p-ISSN : 2460-173X
e-ISSN : 2598-5841

Pengaruh Komposisi *Split Data* Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma *Machine Learning*

Rian Oktafiani[1)], Arief Hermawan[2)], Donny Avianto[3)]

[1)2)] Magister Teknologi Informasi, Fakultas Pascasarjana, Universitas Teknologi Yogyakarta
Jalan Siliwangi, Sleman, Daerah Istimewa Yogyakarta, Indonesia
[1)] rian.oktafiani@student.uty.ac.id
[2)] ariefdb@uty.ac.id

[3)]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Teknologi Yogyakarta
Jalan Siliwangi, Sleman, Daerah Istimewa Yogyakarta, Indonesia
[3)] donny@uty.ac.id

Abstrak

Hasil klasifikasi kanker payudara yang tidak tepat dan memiliki akurasi rendah berpotensi membahayakan nyawa pasien. Rasio split data training dan testing mempengaruhi akurasi klasifikasi. Pemilihan rasio split data yang tidak tepat dapat menurunkan akurasi model. Penelitian ini bertujuan menemukan komposisi data terbaik untuk hasil klasifikasi kanker payudara yang baik. Metode yang digunakan adalah holdout dan k-fold cross validation. Algoritma klasifikasi yang dibandingkan adalah SVM, Random Forest, dan Naïve Bayes. Hasil penelitian menunjukkan performa akurasi yang berbeda pada ketiga algoritma tergantung pada metode validasi. Skema holdout validation dengan rasio 75%:25% menghasilkan akurasi terbaik untuk SVM, yaitu 98.89%. Algoritma Random Forest mencapai akurasi terbaik pada rasio split data 55%:45%, yaitu 95.85%. Namun, Naïve Bayes memiliki performa akurasi yang lebih baik saat menggunakan k-fold cross validation dengan akurasi 93.85%. Metode holdout dengan rasio 75:25 terbukti menghasilkan akurasi terbaik untuk klasifikasi data kanker payudara menggunakan SVM. Penelitian selanjutnya dapat menggunakan algoritma deep learning dan memperluas penelitian ke jenis kanker lainnya untuk meningkatkan hasil klasifikasi.

Kata kunci: Kanker Payudara, Klasifikasi, Machine Learning, *Split data*

Abstract

The inaccurate classification results of breast cancer with low accuracy have the potential to endanger patients' lives. The ratio of training and testing data split affects the classification accuracy. Choosing an inappropriate data split ratio can decrease the model's accuracy. This research aims to find the optimal data composition for good breast cancer classification results. The methods used are holdout and k-fold cross-validation. The compared classification algorithms are SVM, Random Forest, and Naïve Bayes. The research results show different accuracy performances for the three algorithms depending on the validation method used. The holdout validation scheme with a 75%:25% data split ratio yields the best accuracy for SVM, which is 98.89%. The Random Forest algorithm achieves the highest accuracy at a 55%:45% data split ratio, with an accuracy of 95.85%. However, Naïve Bayes performs better in terms of accuracy when using k-fold cross-validation, with an accuracy of 93.85%. The holdout method with a 75:25 ratio proves to generate the best accuracy for breast cancer data classification using SVM. Further research can explore the use of deep learning algorithms and expand the study to other types of cancer to improve classification outcomes.

Keywords: *Breast Cancer, Classification, Machine Learning, Split data*

DOI: 10.34128/jsi.v9i1.622
Received: 29 April 2023
Accepted: 25 Juni 2023

1. PENDAHULUAN

Penyebab utama kematian di dunia ini salah satunya karena kanker. Kanker adalah suatu kondisi dimana gen yang mengatur regenerasi sel dalam tubuh manusia rusak dan tumbuh secara tidak normal [1]. Suatu jenis kanker yang berkembang di sel payudara disebut kanker payudara [2]. Terdapat dua jenis kanker payudara yang dapat diklasifikasikan yaitu kanker jinak dan ganas [3]. Berdasarkan data *Global Cancer Observatory* (GLOBOCAN) tahun 2020, dari 19.29 juta kasus kanker, kasus kanker payudara sebanyak 2.261.419 kasus. Kematian akibat kanker payudara sebanyak 684.996 jiwa [4].

Pendekatan tradisional untuk diagnosis kanker sangat tergantung pada pengalaman dokter dan inspeksi visual mereka, yang mana masih terbatas karena adanya potensi kesalahan manusia [5]. Metode untuk mendeteksi jenis kanker payudara, yang menggunakan pembelajaran mesin dapat membantu ahli patologi dalam melakukan tes secara efektif [6]. Salah satu metode *data mining* yang dapat membantu dalam pendeteksian kanker adalah pendekatan menggunakan klasifikasi pembelajaran mesin [7]. Namun, hasil klasifikasi yang tidak tepat dan memiliki akurasi yang rendah dapat menimbulkan diagnosis yang salah dan membahayakan nyawa pasien selain itu rasio *split data training* dan *data testing* mempengaruhi hasil akurasi suatu klasifikasi [8].

Penelitian sebelumnya yang membahas mengenai komposisi *split data*, data *training* dan *testing* yang dapat mempengaruhi nilai akurasi hasil pengolahan data yaitu, penelitian oleh [9] menggunakan rasio 50%:50%, 60%:40%, 70%:30%, 73%:27%, 75%:25%, 80%:20%, 83%:17%, 85%:15%, dan 90%:10% untuk data pelatihan dan pengujian. Algoritma klasifikasi *Decision Tree* C4.5 untuk mengolah data penderita penyakit liver ini menggunakan *Principal Component Analysis* (PCA) dan hasil penelitian menunjukkan bahwa rasio data pelatihan terhadap pengujian adalah 90%:10%, dan jumlah komponen PCA=8 dengan akurasi terbaik 78,40%. Kemudian penelitian oleh [10] yang melakukan komparasi Algoritma *Decision Tree, Naive Bayes, dan K-Nearest Neighbors* dan membuat perbandingan menggunakan teknik *Hold-Out* dan *K-Fold cross validation.* Temuan pengujian menunjukkan bahwa algoritma *K-Nearest Neighbors* secara konsisten mengungguli algoritma *Naive Bayes* dan *Decision Tree* dalam hal performa akurasi, dengan skor 98% dalam teknik *Hold-Out* dan 96% dalam metode *K-Fold.* Menurut penelitian sebelumnya, kinerja hasil klasifikasi dipengaruhi oleh pemilihan persentase data pelatihan dan pengujian yang tepat.

Penelitian ini bertujuan untuk mengevaluasi pengaruh komposisi *split data training dan testing* terhadap performa klasifikasi penyakit kanker payudara pada dataset *Wisconsin Breast Cancer (Diagnostic)* dan membandingkan algoritma *machine learning* yaitu algoritma *Support Vector Machine (SVM), Random Forest* dan *Naïve Bayes yang dapat* digunakan untuk mengelola data numerik pada metode klasifikasi data, hal ini cocok dengan data penelitian ini merupakan data numerik hasil analisis citra digital massa payudara. Ketiga algoritma yang digunakan memiliki keunggulan yaitu metode SVM dapat menentukan *hyperplane* dengan margin terbaik [11], *Random Forest* dapat memilah atribut terbaik [12], sedangkan *Naive Bayes* adalah model yang relatif sederhana dan komputasionalnya cepat untuk dataset besar [13]. Penelitian ini juga bertujuan untuk menghasilkan akurasi terbaik untuk klasifikasi kanker payudara menggunakan komposisi *split data* yang optimal baik itu menggunakan metode *holdout validation* ataupun *k-fold cross validation.*

2. TINJAUAN PUSTAKA

2.1 Klasifikasi dengan *Machine Learning*

Dapat dikatakan bahwa klasifikasi adalah proses pengelompokan data dari suatu objek [14]. Proses klasifikasi data memerlukan identifikasi model atau fungsi untuk membedakan kelas data yang berbeda dalam satu set data [15]. Jadi, dapat didefinisikan bahwa klasifikasi adalah proses melatih/mempelajari fungsi target untuk memetakan setiap atribut atau fitur ke salah satu dari label kelas yang tersedia [16]. Pada *machine learning* memerlukan pembelajaran dari data latih [17], keakuratan hasil pada *machine learning* dapat menurun jika persyaratan tertentu tidak dipenuhi [18].

2.2 *Split data*

Split data adalah teknik yang digunakan untuk mempartisi dataset adalah salah satu dari beberapa aspek yang mempengaruhi seberapa baik kinerja model klasifikasi pada algoritma

pembelajaran mesin [19]. *Split data* merupakan proses untuk membagi antara data latih dan data uji [20]. Metode *holdout validation* dan *k-fold cross validation* dapat digunakan untuk membagi data latih dan data uji. Proses validasi sangat penting untuk dilakukan, tujuannya agar setiap data memiliki peluang sebagai pelatihan data dan pengujian data.

2.2.1 *Holdout Validation*

Pada prinsipnya validasi *holdout* memberikan kesempatan pada setiap data untuk menjadi *data training* dan *data testing*. Sehingga diterapkan uji *holdout validation* untuk mengukur akurasi akan terjadi dua kali [21], seperti yang ditunjukkan pada Persamaan (1):

$$X = \frac{Testing\ A + Testing\ B}{2} \tag{1}$$

Pada Persamaan (1), X merupakan akurasi dari setiap tes yang dihasilkan dari hasil penjumlahan *testing A* dan *testing B* kemudian dibagi dua.

2.2.2 *K-fold Cross Validation*

Data sampel dibagi menjadi K subdivisi yang berbeda, di mana setiap subdivisi berfungsi sebagai subset pelatihan atau subset pengujian. Dalam setiap iterasi *K-fold cross-validation,* subdivisi bergantian digunakan sebagai subset pengujian dan pelatihan. Misalnya, pada iterasi pertama, satu subdivisi digunakan sebagai subset pengujian dan subdivisi lainnya digunakan sebagai subset pelatihan. Proses ini diulang hingga semua subdivisi berfungsi sebagai subset pengujian sekali, dengan demikian, data dipartisi dengan cara ini untuk melakukan validasi silang K-Fold [22]. Secara khusus, proses pelatihan model diulang sebanyak "k" kali, dan kinerja model dihitung sebagai rata-rata dari keseluruhan proses pelatihan [23].

2.3 Algoritma *Support Vector Machine*

Support Vector Machine (SVM) merupakan salah satu teknik dalam *supervised learning* yang sering digunakan untuk regresi dan klasifikasi, seperti *Support Vector Regression* dan *Support Vector Classification.* SVM mencari jarak terdekat untuk menemukan *hyperplane* terbaik dengan margin tertinggi [24]. Adapun tahapan dalam metode *Support Vector Machine* adalah sebagai berikut:

1. Pastikan inisialisasi awal untuk parameter $\alpha = 0.5$, $C = 1$, $\lambda = 0.5$, gamma = 0.5 dan epsilon = 0.001.
2. Menggunakan rumus berikut untuk menghitung matriks

$$D_{ij} = y_i y_j \left(k\left(\overrightarrow{x_i} \cdot \overrightarrow{x_j}\right) + \lambda^2\right) \tag{2}$$

Pada Persamaan (2), D_{ij} adalah elemen matriks data ke-(i, j), y_i dan y_j adalah kelas atau label data ke-i dan ke-j, $\overrightarrow{x_i}$ dan $\overrightarrow{x_j}$ adalah vektor fitur dari sampel ke-i dan ke-j dalam dataset dan λ adalah turunan batas teoritis.

2.4 Algoritma Random Forest

Algoritma *Random Forest* (RF) membagi setiap node secara acak menjadi variabel berdasarkan yang terbaik [3]. Langkah-langkah dalam metode *Random Forest* yaitu, pertama dengan mentukan jumlah pohon (k) yang akan dipilih, dimana k lebih kecil dari m. Pada setiap pohon, pilih N secara acak dari kumpulan data. Sub himpunan prediktor dari setiap pohon dipilih secara acak, dengan m<p, di mana p adalah jumlah total variabel prediktor. Prosedur tahap kedua dan ketiga diulang sebanyak k kali. Hasil kategorisasi pada setiap pohon dengan jumlah suara terbanyak digunakan untuk menentukan hasil prediksi [25].

2.5 *Algoritma Naïve Bayes*

Algoritma Naïve Bayes menggabungkan perhitungan probabilitas [26]. Pada perhitungan menggunakan *Naïve Bayes*, Dataset juga dibagi menjadi data *training* dan *testing*, kemudian dicari nilai rata-rata data pelatihan. Nilai terdekat kemudian diurutkan dari data *training* ke data *testing* untuk memberikan hasil akhir setelah nilai standar deviasi dihitung [27].

$$P(H|E) = P(E|H) \times P(H)P(E) \quad (3)$$

Pada Persamaan (2), P(E) merupakan probabilitas bukti E tanpa mempertimbangkan hipotesis alternatif, P(H) adalah probabilitas awal (priori) dari hipotesis H, tidak dipengaruhi bukti lain, P(H|E) merupakan kemungkinan bersyarat bahwa hipotesis H akan terwujud jika ada bukti untuk E, P(E|H) adalah kemungkinan bukti (E) akan terwujud dan berdampak pada hipotesis (H).

2.6 Confusion Matrix

Confusion matrix menawarkan data tentang perbandingan hasil klasifikasi sistem dengan hasil klasifikasi sebenarnya. *Confusion matrix* ini, yang berbentuk tabel matriks, menggambarkan bagaimana model klasifikasi bekerja ketika dievaluasi pada sekumpulan data yang nilai aktualnya diketahui. Tes akan menghasilkan tiga nilai hasil: *recall,* akurasi, dan presisi [28].

$$Akurasi = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

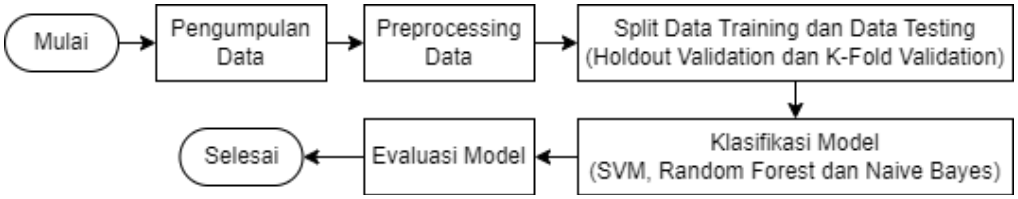
$$Presisi = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TN}{TP+FN} \quad (6)$$

Keterangan pada Persamaan (4), (5) dan (6) yaitu TP *(True Positif)* adalah prediksi kelas dan jumlah data di kelas aktual adalah positif, FN *(False Negative)* adalah jumlah data di kelas sebenarnya positif, tetapi kelas yang diproyeksikan negatif, FP *(False Positive)* adalah kelas yang diproyeksikan positif sedangkan kelas sebenarnya memiliki jumlah data negatif, dan TN *(True Negative)* adalah kelas aktual dan prediksi memiliki jumlah data negatif.

3. METODE PENELITIAN

Pada penelitian ini memiliki beberapa tahapan penelitian. Pada gambar 1 merupakan tahapan penelitian.



Gambar 1. Tahapan Penelitian

Pada Gambar 1, dijelaskan tahapan dalam penelitian ini yaitu proses pengumpulan data, kemudian terdapat proses preprocessing yang bertujuan untuk membersihkan data dari *data noise* atau *missing value,* setelah itu dilakukan skema untuk membagi data latih dan data uji menggunakan *holdout validation* dan *k-fold cross validation* , setelah dilakukan *split data*, maka data akan dimodelkan menggunakan model klasifikasi, kemudian data akan diuji dan dilakukan evaluasi.

3.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data terbuka yang bersumber dari situs koleksi dataset UCI *Machine Learning Repository* yang berjudul *Breast Cancer Wisconsin* [29]. Database tersebut merupakan data yang dikumpulkan dari analisis citra digital massa payudara. Dataset berjumlah 569 data, dengan 31 atribut. Terdapat dua jenis golongan kanker, yaitu ganas (*malignant*) dan jinak (*benign*). Pada tabel 1, merupakan dataset kanker payudara yang digunakan untuk penelitian.

p-ISSN : 2460-173X
e-ISSN : 2598-5841

Tabel 1 *Dataset Breast Cancer*

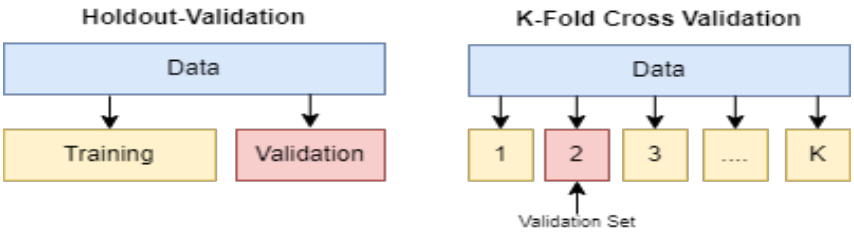
id	*diagnosis*	*radius*	*texture*	*perimeter*	……	*concave points*	*fractal dimension*
842302	2	17.99	10.38	122.8	……	0.1471	0.07871
842517	2	20.57	17.77	132.9	……	0.07017	0.05667
84300903	2	19.69	21.25	130	……	0.1279	0.05999

3.2 *Preprocessing* Data

Preprocessing adalah proses menghilangkan kesalahan atau hal-hal lain yang dianggap tidak relevan dan dapat menurunkan nilai hasil pengolahan data [30]. Pemeriksaan ulang data dilakukan pada tahap *preprocessing/cleaning*, pada tahapan ini terdapat proses menghilangkan redudansi, *outlier*, dan nilai *null* (data kosong) hal ini untuk memastikan data input yang diproses adalah data yang "bersih", sehingga dapat memastikan bahwa hasil perhitungan algoritma *data mining* juga akan memberikan hasil yang sesuai [31].

3.3 *Split Data Training* dan *Data Testing*

Pembelajaran mesin membagi data yang diprosesnya menjadi set pelatihan dan pengujian. Model klasifikasi dilatih menggunakan data pelatihan, dan diuji menggunakan data uji. Skema *Split data* menggunakan *holdout validation* dan *k-fold cross validation*. Berikut pada gambar 2 merupakan tahapan dari metode *holdout validation* dan *k-fold cross validation.*



Gambar 2 Tahapan Metode *Holdout* Validation dan K-Fold Cross Validation

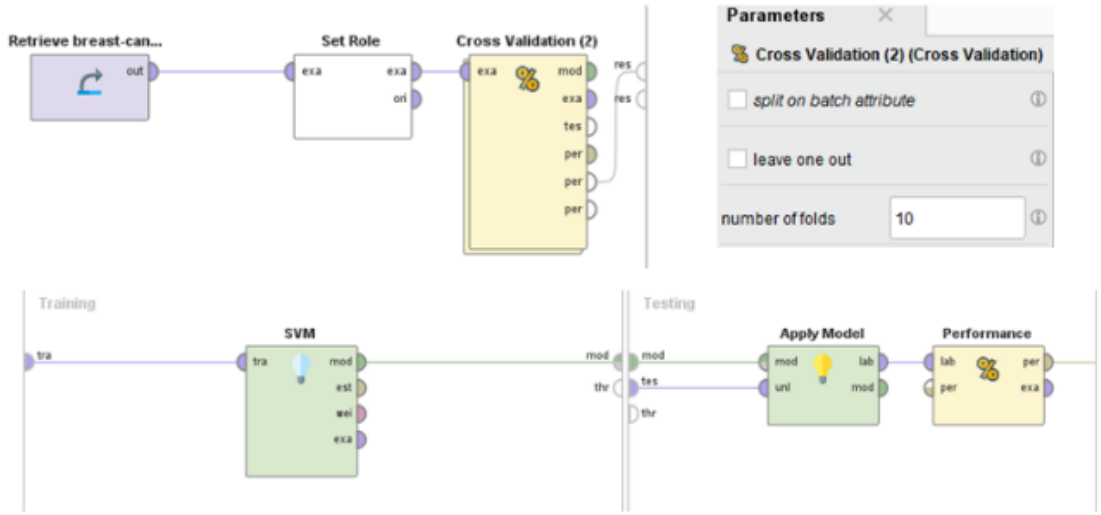
Pada Gambar 2, *Hold-out,* mempartisi kumpulan data menjadi dua bagian, yang pertama disebut "set pelatihan" dan yang lainnya disebut "set validasi" atau tes (yang dikecualikan dari set pelatihan). Sedangkan, *K-fold cross-validation*, melakukan partisi kumpulan data menjadi K kumpulan terpisah dengan ukuran yang K−1 bagian digunakan untuk pelatihan, dan 1 bagian untuk validasi atau pengujian, proses ini diulang K kali dan akhirnya matrik evaluasi ditambahkan. Penelitian ini menggunakan 10 skema *holdout validation* untuk pembagian data latih dan data uji dan dapat dilihat pada Tabel 2.

Tabel 2. Skema *Split data Holdout Validation*

Persentase (Data Latih : Data Uji)	**Data Latih**	**Data Uji**
50%:50%	285	284
55%:45%	313	256
60%:40%	342	227
65%:35%	370	199
70%:30%	399	170
75%:25%	427	142
80%:20%	456	113
83%:17%	473	96
85%:15%	484	85
90%:10%	512	57

Sedangkan untuk Skema *k-fold cross validation* menggunakan 9 skema, yaitu k 2-10. Berikut pada Gambar 3 merupakan tahapan *k-fold cross validation* pada rapidminer menggunakan nilai k=10.

p-ISSN : 2460-173X
e-ISSN : 2598-5841

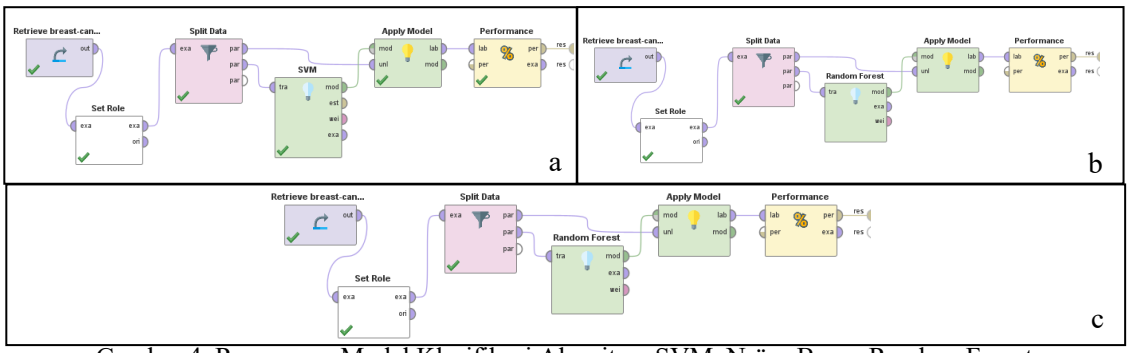


Gambar 3 Tahapan K-Fold Cross Validation Menggunakan Rapid Miner

Pada Gambar 3, pertama dilakukan input data *breast cancer,* kemudian dilakukan *set role* untuk menentukan label dan pada penelitian ini menggunakan atribut diagnosis sebagai label dengan label M atau *Malignant* (Ganas) dan B atau *Benign* (Jinak), kemudian dilakukan proses *k-fold cross validation* dengan menginputkan nilai lipatan (*fold*) atau nilai k =10. Pada *training* data menggunakan metode SVM dan pada *testing*, dilakukan *apply model* dan dilakukan evaluasi performa.

3.4 Model Klasifikasi

Tiga algoritma *machine learning* yaitu *Support Vector Machine, Random Forest* dan *Naïve Bayes* digunakan dalam penelitian ini. Klasifikasi model ini menggunakan *tool RapidMiner* untuk pengolahan data klasifikasi kanker payudara. Pada model klasifikasi SVM, menggunakan kernel linear dengan *kernel cache* = 200, *convergence epsilon* = 0.001 dan *max iteration* = 100000. Gambar 4 merupakan rancangan pengujian algoritma SVM menggunakan *RapidMiner.*



Gambar 4. Rancangan Model Klasifikasi Algoritma SVM, Naïve Bayes Random Forest

Pada Gambar 4 merupakan rancangan model klasifikasi menggunakan algoritma SVM (a), Naïve Bayes (b) dan Random Forest (c) menggunakan *RapidMiner* memiliki tahapan yaitu *input data set breast cancer,* kemudian *set role* yang digunakan untuk menentukan label untuk klasifikasi dan pada penelitian ini menggunakan atribut diagnosis sebagai label yaitu *Malignant* (ganas) dan *benign* (jinak). Kemudian dilakukan *split data* untuk membagi data latih dan data uji dan dilakukan berbagai skema percobaan *split data* dengan teknik *holdout validation* yang telah ditentukan. Kemudian data akan diolah atau dilakukan *apply model* menggunakan algoritma dan dilakukan pengujian serta evaluasi untuk melihat performa akurasi algoritma menggunakan *confusion matrix*. Pengolahan data menggunakan algoritma *Random Forest* menggunakan *number of trees* = 100, dan *maximal depth* = 10.

p-ISSN : 2460-173X
e-ISSN : 2598-5841

4. PEMBAHASAN

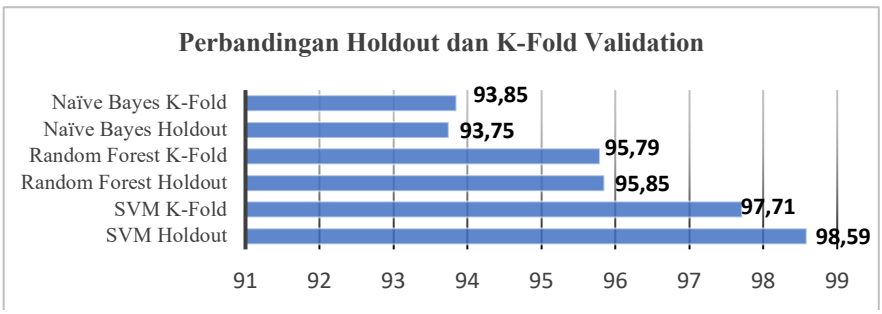
Data yang digunakan dalam penelitian ini adalah *Dataset Breast Cancer Wisconsin* yang berjumlah 569 data dengan 31 atribut. Pada saat proses *preprocessing* data dilakukan pengecekan *missing value,* dan dataset ini tidak ada *missing value*. Namun saat proses *preprocessing* dilakukan *featured selection,* yaitu proses pemilihan fitur yang berpengaruh terhadap klasifikasi dan mengesampingkan fitur yang tidak berpengaruh. Dari 31 atribut atau fitur yang digunakan dalam penelitian ini hanya 10 atribut yang digunakan yaitu *Diagnosis, Radius, Texture, Area, Smoothness, Compactness, Concavity, Concave, Concave Point, Fractal Dimension*. Atribut diagnosis digunakan sebagai label klasifikasi yaitu *Malignant* (ganas) dan *Benign* (jinak).

Setelah dilakukan proses *preprocessing,* kemudian dilakukan proses *split data* menggunakan skema *holdout* dan *k-fold cross validation.* Terdapat 10 skema persentase *split data training* dan *data testing*, untuk *holdout validation* dan 9 skema *split data* untuk *k-fold cross validation* yaitu k=2-10. Setelah penentuan skema *split data training* dan *data testing*, dilakukan proses pengolahan data menggunakan *tool RapidMiner Studio* dan menggunakan model klasifikasi menggunakan algoritma SVM, *Random Forest* dan *Naïve Bayes*. Tabel 3 merupakan hasil pengujian pengolahan data menggunakan algoritma SVM, *Random Forest* dan *Naïve Bayes* dan menggunakan skema *holdout validation.*

Tabel 3. Hasil Pengujian Menggunakan Skema *Holdout Validation* VS *K-Fold Cross Validation*

Holdout Validation				***K-Fold Cross Validation***			
Perban-dingan (%)	**SVM (%)**	***Random Forest* (%)**	***Naïve Bayes* (%)**	**K-Fold (k)**	**SVM (%)**	***Random Forest* (%)**	***Naïve Bayes* (%)**
50:50	95.07	95.44	92.25	10	96.84	94.91	93.51
55:45	98.05	**95.85**	**93.75**	9	97.18	95.08	93.32
60:40	97.81	95.60	92.98	8	97.54	95.6	92.79
65:35	97.49	95.68	91.96	7	**97.71**	95.44	93.33
70:30	94.74	94.47	92.40	6	97.36	**95.79**	92.97
75:25	**98.59**	93.44	92.25	5	97.37	94.91	93.33
80:20	97.35	94.30	91.15	4	97.02	95.61	92.79
83:17	97.94	94.92	92.78	3	97.01	95.78	**93.85**
85:15	96.51	94.00	91.86	2	96.48	95.25	92.97
90:10	94.74	93.55	89.41				

Berdasarkan Tabel 3 yang merupakan hasil pengujian dan perbandingan algoritma SVM, *Random Forest* dan *Naïve Bayes* dan menggunakan skema *split data holdout validation* dan *k-fold cross validation* didapatkan hasil akurasi terbaik untuk algoritma SVM yaitu 98.85% dengan perbandingan *split data* 75%:25%. Hasil akurasi terbaik untuk algoritma *Random Forest* dan *Naïve Bayes* terdapat pada komposisi *split data* 55%:45% yaitu *Random Forest* 95.85% dan *Naïve Bayes* 93.75%. Pada skema *split data* untuk *k-fold cross validation* yaitu k=2-10, hasil pengujian menghasilkan akurasi terbaik untuk algoritma SVM yaitu 97.71% dan pada skema k-fold k=7, kemudian akurasi algoritma *Random Forest* yaitu 95.79% pada skema k-fold k=6, akurasi algoritma *Naïve Bayes* yaitu 93.85% pada skema k-fold k=3. Pada Gambar 5 merupakan penjelasan perbandingan hasil pengujian menggunakan *holdout* dan *k-fold cross validation*.



Gambar 5. Perbandingan Hasil Pengujian *Holdout* dan *K-fold cross validation*

Pada Gambar 5, performa akurasi yang dihasilkan oleh skema *holdout validation* lebih unggul dibandingkan menggunakan skema *k-fold cross validation* untuk algoritma *Random Forest* dengan akurasi terbaik 95.85% dan SVM 98.89%, sedangkan untuk algoritma *Naïve Bayes* performa akurasi yang dihasilkan lebih unggul saat menggunakan skema *k-fold cross validation* yaitu menghasilkan akurasi 93.85%. Berdasarkan penelitian yang telah dilakukan, komposisi *split data* dan akurasi yang dihasilkan tergantung dengan kebutuhan dan karakteristik dataset yang digunakan, serta perlakuan terhadap data seperti menggunakan reduksi dimensi atau seleksi fitur juga berpengaruh terhadap akurasi yang akan dihasilkan. Pada penelitian, performa akurasi skema *holdout validation* lebih baik daripada skema *k-fold cross validation* untuk Random Forest (95.85%) dan SVM (98.89%). Namun, skema *k-fold cross validation* memberikan performa akurasi lebih baik untuk Naïve Bayes (93.85%). Akurasi dipengaruhi oleh skema validasi dan algoritma yang digunakan.

Pada penelitian, skema *holdout validation* sederhana dengan waktu komputasi yang cepat, namun hasil yang diperoleh bisa tergantung pada pembagian acak awal data, yang dapat menghasilkan variasi dalam evaluasi model. Sedangkan pada *k-fold cross validation* lebih stabil dan konsisten karena mengulangi proses pelatihan dan pengujian sebanyak K kali, sehingga dapat memberikan estimasi yang lebih akurat tentang kinerja model. Namun, *k-fold cross validation* membutuhkan waktu komputasi yang lebih lama, terutama jika nilai K sangat besar atau dataset sangat besar.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan yaitu klasifikasi kanker payudara menggunakan *dataset Breast Cancer Wisconsin* yang dengan membandingkan algoritma *Support Vector Machine* (SVM), *Random Forest* dan *Naïve Bayes* dilakukan analisis dan evaluasi komposisi *split data* dengan teknik *holdout validation* dan *k-fold cross validation* secara keseluruhan menghasilkan performa akurasi yang dihasilkan oleh skema *holdout validation* lebih unggul dibandingkan menggunakan skema *k-fold cross validation* untuk algoritma SVM 98.89% pada persentase *split data* 75%:25% dan *Random Forest* dengan akurasi terbaik 95.85% pada skema 55%:45%, sedangkan untuk algoritma *Naïve Bayes* performa akurasi yang dihasilkan lebih unggul saat menggunakan skema *k-fold cross validation* yaitu menghasilkan akurasi 93.85%. Pemilihan komposisi *split data* yang baik dapat membantu menghindari *overfitting* atau *underfitting,* sehingga model klasifikasi dapat menghasilkan kinerja yang lebih baik pada data yang belum pernah dilihat sebelumnya. Berbagai hal dapat dilakukan untuk penelitian selanjutnya untuk memperbaiki klasifikasi kanker payudara yaitu menggunakan algoritma *deep learning* dan dilakukan penelitian terhadap jenis kanker lainnya.

DAFTAR PUSTAKA

[1] N. P. W. P. Sari, "Women Living With Breast and Cervical Cancer in the Community: The Face of Surabaya Nowadays," *Indones. J. ofr Cancer*, vol. 12, no. 4, pp. 116–122, 2018, doi: https://doi.org/10.33371/ijoc.v12i4.605.

[2] Komite Penanggulangan Kanker Nasional, *Panduan Penatalaksanaan Kanker Payudara*. Kementerian Kesehatan Republik Indonesia.

[3] A. M. Widodo, N. Anwar, B. Irawan, A. Wisnujati, and L. Meria, "Komparasi Performansi Algoritma Pengklasifikasi KNN, Bagging Dan Random Forest Untuk Prediksi Kanker Payudara," *Proceeding KONIK (Konferensi Nas. Ilmu Komputer)*, vol. 5, pp. 367–372, 2021.

[4] Globocan, "Breast Cancer Fact Sheet," 2020.

[5] A. Ed-daoudy and K. Maalmi, "Breast cancer classification with reduced feature set using association rules and support vector machine," *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 9, no. 1, 2020, doi: https://doi.org/10.1007/s13721-020-00237-8.

[6] R. Erwandi and S. Suyanto, "Klasifikasi Kanker Payudara Menggunakan Residual Neural Network," *Indones. J. Comput.*, vol. 5, no. 1, pp. 45–52, 2020, doi: https://doi.org/10.34818/INDOJC.2020.5.1.373.

[7] D. T. Artha, S. Adinugroho, and P. P. Adikara, "Klasifikasi Pengidap Kanker Payudara

Menggunakan Metode Voting Based Extreme Learning Machine (V-ELM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2180–2186, 2019.

[8] V. R. Joseph, "Optimal ratio for data splitting," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 15, no. 4, pp. 531–538, 2022, doi: https://doi.org/10.1002/sam.11583.

[9] B. N. Azmi, A. Hermawan, and D. Avianto, "Analisis Pengaruh Komposisi Data Training dan Data Testing pada Penggunaan PCA dan Algoritma Decision Tree untuk Klasifikasi Penderita Penyakit Liver," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 4, no. 4, pp. 281–290, 2023, doi: https://doi.org/10.35746/jtim.v4i4.298.

[10] M. A. Jabbar, E. Hasmin, Sunardi, C. Susanto, and W. Musu, "Komparasi Algoritma Decision Tree, Naive Bayes dan KNN dalam Klasifikasi Kanker Payudara," *Comput. Sci. Res. Its Dev. J.*, vol. 14, no. 3, pp. 258–270, 2022, doi: https://doi.org/10.22303/csrid.14.3.2022.258-270.

[11] S. Adi and A. Wintarti, "Komparasi Metode Support Vector Machine (SVM), K-Nearest Neighbors (KNN), dan Random Forest (RF) Untuk Prediksi Penyakit Gagal Jantung," *MATHunesa J. Ilm. Mat.*, vol. 10, no. 2, pp. 258–268, 2022, doi: https://doi.org/10.26740/mathunesa.v10n2.p258-268.

[12] V. R. Sari, F. Firdausi, and Y. Azhar, "Perbandingan Prediksi Kualitas Kopi Arabika dengan Menggunakan Algoritma SGD, Random Forest dan Naive Bayes," *Edumatic J. Pendidik. Inform.*, vol. 4, no. 2, pp. 1–9, 2020, doi: https://doi.org/10.29408/edumatic.v4i2.2202.

[13] M. G. Pradana, P. H. Saputro, and D. P. Wijaya, "Komparasi Metode Support Vector Machine Dan Naïve Bayes Dalam Klasifikasi Peluang Penyakit Serangan Jantung," *Indones. J. Bus. Intell.*, vol. 5, no. 2, pp. 87–91, 2022, doi: http://dx.doi.org/10.21927/ijubi.v5i2.2659.

[14] N. M. Putry, "Komparasi Algoritma KNN Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," *Evolusi J. Sains Dan Manaj.*, vol. 10, no. 1, pp. 45–57, 2022, doi: https://doi.org/10.31294/evolusi.v10i1.12514.

[15] N. B. Putri and A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing," *Komputika J. Sist. Komput.*, vol. 11, no. 1, pp. 59–66, 2022, doi: https://doi.org/10.34010/komputika.v11i1.4350.

[16] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, pp. 437–444, 2020, doi: http://dx.doi.org/10.30865/mib.v4i2.2080.

[17] C. Chazar and B. Erawan, "Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine," *Inf. (Jurnal Inform. Dan Sist. Informasi)*, vol. 12, no. 1, pp. 67–80, 2020.

[18] P. D. Kusuma, *Machine Learning Teori, Program, dan Studi Kasus*. Sleman: Deepublish, 2020.

[19] A. Nurhopipah and U. Hasanah, "Dataset Splitting Techniques Comparison For Face Classification on CCTV Images," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 14, no. 4, pp. 341–352, 2020, doi: https://doi.org/10.22146/ijccs.58092.

[20] A. Yasin, A. Yuniarti, and Y. A. Nugroho, "Efektifitas Algoritma Data Mining dalam Menentukan Pendonor Darah Potensial," *Syntax J. Inform.*, vol. 11, no. 01, pp. 12–22, 2022, doi: https://doi.org/10.35706/syji.v11i01.6595.

[21] F. Tempola, R. Rosihan, and R. Adawiyah, "Holdout Validation for Comparison Classfication Naïve Bayes and KNN of Recipient Kartu Indonesia Pintar," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1125, no. 1, 2021, doi: https://dx.doi.org/10.1088/1757-899X/1125/1/012041.

[22] P. I. Nainggolan, D. S. Prasvita, and D. S. Bukit, "Klasifikasi Informasi Kesehatan Pada Data Media Sosial Menggunakan Support Vector Machine dan K-Fold Cross Validation," *Malikussaleh J. Mech. Sci. Technol.*, vol. 5, no. 2, pp. 34–38, 2021, doi: https://doi.org/10.29103/mjmst.v5i2.6317.

[23] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode Smote Random Forest

Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data," *J. Gaussian*, vol. 9, no. 3, pp. 227–236, 2020, doi: https://doi.org/10.14710/j.gauss.9.3.227-236.

[24] M. N. Akbar, N. A. S. Yusuf, Nasrullah, and Mubarak, "Analisis Sentimen Pengguna Indihome dengan Metode Klasifikasi Support Vector Machine (SVM)," *J. Software, Hardw. Inf. Technol.*, vol. 2, no. 1, pp. 13–21, 2022, doi: https://doi.org/10.24252/shift.v2i1.18.

[25] U. Erdiansyah, A. I. Lubis, and K. Erwansyah, "Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil," *J. Media Inform. Budidarma*, vol. 6, no. 1, pp. 208–214, 2022, doi: http://dx.doi.org/10.30865/mib.v6i1.3373.

[26] L. W. Astuti, I. Saluza, F. Faradilla, and M. F. Alie, "Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes," *J. Inform. Glob.*, vol. 11, no. 2, pp. 63–67, 2021, doi: https://doi.org/10.36982/jiig.v11i2.1235.

[27] A. P. Ayudhitama and U. Pujianto, "Analisa 4 Algoritma Dalam Klasifikasi Liver Menggunakan Rapidminer," *J. Inform. Polinema*, vol. 6, no. 2, pp. 1–9, 2020, doi: https://doi.org/10.33795/jip.v6i2.274.

[28] D. S. Suparno, "Pengenalan Pola Untuk Mengetahui Jumlah Target Pengunjung Mall Berdasarkan Usia, Gender, Pendapatan Pertahun, Pengeluaran, Tujuannya Untuk Mempermudah Mengetahui Target Pasar Menggunakan Metode EDA, K-Means, Hierarchial Clustering, Confusion Matrix," *Sains, Apl. Komputasi dan Teknol. Inf.*, vol. 3, no. 2, pp. 61–69, 2023, doi: http://dx.doi.org/10.30872/jsakti.v3i2.4445.

[29] W. Wolberg, O. Mangasarian, N. Street, and W. Street, "Breast Cancer Wisconsin (Diagnostic)." UCI Machine Learning Repository, 1995, doi: https://doi.org/10.24432/C5DW2B.

[30] T. Hidayat, M. Priyatna, A. Sutanto, A. Al Khudri, and R. Komaruddin, "Informasi Sebaran Titik Panas Berbasis WebGIS untuk Pemantauan Kebakaran Hutan dan Lahan di Indonesia," *J. Teknol. Lingkung.*, vol. 20, no. 1, pp. 105–112, 2019.

[31] S. Widaningsih, "Penerapan Data Mining untuk Memprediksi Siswa Berprestasi dengan Menggunakan Algoritma K Nearest Neighbor," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, pp. 2598–2611, 2022, doi: https://doi.org/10.35957/jatisi.v9i3.859.

Biodata Penulis

Rian Oktafiani, **P**enulis merupakan mahasiswa program studi Magister Teknologi Informasi di Universitas Teknologi Yogyakarta. Penulis memiliki minat penelitian dibidang *data mining* dan *information system.*

Arief Hermawan, penulis bekerja sebagai dosen tetap program studi Teknologi Informasi di Universitas Teknologi Yogyakarta. Penulis memiliki minat penelitian dibidang *neural network, data mining* dan *information system.*

Donny Avianto, penulis bekerja sebagai dosen tetap program studi Informatika di Universitas Teknologi Yogyakarta. Penulis memiliki minat penelitian dibidang *neural network, data mining* dan *information technology.*