. Jadi dalam kasus klasifikasi biner, perlu untuk mengambil nilai k yang ganjil sehingga instance mayoritas akan muncul. Formulasi urutan proses kerja K-NN adalah sebagai berikut[13] :

1. Menentukan parameter k (jumlah tetangga yang paling dekat)
2. Menghitung kuadrat jarak euclidean(euclidean distance) masing-masing obyek terhadap data sampel yang diberikan.

$$d_i = \sqrt{\sum_{i=1}^{p}(X_{2i} - X_{1i})^2} \quad (4)$$

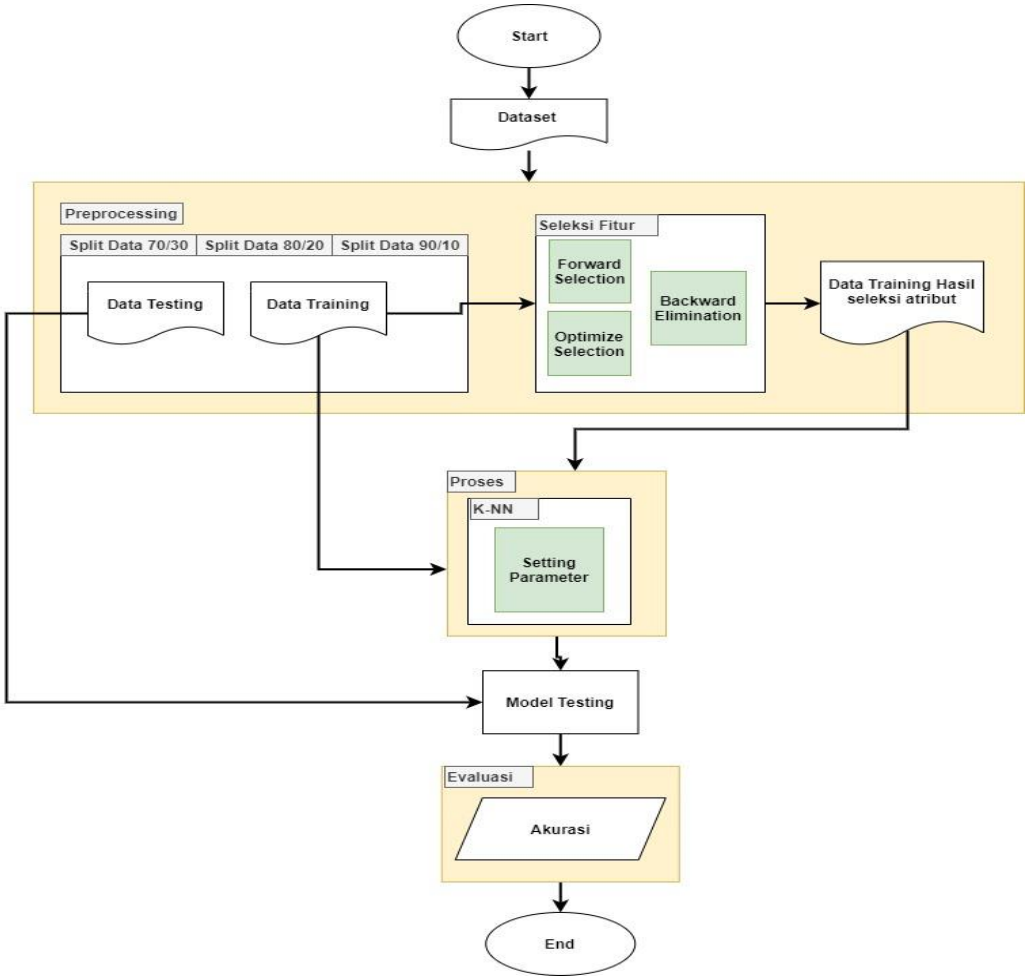
$Keterangan$:
$X_1 =$ Sampel Data
$X_2 =$ Data Testing
$i =$ Variable Data
$d =$ Jarak
$p =$ Dimensi Data

3. Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak euclidean terkecil
4. Mengumpulkan kategori y (klasifikasi nearest neighbor)

Setelah pemrosesan dengan algoritma KNN, selanjutnya pada tahap evaluasi akan dihasilkan akurasi berdasarkan confusion matrix dari pengujian model. Confussion matrix akan menghasilkan akurasi mulai dari prediksi positif yang positif (True Positif), prediksi positif yang negatif (false negative), prediksi negatif yang positif (false positif), dan prediksi negatif yang negatif (false negative). Akurasi akan dihitung dari seluruh prediksi yang benar dibandingkan dengan seluruh data testing. Semakin tinggi nilai akurasi, semakin baik pula model yang dihasilkan[28]. Metrik yang digunakan pada penelitian ini untuk mengukur akurasi didefinisikan pada persamaan (5).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

Kemudian hasil akurasi K-NN akan dibandingkan dengan akurasi algoritma lainnya seperti Naive bayes, Decision Tree, Random Forest, dan Neural Network,serta metode yang digunakan pada penelitian sebelumnya yaitu Support Vector Machine[9]. Alur penelitian dapat dilihat pada Gambar 1.



Gambar 1. Alur penelitian dalam meningkatkan kinerja K-NN untuk klasifikasi kanker payudara menggunakan forward selection.

3. HASIL DAN PEMBAHASAN

Eksperimen dilakukan menggunakan komputer personal Intel Core i3 4GB RAM, 128 MB SSD, sistem operasi Windows 10, dan Rapidminer 9.4.0 Pada tahapan pertama, eksperimen dilakukan untuk mendapatkan bobot atribut dengan forward selection menggunakan metode K-NN dengan mengatur nilai K secara manual sebanyak 5 Kali pada masing-masing percobaan pemisahan proporsi data secara manual. Nilai K yang diatur lebih baik bernilai ganjil sesuai dengan pernyataan Goldberger dalam penelitian [15] yaitu 1,3,5,7,9. Pemisahan proporsi data yang pertama adalah 90% data training dan 10% data testing %. Bobot data yang didapat pada percobaan forward selection terhadap K-NN terhadap proporsi data 90/10% dapat dilihat pada Tabel 3.

Tabel 3.Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan K-NN+FS Proporsi data 90/10%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	0	1	1	1	0
BMI	0	0	0	0	0
Glucose	1	0	1	0	0
Insulin	0	0	0	0	0
HOMA	0	0	0	0	0
Leptin	1	0	1	0	0
Adiponectin	0	0	0	0	0
Resistin	0	0	0	0	1
MCP.1	0	0	0	0	0

Pada Tabel 3. Dapat dilihat bahwa pada algoritma K-NN untuk K bernilai 1 terdapat dua atribut terbaik yaitu glucose dan adiponectin, sedangkan pada K bernilai 3 dan 7 terdapat hanya satu atribut yaitu Age, sedangkan pada K bernilai 5 terdapat 3 attibut yang terpilih yaitu age, glucose, dan leptin, dan pada K bernilai 9 hanya atribut resistin yang terpilih.

Pada tahapan yang ke dua, eksperimen dilakukan untuk mendapatkan bobot atribut dengan forward selection menggunakan metode K-NN dengan mengatur nilai K secara manual sebanyak 5 Kali pada masing-masing percobaan pemisahan proporsi data secara manual. Nilai K yang diatur masih sama dengan eksperimen yang pertama. Pemisahan proporsi data yang kedua adalah 80% data training dan 20% data testing. Bobot data yang didapat pada percobaan forward selection terhadap K-NN terhadap proporsi data 80/20% dapat dilihat pada Tabel 4.

Tabel 4.Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+FS proporsi data 80/20%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	0	1	1	1	1
BMI	0	0	0	1	0
Glucose	1	1	1	0	0
Insulin	0	0	0	0	0
HOMA	1	1	0	0	0
Leptin	0	0	1	0	0
Adiponectin	0	0	0	0	0
Resistin	0	0	0	0	0
MCP.1	0	0	0	0	0

Pada Tabel 4. Dapat dilihat bahwa pada algoritma K-NN untuk K bernilai 1 terdapat dua atribut terbaik yaitu glucose dan HOMA, sedangkan pada K bernilai 3 terdapat tiga atribut yaitu Age,glucose dan HOMA, sedangkan pada K bernilai 5 terdapat 3 attibut yang terpilih yaitu age, glucose, dan leptin, sedangkan pada K bernilai 7 atribut yang terpilih ada dua yaitu age, dan BMI, dan pada K bernilai 9 hanya atribut age yang terpilih.

Pada tahapan yang ke tiga, eksperimen dilakukan untuk mendapatkan bobot atribut dengan forward selection menggunakan metode K-NN dengan mengatur nilai K secara manual sebanyak 5 Kali pada masing-masing percobaan pemisahan proporsi data secara manual. Nilai K yang diatur masih sama dengan eksperimen yang pertama. Pemisahan proporsi data yang ketiga adalah 70% data training dan 30% data testing. Bobot data yang didapat pada percobaan forward selection terhadap K-NN terhadap proporsi data 70/30% dapat dilihat pada Tabel 5.

Tabel 5.Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+FS proporsi data 70/30%

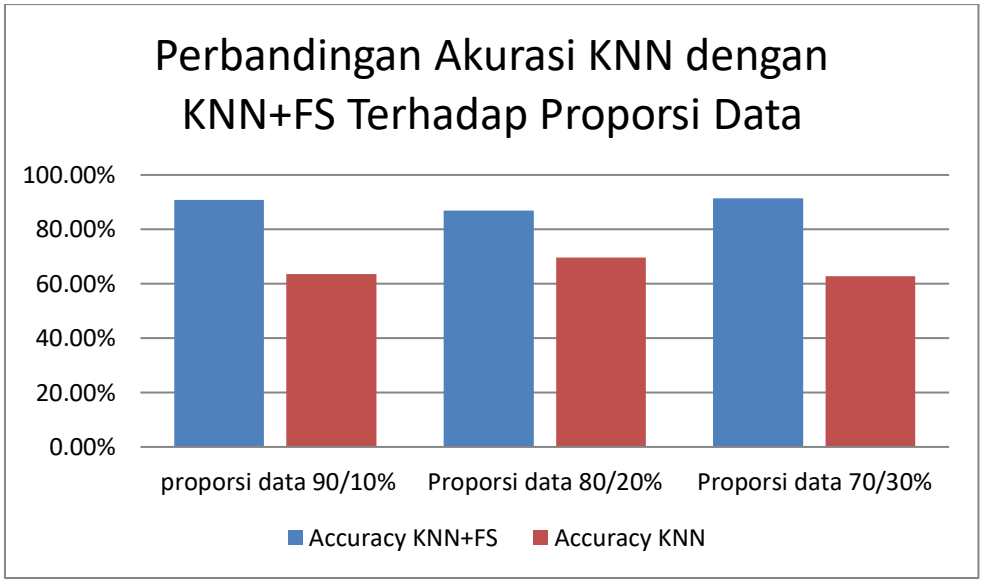
Attribute	K=1	K=3	K=5	K=7	K=9
Age	1	0	1	1	1
BMI	0	0	0	0	0
Glucose	1	1	1	0	0
Insulin	1	1	1	0	0
HOMA	0	0	1	0	0
Leptin	0	0	0	0	0
Adiponectin	1	0	1	0	0
Resistin	1	0	1	0	0
MCP.1	0	0	0	0	0

Pada Tabel 5. Dapat dilihat bahwa pada algoritma K-NN untuk K bernilai 1 terdapat lima atribut terbaik yaitu age, glucose, insulin, adiponectin, dan resistin, sedangkan pada K bernilai 3 hanya terdapat dua atribut yaitu glucose dan insulin, sedangkan pada K bernilai 5 terdapat 6 attibut yang terpilih yaitu age, glucose, insulin, HOMA, adiponectin dan resistin, sedangkan pada K bernilai 7 dan 9 atribut yang hanya age. Perbedaan akurasi K-NN dengan K-NN+FS terhadap proporsi datadapat dilihat pada Tabel 6.

Tabel 6. Perbedaan Akurasi K-NN dengan K-NN+FS pada proporsi data.

K	Proporsi Data 90/10%		Proporsi Data 80/20%		Proporsi Data 70/30%	
	Accuracy K-NN	Accuracy K-NN+FS	Accuracy K-NN	Accuracy K-NN+FS	Accuracy K-NN	Accuracy K-NN+FS
1	**63,64%**	72,73%	**69,57%**	69,57%	**62,86%**	**91,43%**
3	36,36%	81,82%	52,17%	69,57%	48,57%	71,43%
5	36,36%	**90,91%**	43,48%	**86,96%**	40%	88,57%
7	36,36%	81,82%	47,83%	78,26%	48,57%	74,29%
9	36,36%	90,91%	52,17%	73,91%	57,14%	74,29%

Pada Tabel 6. Dapat dilihat bahwa selisih akurasi antara KNN dengan KNN+FS cukup besar untuk semua variasi nilai K yang dipilih. Nilai akurasi paling tinggi adalah 91,43% diperoleh dari kombinasi KNN+FS dengan nilai K=1 dan proporsi data sebesar 70% data training dan 30% data testing dengan atribut yang terpilih adalah age, glucose, insulin, adiponectin, resistin. Secara umum perbandingan akurasi KNN dengan KNN+FS terhadap perbedaan proporsi data dapat digambarkan pada Gambar 2.



Gambar 2. Perbandingan Akurasi KNN dengan KNN+FS Terhadap Proporsi Data

Pada tahapan yang ke empat eksperimen dilakukan untuk mendapatkan bobot atribut dengan metode feature selection yang lain yaitu menggunakan metode backward elimination menggunakan metode K-NN. Pemisahan proporsi data dilakukan secara manual. Pemisahan proporsi data adalah 90% data training dan 10% data testing. Bobot data yang didapat pada percobaan backward elimination terhadap KNN dengan proporsi data 90/10% dapat dilihat pada Tabel 7, sedangkan perbandingan akurasinya dapat dilihat pada Tabel 7.

Tabel 7. Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+BE proporsi data 90/10%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	1	1	1	1	1
BMI	1	0	0	1	1
Glucose	1	1	1	1	1
Insulin	1	1	1	1	1
HOMA	1	1	1	1	1
Leptin	0	1	0	1	1
Adiponectin	1	1	1	1	1

Resistin	1	1	1	0	0
MCP.1	0	0	0	0	0

Pada Tabel 7. Dapat dilihat bahwa pada algoritma K-NN untuk K bernilai 1 terdapat dua atribut yang tidak bercorelasi yaitu leptin dan homa, sedangkan pada K bernilai 3 terdapat satu atribut yaitu Age,glucose dan HOMA, sedangkan pada K bernilai 5 terdapat 3 attibut yang terpilih yaitu BMI, sedangkan pada K bernilai 7 atribut yang terdapat dua variable yaitu resistin, dan MCP.1, dan pada K bernilai 9 juga terdapat 2 atribut yaitu resistin dan MCP.1.

Pada tahapan yang ke lima eksperimen dilakukan untuk mendapatkan bobot atribut dengan metode feature selection yang lain yaitu menggunakan metode backward elimination menggunakan metode K-NN. Pemisahan proporsi data dilakukan secara manual. Pemisahan proporsi data adalah 80% data training dan 20% data testing. Bobot data yang didapat pada percobaan backward elimination terhadap KNN dengan proporsi data 80/20% dapat dilihat pada Tabel 8.

Tabel 8. Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+BE proporsi data 80/20%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	1	1	1	1	1
BMI	1	1	1	1	1
Glucose	1	1	1	1	1
Insulin	0	0	0	0	0
HOMA	1	1	1	1	1
Leptin	1	1	1	1	1
Adiponectin	1	1	1	1	1
Resistin	1	1	1	1	1
MCP.1	0	0	0	1	0

Pada Tabel 8. Dapat dilihat bahwa pada algoritma K-NN semua nilai K menunjukkan atribut insulin tidak berkorelasi, sedangkan attribut MCP.1 hanya berkorelasi pada K=7.

Pada tahapan yang ke enam eksperimen dilakukan untuk mendapatkan bobot atribut dengan metode feature selection yang lain yaitu menggunakan metode backward elimination menggunakan metode K-NN. Pemisahan proporsi data dilakukan secara manual. Pemisahan proporsi data adalah 70% data training dan 30% data testing. Bobot data yang didapat pada percobaan backward elimination terhadap KNN dengan proporsi data 70/30% dapat dilihat pada Tabel 9.

Tabel 9. Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+BE proporsi data 70/30%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	1	1	1	1	1
BMI	1	0	0	1	1
Glucose	1	1	1	1	1
Insulin	1	1	1	1	1
HOMA	1	1	1	1	1
Leptin	0	1	0	1	1
Adiponectin	1	1	1	1	1
Resistin	1	1	1	0	0
MCP.1	0	0	0	0	0

Pada Tabel 9. Dapat dilihat bahwa pada algoritma K-NN pada K bernilai 1 terdapat 2 atribut yang tidak berkorelasi yaitu leptin dan MCP.1, pada K bernilai 3 juga terdapat 2 atribut yang tidak berkorelasi yaitu BMI, dan MCP.1, kemudian pada K bernilai 5 ada tiga atribut yang tidak berkorelasi yaitu BMI,Leptin, dan MCP.1, sedangkan untuk K bernilai 7 dan 9 hanya atribut resistin dan MCP.1 yang tidak berkorelasi. Pada proporsi data 70/30% semua percobaan menunjukkan attribute MCP.1 tidak berkorelasi.

Pada tahapan yang ke tujuh eksperimen dilakukan untuk mendapatkan bobot atribut dengan metode feature selection yang lain yaitu menggunakan metode optimize selection menggunakan metode K-NN. Pemisahan proporsi data dilakukan secara manual. Pemisahan proporsi data adalah 90% data training dan 10% data testing. Bobot data yang didapat pada percobaan optimize selection terhadap KNN dengan proporsi data 90/10% dapat dilihat pada Tabel 10.

Tabel 10. Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+OS proporsi data 90/10%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	0	1	1	1	1
BMI	0	0	0	0	1
Glucose	1	0	0	0	1
Insulin	0	0	0	0	1
HOMA	0	0	0	0	1
Leptin	1	0	0	0	0
Adiponectin	0	0	0	0	0
Resistin	0	0	1	0	1
MCP.1	0	0	0	0	0

Pada Tabel 10. Dapat dilihat bahwa pada algoritma K-NN pada K bernilai 1 terdapat hanya dua atribut yang berkorelasi yaitu glucose dan leptin, pada K bernilai 3 dan 7 hanya terdapat satu atribut yang berkorelasi yaitu age. Kemudian pada K bernilai 5 ada dua atribut yang berkorelasi yaitu age dan resistin, sedangkan untuk K bernilai 9 hanya 3 atribut yang tidak berkorelasi yaitu leptin, adiponectin, dan MCP.1.

Pada tahapan yang ke delapan eksperimen dilakukan untuk mendapatkan bobot atribut dengan metode feature selection yang lain yaitu menggunakan metode optimize selection menggunakan metode K-NN. Pemisahan proporsi data dilakukan secara manual. Pemisahan proporsi data adalah 80% data training dan 20% data testing. Bobot data yang didapat pada percobaan optimize selection terhadap KNN dengan proporsi data 80/20% dapat dilihat pada Tabel 11.

Tabel 11. Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+OS proporsi data 80/20%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	0	1	0	1	1
BMI	0	0	0	0	0
Glucose	1	1	1	1	0
Insulin	0	0	0	0	0
HOMA	1	1	0	0	0
Leptin	0	0	0	1	0
Adiponectin	0	0	1	0	0
Resistin	0	0	0	0	0
MCP.1	0	0	0	0	0

Pada Tabel 11. Dapat dilihat bahwa pada algoritma K-NN pada K bernilai 1 terdapat hanya dua atribut yang berkorelasi yaitu glucose dan HOMA, pada K bernilai 3 ada 3 attribut yang berkorelasi yaitu age, glucose, dan HOMA. Kemudian pada K bernilai 5 terdapat dua atribut yang berkorelasi yaitu glucose dan adiponectin, pada K bernilai 7 terdapat 3 attribut yang berkorelasi yaitu age, leptin, dan glucose. Sedangkan untuk K bernilai 9 hanya 1 atribut yang berkorelasi yaitu age.

Pada tahapan yang ke sembilan eksperimen dilakukan untuk mendapatkan bobot atribut dengan metode feature selection yang lain yaitu menggunakan metode optimize selection menggunakan metode K-NN. Pemisahan proporsi data dilakukan secara manual. Pemisahan proporsi data adalah 70% data training dan 30% data testing. Bobot data yang didapat pada percobaan optimize selection terhadap KNN dengan proporsi data 70/30% dapat dilihat pada Tabel 12.

Tabel 12. Perbedaan bobot attribut terhadap perbedaan nilai K pada percobaan KNN+OS proporsi data 70/30%

Attribute	K=1	K=3	K=5	K=7	K=9
Age	1	0	1	1	1
BMI	0	0	0	0	1
Glucose	1	1	1	0	1
Insulin	0	1	1	0	1
HOMA	1	0	1	0	1
Leptin	0	0	0	0	1
Adiponectin	1	0	1	0	1
Resistin	1	0	1	0	0
MCP.1	0	0	0	0	0

Pada Tabel 12. Dapat dilihat bahwa pada algoritma K-NN pada K bernilai 1 terdapat empat atribut yang berkorelasi yaitu age, glucose, HOMA, Adiponectin, Resistin. Pada K bernilai 3 hanya ada dua attribut yang berkorelasi yaitu glucose, dan Insulin. Kemudian pada K bernilai 5 terdapat enam atribut yang berkorelasi yaitu age, glucose, insulin, HOMA, adiponectin, dan resistin. pada K bernilai 7 hanya terdapat 1 attribut yang berkorelasi yaitu age. Sedangkan untuk K bernilai 9 terdapat enam atribut yang berkorelasi yaitu age, BMI, Glucose, Insulin, HOMA, Leptin, dan Adiponectin. Perbandingan akurasi K-NN+BE, dengan K-NN+OS terhadap proporsi data dapat dilihat pada Tabel 13.

Tabel 13. Perbandingan Akurasi K-NN dengan Berbagai Metode Feature Selection Terhadap Proporsi Data.

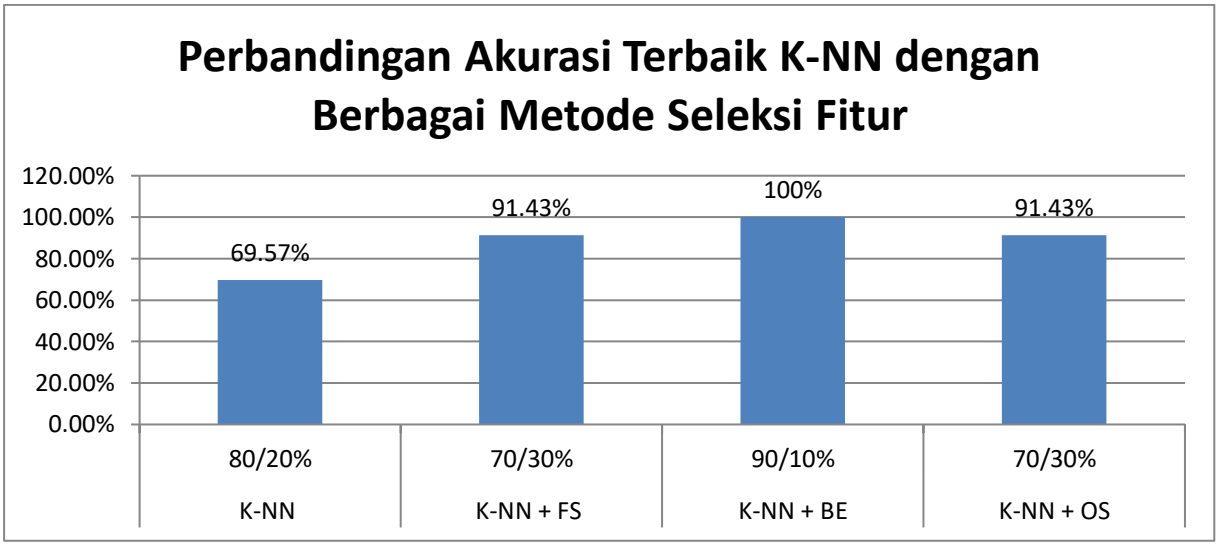
K	Proporsi Data 90/10%		Proporsi Data 80/20%		Proporsi Data 70/30%	
	Accuracy K-NN+ BE	Accuracy K-NN+OS	Accuracy K-NN+ BE	Accuracy K-NN+OS	Accuracy K-NN+BE	Accuracy K-NN+OS
1	90,91%	72,73%	78,26%	69,57%	80,00%	**91,43%**
3	81,82%	81,82%	78,26%	82,61%	80,00%	71,43%
5	81,82%	81,82%	78,26%	78,26%	82,86%	66,87%
7	72,73%	81,82%	73,91%	**86,96%**	77,14%	74,29%
9	**100,00%**	**90,91%**	**82,61%**	73,29%	**82,86%**	74,29%

Pada Tabel 13. Dapat dilihat bahwa pada proporsi data 90/10% kombinasi K-NN+BE memiliki akurasi tertinggi yaitu 100% yang artinya K-NN mengalami overfitting, sedangkan akurasi tertinggi kombinasi metode K-NN+OS pada proporsi data 90/10% adalah 90,91%. Kemudian pada proporsi data 80/20% akurasi tertinggi dari kombinasi K-NN+BE adalah 82,61% dan K-NN+OS adalah 86,96%. Kemudian pada proporsi data 70/30% akurasi tertinggi kombinasi K-NN+BE adalah 82,86%, dan K-NN+OS adalah 91,43%. Secara umum perbandingan akurasi KNN tertinggi dengan berbagai metode feature selection dapat dilihat pada Tabel 14.

Tabel 14.Perbandingan Akurasi Terbaik K-NN dengan berbagai metode seleksi fitur

Method	Nilai K	Proporsi Data	Akurasi Terbaik
K-NN	1	80/20%	69,57%
K-NN + FS	1	70/30%	91,43%
K-NN + BE	9	90/10%	100%
K-NN + OS	1	70/30%	91,43%

Pada Tabel 14. Sangat jelas terlihat bahwa akurasi tertinggi adalah KNN+BE yaitu 100% yang mana menunjukkan KNN+BE mengalami overfittinh. Sedangkan K-NN+ FS memiliki akurasi tertinggi yang sama dengan K-NN+OS sebesar 91,43% dengan proporsi data yang sama dan nilai K yang sama. Perbandingan akurasi terbaik dapat dilihat pada Gambar 3.



Gambar 3. Perbandingan Akurasi Terbaik K-NN dengan Berbagai Metode Seleksi FiturTerhadap Proporsi Data

4. SIMPULAN DAN SARAN

Berdasarkan seluruh hasil pada percobaan ini dapat disimpulkan bahwa seleksi fitur sangat mempengaruhi secara positif terhadap kinerja K-NN. Selain itu Proporsi data juga sangat menentukan keputusan pemilihan fitur yang tepat pada forward selection, dan juga pada kinerja K-NN. Pada penelitian ini dengan menggunakan dataset kanker payudara Coimbra akurasi tertinggi diperoleh berdasarkan kombinasi metode K-NN dengan metode seleksi fitur forward selection dan optimize selection yaitu 91,43% pada proporsi data 70/30% dan nilai K=1. Sedangkan kombinasi K-NN dengan Backward selection mengalami overfitting. Hampir seluruh eksperimen menunjukkan attribute MCP.1 tidak mempunyai korelasi, kecuali pada kombinasi metode KNN+BE pada proporsi data 80/20% dengan nilai K sebesar 7. Implementasi berbagai metode fitur seleksi pada K-NN sangat membantu K-NN dalam meningkatkan performansinya. Penelitian dimasa mendatang tantangan penelitian datang dari berbagai aspek diantaranya penggunaan algoritma metaheuristik dalam pemilihan nilai k yang optimal mungkin dapat meningkatkan kinerja dari K-NN, selain itu penerapan metode validasi k-fold cross validation dengan variasi nilai k mungkin dapat diterapkan untuk proporsi data yang tepat, sehingga didapatkan hasil yang lebih optimal dan stabil.

DAFTAR PUSTAKA

[1] N. Shukla, M. Hagenbuchner, K. T. Win, and J. Yang, “Breast cancer data analysis for survivability studies and prediction,” *Comput. Methods Programs Biomed.*, vol. 155, pp. 199–208, 2018.

[2] B. K. Singh, “Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm,” *Biocybern. Biomed. Eng.*, vol. 39, no. 2, pp. 393–409, 2019.

[3] 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,” *Breast*, vol. 36, no. 2017, pp. 89–95, 2017.

[4] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, “Prediction of survival and metastasis in breast cancer patients using machine learning classifiers,” *Clin. Epidemiol. Glob. Heal.*, 2018.

[5] P. Galetsi, K. Katsaliaki, and S. Kumar, “Values, challenges and future directions of big data analytics in healthcare: A systematic review,” *Soc. Sci. Med.*, vol. 241, no. July, p. 112533, 2019.

[6] U. R. Gogoi, G. Majumdar, M. K. Bhowmik, and A. K. Ghosh, “Evaluating the efficiency of infrared breast thermography for early breast cancer risk prediction in asymptomatic population,” *Infrared Phys. Technol.*, vol. 99, pp. 201–211, 2019.

[7] E. S. Wahyuni, “Penerapan Metode Seleksi Fitur Untuk Meningkatkan Hasil Diagnosis Kanker Payudara,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 7, no. 1, p. 283, 2016.

[8] E. Purwaningsih, “Application of the Support Vector Machine and Neural Network Model Based on Particle Swarm Optimization for Breast Cancer Prediction,” *SinkrOn*, vol. 4, no. 1, p. 66, 2019.
[9] H. Harafani, “Forward Selection pada Support Vector Machine untuk Memprediksi Kanker Payudara,” *J. Infortech (Information Technol.*, vol. 1, no. 2, pp. 131–139, 2019.
[10] F. S. Nugraha, M. J. Shidiq, and S. Rahayu, “Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara,” *J. Pilar Nusa Mandiri*, vol. 15, no. 2, pp. 149–156, 2019.
[11] A. Fauzi, R. Supriyadi, and N. Maulidah, “Deteksi Penyakit Kanker Payudara dengan Seleksi Fitur berbasis Principal Component Analysis dan Random Forest,” *J. Infortech*, vol. 2, no. 1, pp. 96–101, 2020.
[12] Riska Agustin, Vera Maya Santi, and Bagus Sumargo, “Metode Naive Bayes dalam Mendeteksi Sel Kanker Payudara,” *J. Stat. dan Apl.*, vol. 3, no. 1, pp. 30–38, 2019.
[13] R. Yepriyanto, Kustanto, and Y. R. W. Utami, “SISTEM DIAGNOSA KESUBURAN SPERMA DENGAN METODE K-NEAREST NEIGHBOR (K-NN),” *SINUS*, pp. 33–44, 2015.
[14] T. Wakahara and Y. Yamashita, “k -NN classi fi cation of handwritten characters via accelerated GAT correlation,” vol. 47, pp. 994–1001, 2014.
[15] W. Cherif, “Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis,” *Procedia Comput. Sci.*, vol. 127, pp. 293–299, 2018.
[16] A. Bustamam, A. Bachtiar, and D. Sarwinda, “Selecting Features Subsets Based on Support Vector Machine-Recursive Features Elimination and One Dimensional-Naïve Bayes Classifier using Support Vector Machines for Classification of Prostate and Breast Cancer,” *Procedia Comput. Sci.*, vol. 157, pp. 450–458, 2019.
[17] F. D. Astuti, “Seleksi Fitur Forward Selection pada Algoritma Naive Bayes untuk Klasifikasi Benih Gandum,” *J. Inf. Interaktif*, vol. 3, no. 1, pp. 161–166, 2018.
[18] M. F. Nugroho and S. Wibowo, “Fitur Seleksi Forward Selection Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes,” *J. Inform. Upgris*, vol. 3, no. 1, pp. 1–8, 2017.
[19] W. Supriyanti and N. Puspitasari, “Implementasi teknik seleksi fitur forward selection pada algoritma klasifikasi data mining untuk prediksi masa studi mahasiswa politeknik indonusa surakarta,” vol. 4, 2018.
[20] M. Zhu and J. Song, “An embedded backward feature selection method for MCLP classification algorithm,” *Procedia Comput. Sci.*, vol. 17, pp. 1047–1054, 2013.
[21] F. Asdaghi and A. Soleimani, “An effective feature selection method for web spam detection,” *Knowledge-Based Syst.*, vol. 166, pp. 198–206, 2019.
[22] R. T. Prasetio and E. Ripandi, “Optimasi Klasifikasi Jenis Hutan Menggunakan Deep Learning Berbasis Optimize Selection,” *J. Inform.*, vol. 6, no. 1, pp. 100–106, 2019.
[23] M. Patrício *et al.*, “Using Resistin, glucose, age and BMI to predict the presence of breast cancer,” *BMC Cancer*, vol. 18, no. 1, pp. 1–8, 2018.
[24] M. Hasan, “Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naïve Bayes Berbasis Forward Selection,” *Ilk. J. Ilm.*, vol. 9, no. 3, p. 317, 2017.
[25] I. C. R. Drajana, “Metode Support Vector Machine Dan Forward Selection Prediksi Pembayaran Pembelian Bahan Baku Kopra,” *Ilk. J. Ilm.*, vol. 9, no. 2, p. 116, 2017.
[26] A. Bode, “K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika,” *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017.
[27] M. Nanja and P. Purwanto, “Metode K-Nearest Neighbor Berbasis Forward Selection Untuk Prediksi Harga Komoditi Lada,” *J. Pseudocode*, vol. 2, no. 1, pp. 53 – 64, 2015.
[28] H. Saleh, “Prediksi Kebangkrutan Perusahaan Menggunakan Algoritma C4.5 Berbasis Forward Selection,” *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 173–180, 2017.