Laman web jurnal: https://ejournal.unama.ac.id/index.php/processor

Processor: Jurnal Ilmiah Sistem Informasi, Teknologi Informasi dan Sistem Komputer

P-ISSN: 1907-6738 | E-ISSN: 2528-0082

Perbandingan Klasifikasi Penyakit Kanker Paru-paru menggunakan *Support Vector Machine* dan *K-Nearest Neighbor*

Sri Indra Maiyanti[1], Des Alwine Zayanti[2], Yuli Andriani[3], Bambang Suprihatin[4], Anita Desiani[5], Aulia Salsabila[6], Nyayu Chika Marselina[7]

[1,2,3,4,5,6,7]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih KM. 32, Indralaya, Sumatra Selatan 30862, Indonesia.

ABSTRACT

Lung cancer is a condition where cells grow uncontrollably in the lungs due to carcinogens. Lung cancer is the first cause of death in men and women's second cause of death. One way to reduce the death rate due to lung cancer is to carry out early detection, that is classification. The process of identifying and grouping objects with the same characteristics or characteristics into several predetermined classes is called classification. Several algorithms widely used in the classification process are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). SVM has advantages, being able to identify hyperplanes separately to maximize the margin between two or more different classes, but it is difficult to use in large data, while KNN can perform large-scale data separation and is resilient to noise in the data. This study aims to build a model using the SVM and KNN algorithms to classify lung cancer. The lung cancer dataset has a total of 309 data, where data is divided using the percentage split method and k-fold cross validation on each algorithm used. The parameters used in evaluating the model are accuracy, precision, and recall. From the research, the highest accuracy, precision, and recall values were obtained in the SVM algorithm with the percentage split method with consecutive values, namely 95.16%, 88%, and 82.5%. This indicates that the SVM algorithm with the percentage split method performs better in classifying lung cancer than other algorithms and methods.

Keywords: K-Fold Cross Validation, K-Nearest Neighbor, Lung Cancer, Percentage Split, Support Vector Machine

ABSTRAK

Kanker paru-paru merupakan suatu kondisi sel-sel tumbuh secara tidak terkendali pada paru-paru dikarenakan karsinogen. Kanker paru-paru termasuk ke dalam penyebab pertama kematian pada pria dan menjadi penyebab kedua kematian pada wanita. Salah satu cara untuk mengurangi tingkat kematian karena kanker paru-paru adalah dengan melakukan deteksi dini, yakni dengan klasifikasi. Proses mengidentifikasi dan mengelompokkan objek dengan ciri atau karakteristik yang sama ke dalam beberapa kelas yang telah ditentukan disebut dengan klasifikasi. Beberapa algoritma yang banyak digunakan dalam proses klasifikasi adalah *Support Vector Machine* (SVM) dan *K-Nearest Neighbor* (KNN). SVM memiliki kelebihan, yakni mampu mengidentifikasi *hyperplane* secara terpisah sehingga memaksimalkan *margin* antara dua kelas atau lebih yang berbeda, tetapi sulit digunakan dalam data yang berukuran besar, sedangkan KNN dapat melakukan pemisahan data yang berskala besar dan tangguh terhadap *noise* pada data. Penelitian ini bertujuan untuk membangun model dengan menggunakan algoritma SVM dan KNN pada klasifikasi penyakit kanker paru-paru. Dataset penyakit kanker paru-paru memiliki jumlah data sebanyak 309 data dimana data dibagi dengan menggunakan metode *percentage split* dan *k-fold cross validation* pada masing-masing algoritma yang digunakan. Parameter yang digunakan dalam mengevaluasi model adalah akurasi, presisi, dan *recall*. Dari penelitian yang dilakukan, nilai akurasi, presisi, dan *recall* tertinggi diperoleh pada algoritma SVM metode *percentage split* dengan nilai secara berturut-turut, yakni 95,16%, 88%, dan 82,5%. Hal tersebut mengindikasikan bahwa algoritma SVM dengan metode *percentage split* memiliki performa yang lebih baik dalam melakukan klasifikasi penyakit kanker paru-paru dibandingkan algoritma dan metode lainnya.

Kata Kunci: Kanker Paru-Paru, *K-Fold Cross Validation, K-Nearest Neighbor, Percentage Split, Support Vector Machine*

1. PENDAHULUAN

Kanker paru-paru adalah kondisi sejumlah karsinogen menyebabkan sel-sel tumbuh secara tidak terkendali pada paru-paru [1]. Secara global, kanker paru-paru menjadi penyebab pertama kematian akibat kanker pada pria dan penyebab kedua kematian akibat kanker pada wanita [2]. Salah satu cara menanggulangi penyakit kanker paru-paru agar tingkat kematian tidak terlalu tinggi adalah dengan melakukan pencegahan dan melakukan deteksi dini [3]. Klasifikasi merupakan salah satu cara dalam melakukan deteksi dini. Klasifikasi adalah proses mengidentifikasi dan menggolongkan objek dengan karakteristik yang sama ke dalam beberapa kelas [4]. Dalam melakukan klasifikasi, dibutuhkan suatu algoritma tertentu agar proses klasifikasi bekerja lebih optimal.

Salah satu algoritma yang banyak digunakan dalam menyelesaikan tugas klasifikasi adalah *Support Vector Machine* (SVM). Algoritma SVM adalah algoritma yang menerapkan pemetaan nonlinear untuk mengubah *data training* asli ke skala yang lebih tinggi [5]. SVM memiliki kelebihan, yaitu mampu mengenali *hyperplane* secara terpisah sehingga memaksimalkan batas antara dua kelas atau lebih yang berbeda [6]. SVM telah diterapkan pada beberapa penelitian, seperti penelitian Ahmed et al, 2019 [7] mengenai deteksi kanker paru-paru dengan menggunakan dataset yang diperoleh dari UCI *machine learning repository* memberikan hasil akurasi sebesar 79,40%, tetapi tidak menampilkan hasil presisi dan *recall* dalam penelitian yang dilakukan. Francis dan Babu, 2019 [8] mengenai prediksi prestasi pada siswa dengan menggunakan data yang dikumpulkan berdasarkan demografis, akademik, perilaku dan fitur tambahan yang menghasilkan tingkat akurasi sebesar 66,03%, tetapi tidak menampilkan hasil presisi dan *recall*. Penelitian Woldemichael dan Menaria, 2018 [9] melakukan klasifikasi pada penyakit diabetes dengan hasil akurasi sebesar 81,69%, tetapi tidak menampilkan hasil presisi dan *recall*.

Selain itu, SVM juga memiliki kelemahan, yakni sulit digunakan pada data yang berukuran besar dan memiliki perhitungan yang kompleks [10]. Algoritma lain yang mampu melakukan pemisahan pada data yang berskala besar adalah *K-Nearest Neighbor* (KNN). Algoritma KNN merupakan algoritma klasifikasi berdasarkan sejumlah k data yang terdekat [11]. Kelebihan dari KNN, yaitu tangguh terhadap data latih yang memiliki *noise* dan efektif jika digunakan pada data latih dengan jumlah yang besar [12]. KNN telah diterapkan pada beberapa penelitian, antara lain Bharati et al, 2020 [13] yang melakukan klasifikasi kanker paru-paru menggunakan data dari UCI *Machine Learning Repository* dengan nilai akurasi dan *recall* sebesar 44%, serta presisi sebesar 58%. Devika et al, 2019 [14]. melakukan klasifikasi pada penyakit ginjal kronis memberikan hasil nilai akurasi, presisi, dan *recall* sebesar 87%. KNN memiliki kekurangan, yakni perlu ditentukannya nilai k terbaik yang menyatakan jumlah tetangga terdekat dan biaya komputasi yang cukup tinggi karena perhitungan jarak yang harus dilakukan pada setiap data latih [15]. Adapun dataset yang banyak digunakan dalam melakukan klasifikasi adalah *lung cancer* yang dapat diakses di Kaggle melalui link https://www.kaggle.com/datasets/h13380436001/h-lung-cancer. Dataset tersebut memiliki dua label, yakni ya yang berarti mengidap kanker paru-paru dan tidak yang berarti tidak mengidap kanker paru-paru, serta memiliki dimensi fitur yang cukup banyak, yakni 16 fitur, sehingga diperlukan algoritma tertentu untuk melakukan klasifikasi pada dua kelas dan dapat bekerja pada dimensi yang cukup besar.

Pada penelitian ini akan dilakukan perbandingan kinerja antara algoritma SVM dan KNN pada kanker paru-paru menggunakan dataset *lung cancer*. Implementasi dua algoritma dilakukan dengan tujuan untuk melihat algoritma klasifikasi yang lebih baik dan efektif dalam melakukan klasifikasi kanker paru-paru. Terdapat dua metode yang digunakan untuk membagi data sebelum proses pelatihan model, yakni metode *percentage split* dan *k-fold cross validation*. Pengukuran kinerja dari setiap algoritma klasifikasi dan metode pembagian data dilakukan dengan menggunakan akurasi, presisi, dan *recall*.

2. TINJAUAN PUSTAKA

2.1. Support Vector Machine

SVM merupakan algoritma yang menggunakan pemetaan nonlinear untuk mengubah data latih ke skala atau ukuran yang lebih tinggi [16]. SVM adalah metode yang ampuh untuk membangun *classifier*. Ini bertujuan untuk membuat batas keputusan antara dua kelas sehingga memungkinkan prediksi label dari satu atau beberapa fitur vektor [17]. SVM memetakan data *input* nonlinear ke beberapa ruang dimensi yang lebih tinggi, dimana data dapat dipisahkan secara linier, sehingga memberikan kinerja klasifikasi atau regresi yang besar [18]. Konsep sederhana SVM, yakni usaha mencari *hyperplane* terbaik yang berfungsi sebagai batas dari dua buah kelas berdasarkan *support vectors* yang merupakan vektor data berjarak paling mendekati *hyperplane* dan batas yang menyatakan *hyperplane* pemisah [19].

Beberapa penelitian sebelumnya yang menerapkan SVM dalam melakukan klasifikasi untuk mendeteksi penyakit kanker paru-paru diantaranya, penelitian Abdullah et al, 2021 [20]. Menggunakan dataset yang diperoleh dari UCI *machine learning repository* dengan menerapkan metode *k-fold cross-validation* dengan jumlah k, yaitu 10 dan menggunakan bantuan aplikasi Weka memberikan hasil akurasi, presisi, dan *recall* masing-masing sebesar 95,56%, 95,6%, dan 95,4%. Faisal et al, 2019 [21]. Yang melakukan perbandingan kinerja beberapa metode *machine learning*, yaitu *Naïve Bayes*, SVM, *Multi-Layer Perceptron*, *Decision Tree*, dan *Gradient-boosted Decision Tree* menggunakan metode *k-fold cross-validation* dengan jumlah k, yaitu 10 menggunakan bantuan alat *Rapid Miner* memberikan hasil akurasi, presisi, dan *recall* untuk SVM, yaitu 79,17%, 66,07% dan 79,71%. Penelitian lainnya yang dilakukan oleh Goel A dan Srivastava S, 2016 [22]. Menggunakan dataset yang diperoleh dari *Kent Ridge Biomedical dataset* dengan metode *k-fold cross-validation* dengan jumlah *k* sebanyak 10 memberikan hasil akurasi, presisi, dan *recall* masing-masing sebesar 92,61%, 92,6%, dan 92,6%.

Penelitian yang dilakukan oleh Sathiyapriya E dan Venila M, 2017 [23]. Menggunakan metode *percentage split* dalam mendeteksi penyakit kanker paru-paru memberikan hasil akurasi, presisi dan *recall* masing-masing sebesar 86%, 85%, dan 97,2%. Penelitian lainnya yang dilakukan oleh Thallam et al, 2020 [24]. Menggunakan metode *percentage split* dalam mendeteksi penyakit kanker paru-paru memperoleh hasil akurasi sebesar 95%, tetapi tidak memaparkan nilai presisi, dan *recall*. Penelitian yang dilakukan oleh Kareem et al, 2021 [25]. Menggunakan metode *percentage split* dengan *training data* sebesar 70% dan *testing data* sebesar 30% memberikan hasil akurasi sebesar 82,02%, tetapi tidak memaparkan nilai presisi dan *recall*.

2.2. K-Nearest Neighbor

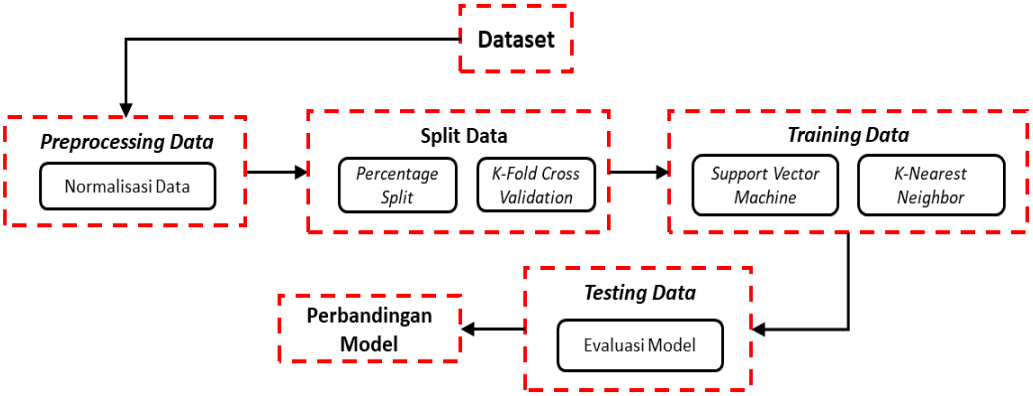
KNN merupakan algoritma klasifikasi terhadap objek berdasarkan data latih dengan menggunakan jarak terdekat atau kemiripan terhadap objek [26]. Sebuah titik pada ruang ini ditandai kelas tertentu. Kelas tersebut merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat dari titik tersebut yang mana dekat atau jauhnya tetangga biasanya dihitung menggunakan jarak *Euclidean*. KNN memiliki kelebihan, seperti tangguh terhadap *noise* dan efektif digunakan pada data latih berskala besar. Adapun kelemahan dari KNN, yakni perlu ditentukannya nilai k yang paling optimal yang menyatakan jumlah tetangga terdekat dan biaya komputasi yang cukup tinggi karena perhitungan jarak yang harus dilakukan pada setiap data latih [27].
Algoritma KNN banyak digunakan untuk klasifikasi dan regresi dalam pengenalan pola dan konsistensi dalam data. KNN adalah algoritma pembelajaran yang diawasi. KNN adalah jenis pembelajaran berbasis memori karena hipotesis dibangun secara langsung dari contoh pelatihan Para tetangga berasal dari kumpulan objek dari kelas yang diketahui. Jika $k = 1$, maka tunggal tetangga terdekat ditugaskan ke kelas. Secara umum skema pembobotan. Garis lurus akan selalu dihasilkan ketika ada jarak terpendek antara 2 tetangga dan jarak tersebut disebut jarak *Euclidean* [14].

Terdapat beberapa penelitian sebelumnya yang menerapkan algoritma KNN dalam melakukan klasifikasi menggunakan dataset penyakit kanker paru-paru diantaranya, penelitian Maleki et al, 2021 [28]. menggunakan data kanker paru-paru yang diperoleh dari situs *data world* dan metode yang digunakan adalah *k-fold cross-validation* dengan jumlah k, yaitu 10 memberikan hasil akurasi sebesar 96,40% namun tidak memaparkan nilai presisi dan *recall*. Penelitian Abdullah et al, 2021 [20]. Menggunakan dataset yang diperoleh dari UCI *machine learning repository* menggunakan metode *k-fold cross-validation* dengan jumlah *k* adalah 10 menghasilkan nilai akurasi, presisi, dan *recall* masing-masing, yaitu 89,65%, 89,8%,dan 89,7%. Patra R, 2020 [29]. Menggunakan

dataset yang diperoleh dari UCI *machine learning repository* menggunakan metode *k-fold cross-validation* dengan jumlah *k*, yaitu 10, memberikan hasil nilai akurasi, presisi, dan *recall* sebesar 75%, 73%, dan 75%.
Penelitian lainnya yang dilakukan oleh Sathiyapriya E dan Venila M, 2017 [23]. menggunakan metode *percentage split* menghasilkan nilai akurasi 86%, presisi 87,7%, dan *recall* 93,8%. Penelitian yang dilakukan oleh Thallam et al, 2020 [24]. Menggunakan metode *percentage split* dengan 80% *training data* dan 20% *testing data* menghasilkan nilai akurasi sebesar 97%, tetapi tidak memaparkan nilai presisi dan *recall*.

3. METODOLOGI PENELITIAN

Rangkaian proses yang akan dilakukan dalam penelitian klasifikasi penyakit kanker paru-paru menggunakan algoritma SVM dan KNN ditunjukkan pada Gambar 1.



Gambar 1. Rangkaian Proses yang Dilakukan dalam Penelitian

Berdasarkan Gambar 1, tahapan dilakukan pada penelitian ini adalah pengumpulan data, *preprocessing data*, pemisahan data menggunakan metode *percentage split* dan *k-fold cross validation*, melakukan *training* dan *testing* pada data, mengevaluasi model menggunakan *confusion matrix*, dan melakukan komparasi dari model yang dibentuk. Adapun perangkat keras yang digunakan dalam penelitian ini, yakni komputer dengan sistem operasi Windows 11 64 bit, processor Intel ® Core ™ i5-10210U @1.60GHz, 2.11GHz, RAM 8GB, memori 512GB, serta grafika NVIDIA GeForce MX350. Perangkat lunak yang digunakan dalam penelitian ini adalah *Google Colaboratory* dengan bahasa pemrograman *Python.*

3.1. Deskripsi Data

Data yang digunakan dalam penelitian ini adalah dataset yang diperoleh dari situs Kaggle (https://www.kaggle.com/datasets/h13380436001/h-lung-cancer) format csv. Data terdiri dari 309 data yang terbagi menjadi 270 untuk 'Ya' dan 39 untuk 'Tidak'. Atribut-atribut yang digunakan pada data kanker paru-paru beserta tipe atribut ditunjukkan pada Tabel 1.

Tabel 1. Atribut dan Nilai Atribut

Nama Atribut	**Tipe Atribut**	**Partisi Nilai**
Gender	Nominal	F: *Female*, M: *Male*
Age	Numerik	1: Tidak, 2: Ya
Smoking	Nominal	1: Tidak, 2: Ya
Yellow finger	Nominal	1: Tidak, 2: Ya
Anxiety	Nominal	1: Tidak, 2: Ya
Peer pressure	Nominal	1: Tidak, 2: Ya
Chronic disease	Nominal	1: Tidak, 2: Ya
Fatigue	Nominal	1: Tidak, 2: Ya
Allergy	Nominal	1: Tidak, 2: Ya
Wheezing	Nominal	1: Tidak, 2: Ya
Alcohol	Nominal	1: Tidak, 2: Ya
Consuming	Nominal	1: Tidak, 2: Ya
Coughing	Nominal	1: Tidak, 2: Ya
Shortness of breath	Nominal	1: Tidak, 2: Ya
Chest pain	Nominal	1: Tidak, 2: Ya
Lung cancer	Nominal	*Yes* dan *No*

Berdasarkan Tabel 1, terdapat 16 atribut pada dataset penyakit kanker paru-paru dimana 15 atribut bertipe nominal, sedangkan 1 atribut bertipe numerik. Pada atribut *lung cancer* terdapat dua kelas, yakni *yes* untuk yang mengidap penyakit kanker paru-paru dan *no* untuk yang tidak mengidap penyakit kanker paru-paru. Dapat diketahui, terdapat dua buah kelas pada penelitian klasifikasi penyakit kanker paru-paru, yakni kelas *yes* dan *no.*

3.2. Preprocessing Data

Data terdiri dari 16 atribut yang mana atribut-atribut yang tidak memiliki pengaruh akan dihapus. Pada penelitian klasifikasi penyakit kanker paru-paru digunakan 15 atribut dari yang semula 16 atribut, yakni *age, smoking, yellow finger, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol, consuming, coughing, shortness of breath, swallowing difficulty,* dan *chest pain*. Dari data yang digunakan, atribut-atribut memiliki *range* yang berbanding jauh satu sama lain, sehingga dilakukan normalisasi data untuk menyamakan *range* dari setiap atribut pada dataset. Normalisasi data dilakukan dengan menggunakan normalisasi *min-max* yang merupakan teknik penskalaan menggunakan nilai minimum dan maksimum dari fitur untuk mengubah skala nilai ke dalam suatu *range* dikarenakan *range* dari fitur-fitur yang ada terlampau jauh, biasanya *range* yang dipakai 0 hingga 1 atau -1 [30]. Adapun rumus perhitungan normalisasi *min-max* yang dapat dilihat pada Persamaan (1) [31].

$$Z^* = \left(\frac{Z - \min(Z)}{\max(Z) - \min(Z)}\right) \tag{1}$$

dimana, Z^* adalah hasil normalisasi *min-max*, Z adalah data sebelum dinormalisasi, $\min(Z)$ adalah nilai minimum dari keseluruhan data, dan $\max(Z)$ adalah nilai maksimum dari keseluruhan data.

3.3. Penerapan Klasifikasi

Setelah proses *preprocessing data*, dilakukan klasifikasi kanker paru-paru menggunakan algoritma SVM dan KNN. Sebelum klasifikasi dilakukan, data dibagi terlebih dahulu menggunakan metode *percentage split* dan metode *k-fold cross validation. Percentage split* adalah metode membagi data latih dan data uji berdasarkan persentase tertentu, sedangkan *k-fold cross validation* merupakan metode membagi data latih dan data uji ke dalam *n* subhimpunan [32]. Pada metode *percentage split,* data akan dibagi menjadi data latih dan data uji, sebesar 80% dan 20% secara acak. Kemudian, pada metode *k-fold cross validation* digunakan nilai k sebesar 10 yang berarti data dibagi menjadi 10 kelompok dengan 9 kelompok merupakan data latih dan 1 kelompok merupakan data latih dan data uji. Metode *percentage split* dan *k-fold cross validation* akan diterapkan pada masing-masing algoritma yang digunakan pada penelitian klasifikasi penyakit kanker paru-paru, yakni algoritma SVM dan KNN. Berdasarkan [33], langkah-langkah perhitungan SVM, yakni sebagai berikut:

1. Terdapat $\overrightarrow{x_i} \in \overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_n}$ dimana x_i merupakan data dari n atribut dan dua kelas, $y_i \in +1, -1$.
2. Asumsikan data linear dan kelas dapat dipisahkan oleh *hyperplane,* seperti yang ditunjukkan pada Persamaan (2).

   $$\vec{w}.\vec{x} + b = 0 \tag{2}$$

   Dari Persamaan (2) akan diperoleh Persamaan (3) dan Persamaan (4).

   $$\vec{w}.\vec{x} + b \geq 1, \text{ untuk kelas } +1 \tag{3}$$

   $$\vec{w}.\vec{x} + b \geq -1, \text{ untuk kelas } -1 \tag{4}$$

   dimana, $\vec{w}$ adalah bobot, b adalah bias, dan $\vec{x}$ adalah nilai input.
3. Mencari *hyperplane* pemisah dengan memaksimalkan jarak dari kedua kelas menggunakan pencarian titik minimum, seperti pada Persamaan (5).

   $$min_w \frac{1}{2}\left(||\vec{w}|| \right)^2 \tag{5}$$

   Dengan kendala pada Persamaan (6).

   $$y_i(\vec{w}.\overrightarrow{x_i} + b) - 1 \geq 0 \tag{6}$$

   dimana, $\vec{w}$ adalah bobot, b adalah bias, $\overrightarrow{x_i}$ adalah nilai input ke-i, dan y_i adalah kelas target ke-i.
4. Untuk mengatasi data nonlinear, digunakan kernel yang mentransformasikan *input space* ke dalam *feature space*. Pada penelitian ini digunakan kernel *polynomial*, seperti pada Persamaan (7).

   $$K(\vec{x}, \vec{y}) = (\vec{x}.\vec{y} + c)^d \tag{7}$$

   dimana $\vec{x}$ dan $\vec{y}$ adalah vektor dari ruang fitur, d adalah derajat, dan c adalah parameter bebas.

Selain menerapkan algoritma SVM, diterapkan juga algoritma KNN yang mana kedua algoritma tersebut akan dibandingkan untuk mengetahui algoritma yang lebih baik dalam prediksi kanker paru-paru. Berdasarkan [34], langkah-langkah melakukan perhitungan menggunakan algoritma KNN, yakni sebagai berikut:

1. Menentukan nilai k yang digunakan dimana k adalah jumlah tetangga terdekat yang akan digunakan selama proses *training* data.
2. Melakukan perhitungan jarak antara data latih dan data uji menggunakan jarak *Euclidean*, seperti pada Persamaan (8).

   $$Euc = \sqrt{\left(\sum_{i=1}^{n}(p_i - q_i)^2\right)} \tag{8}$$

   dimana, p_i adalah data latih, q_i adalah data uji, i adalah variabel data, n adalah ukuran data.
3. Mengurutkan jarak yang terbentuk.
4. Menentukan jarak terdekat sampai urutan k.
5. Memasangkan kelas yang bersesuaian.
6. Mencari jumlah kelas dari tetangga terdekat, lalu menetapkan kelas tersebut sebagai kelas data yang akan dievaluasi.
7.

3.4. Confusion Matrix

Dalam melakukan evaluasi model klasifikasi penyakit kanker paru-paru, digunakan *confusion matrix.* Berdasarkan [27], *confusion matrix* untuk klasifikasi dua kelas dapat dilihat seperti pada Tabel 2.

Tabel 2. *Confusion Matrix*

Kelas		**Prediksi**	
		True	**False**
Aktual	**True**	TP	FN
	False	FP	TN

Keterangan:
TP = Jumlah *record* positif yang diklasifikasikan sebagai positif
FP = Jumlah *record* negatif yang diklasifikasikan sebagai positif
FN = Jumlah *record* positif yang diklasifikasikan sebagai negatif
TN = Jumlah *record* negatif yang diklasifikasikan sebagai negatif

Terdapat beberapa ukuran evaluasi kinerja yang digunakan dalam klasifikasi nilai-nilai dalam *confusion matrix*, yaitu akurasi, presisi, dan *recall* [35]. Berdasarkan [36], rumus untuk menghitung performa dari *confusion matrix* dapat dilihat pada Persamaan (9), Persamaan (10), dan Persamaan (11).

$$asi = \frac{TP+TN}{TP+FP+TN+FN} \quad (9)$$

$$si = \frac{TP}{FP+TP} \quad (10)$$

$$l = \frac{TP}{FN+TP} \quad (11)$$

4. HASIL DAN PEMBAHASAN

Atribut yang menjadi label klasifikasi adalah kanker paru-paru yang memiliki dua kelas, yakni ya dan tidak. Metode yang digunakan untuk melakukan *split data* adalah *percentage split* dengan ukuran split sebesar 80% untuk data uji dan 20% untuk data latih dan *k-fold cross validation* dengan k yang dipilih adalah 10 yang berarti data dibagi menjadi 10 kelompok dengan 9 kelompok merupakan data latih dan 1 kelompok merupakan data latih dan data uji.

4.1. Support Vector Machine

Diperoleh bahwa algoritma SVM dengan metode *percentage split* dapat memprediksi sebanyak 59 data dalam kelas yang benar, namun 3 data diprediksi dalam kelas yang salah. TP, TN, FP, dan FN yang dihasilkan adalah 55, 24, 2, dan 1. Akurasi yang diperoleh pada SVM dengan metode *percentage split* sebesar 95,16%. Kemudian, nilai *recall* yang dihasilkan untuk kedua kelas, yakni 98% untuk kelas ya dan 67% untuk kelas tidak. Nilai presisi untuk kedua kelas, yaitu 96% untuk kelas ya dan 80% untuk kelas tidak. Algoritma SVM menggunakan metode *k-fold cross validation* dapat memprediksi sebanyak 280 data dalam kelas yang benar, namun 29 data diprediksi dalam kelas yang salah. TP, TN, FP, dan FN yang dihasilkan adalah 253, 27, 12, dan 17. Akurasi yang diperoleh pada SVM menggunakan metode *k-fold cross validation* sebesar 90,61%. Kemudian, nilai *recall* yang dihasilkan menggunakan algoritma SVM untuk kedua kelas, yakni 94% untuk kelas ya dan 69% untuk kelas tidak. Nilai presisi untuk kedua kelas, yaitu 95% untuk kelas ya dan 61% untuk kelas tidak.

4.2. K-Nearest Neighbor

Pada algoritma KNN, dilakukan percobaan dengan 9 nilai k, yakni 1, 2, 3, 4, 5, 6, 7, 8, dan 9 untuk mengetahui dengan menggunakan metode *percentage split* dan *k-fold cross validation*. Hasil dari akurasi, presisi, dan *recall* dari 9 nilai *k* dapat dilihat pada Tabel 3.

Tabel 3. Hasil Akurasi, Presisi, dan *Recall* KNN dengan *Percentage Split* dan *K-Fold Cross Validation*

K	**Target Kelas**	***Percentage Split***			***K-Fold Cross Validation***		
		Presisi	***Recall***	**Akurasi**	**Presisi**	***Recall***	**Akurasi**
1	Ya	98%	93%	91,94%	61%	94%	90,29%
	Tidak	60%	86%		95%	64%	
2	Ya	92%	92%	87,10%	57%	91%	89,97%
	Tidak	60%	60%		97%	82%	
3	Ya	95%	96%	91,94%	94%	96%	90,61%
	Tidak	60%	50%		66%	54%	
4	Ya	98%	89%	88,71%	94%	92%	88,03%
	Tidak	40%	80%		52%	62%	
5	Ya	91%	94%	87,10%	92%	94%	88,03%
	Tidak	67%	55%		53%	44%	
6	Ya	93%	95%	88,71%	94%	93%	89,00%

	Tidak	40%	33%		56%	59%	
7	Ya	94%	94%	90,32%	92%	96%	89,00%
	Tidak	62%	62%		60%	38%	
8	Ya	94%	94%	90,32%	94%	96%	90,61%
	Tidak	62%	62%		66%	54%	
9	Ya	88%	98%	87,10%	91%	96%	88,35%
	Tidak	83%	42%		56%	36%	

Dari Tabel 3 dapat dilihat bahwa nilai k terbaik terdapat pada $k = 3$ sehingga nilai akurasi, presisi, dan *recall* metode *percentage split* dan *k-fold cross validation* pada KNN menggunakan $k = 3$ untuk perbandingan terhadap algoritma SVM. Diperoleh, bahwa dengan $k = 3$ algoritma KNN metode *percentage split* dapat memprediksi sebanyak 57 data dalam kelas yang benar, namun 5 data diprediksi dalam kelas yang salah. TP, TN, FP, dan FN yang dihasilkan adalah 54, 3, 3, dan 2. Akurasi yang diperoleh pada KNN menggunakan metode *percentage split* sebesar 91,94%. Kemudian, nilai *recall* yang dihasilkan untuk kedua kelas, yakni 96% untuk kelas ya dan 50% untuk kelas tidak. Nilai presisi untuk kedua kelas, yaitu 95% untuk kelas ya dan 60% untuk kelas tidak. Kemudian, dengan $k = 3$ algoritma KNN metode *k-fold cross validation* dapat memprediksi sebanyak 280 data dalam kelas yang benar, namun 28 data diprediksi dalam kelas yang salah. TP, TN, FP, dan FN yang dihasilkan adalah 259, 21, 18, dan 11. Akurasi yang diperoleh pada KNN dengan metode *k-fold cross validation* sebesar 90,61%. Kemudian, nilai *recall* yang dihasilkan untuk kedua kelas, yakni 96% untuk kelas ya dan 54% untuk kelas tidak. Nilai presisi untuk kedua kelas, yaitu 94% untuk kelas ya dan 66% untuk kelas tidak.

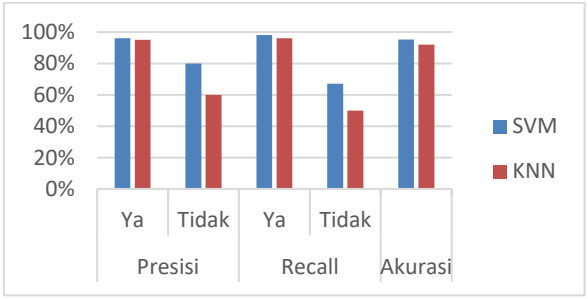
4.3. Perbandingan Hasil Kedua Algoritma

Hasil dari algoritma SVM dan KNN menggunakan metode *percentage split* menunjukkan bahwa kedua metode memiliki kinerja yang baik dalam dalam melakukan prediksi penyakit kanker paru-paru. Perbandingan dari hasil pengukuran menggunakan algoritma SVM dan KNN dengan metode *percentage split* dapat dilihat pada Tabel 4.

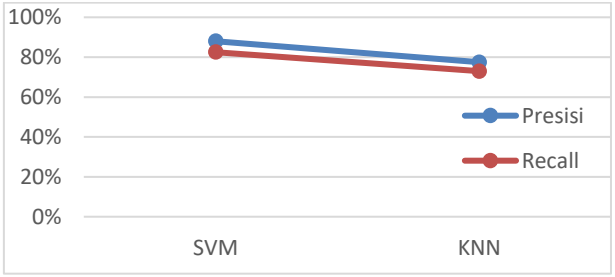
Tabel 4. Nilai Akurasi, Presisi, dan *Recall* pada SVM dan KNN Metode *Percentage Split*

Algoritma	**Presisi (%)**		***Recall* (%)**		**Akurasi (%)**
	Kelas		**Kelas**		
	Ya	**Tidak**	**Ya**	**Tidak**	
SVM	96%	80%	98%	67%	95,16%
KNN	95%	60%	96%	50%	91,94%

Dari Tabel 4, algoritma SVM menghasilkan akurasi lebih tinggi dibandingkan algoritma KNN. Baik menggunakan algoritma SVM maupun KNN, hasil akurasi yang dihasilkan sama-sama di atas 90%. Hasil presisi, *recall*, dan akurasi dari algoritma SVM dan KNN menggunakan *percentage split* dapat dilihat secara ringkas pada Gambar 2. Kemudian untuk nilai rata-rata presisi dan *recall* dari seluruh kelas dengan metode *percentage split* dapat dilihat pada Gambar 3.



Gambar 2. Nilai Presisi dan *Recall* Hasil Prediksi SVM dan KNN Metode *Percentage Split*



Gambar 3. Rata-Rata Seluruh Nilai Presisi dan *Recall* Hasil Prediksi SVM dan KNN Metode *Percentage Split*

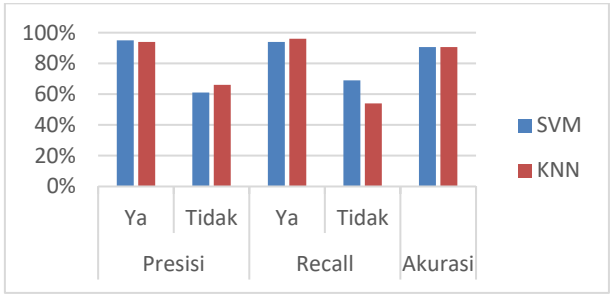
Dari Gambar 3 dapat dilihat nilai presisi dan *recall* yang diperoleh SVM maupun KNN menggunakan metode *percentage split* kurang lebih sama sehingga kedua algoritma tersebut dapat digunakan sebagai prediksi kanker paru-paru. Sama halnya seperti pada metode *percentage split*, hasil dari algoritma SVM dan KNN menggunakan metode *k-fold cross validation* memperlihatkan bahwa

kedua metode memiliki kinerja yang baik dalam melakukan prediksi penyakit kanker paru-paru. Perbandingan dari hasil pengukuran menggunakan algoritma SVM dan KNN dengan metode *k-fold cross validation* terdapat pada Tabel 5.

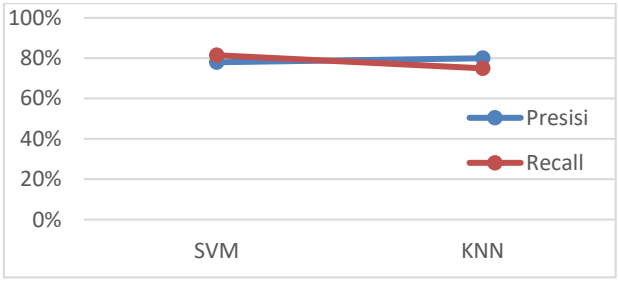
Tabel 5. Nilai Akurasi dan Presisi SVM dan KNN Metode *K-Fold Cross Validation*

Algoritma	**Presisi (%)**		***Recall* (%)**		**Akurasi (%)**
	Kelas		**Kelas**		
	Ya	**Tidak**	**Ya**	**Tidak**	
SVM	95%	61%	94%	69%	90,61%
KNN	94%	66%	96%	54%	90,61%

Dari Tabel 5, algoritma SVM dan KNN masing-masing memiliki nilai akurasi yang sama. Hasil presisi, *recall*, dan akurasi dari algoritma SVM dan KNN menggunakan *k-fold cross validation* dapat dilihat secara ringkas pada Gambar 4. Kemudian untuk nilai rata-rata presisi dan *recall* dari seluruh kelas menggunakan metode *k-fold cross validation* dapat dilihat pada Gambar 5.



Gambar 4. Nilai Presisi dan *Recall* Hasil Prediksi SVM dan KNN Metode *K-Fold Cross Validation*



Gambar 5. Rata-rata Seluruh Nilai Presisi dan *Recall* Hasil Prediksi SVM dan KNN Metode *K-Fold Cross Validation*

Dari Gambar 5 dapat dilihat nilai rata-rata presisi menggunakan metode *percentage split* pada algoritma KNN lebih tinggi dibandingkan pada algoritma SVM. Kemudian, nilai rata-rata *recall* pada algoritma SVM lebih tinggi dibandingkan algoritma KNN dengan menggunakan metode *k-fold cross validation*. Berdasarkan keseluruhan hasil yang diperoleh, baik menggunakan algoritma SVM maupun algoritma KNN, kedua algoritma mampu melakukan klasifikasi kanker paru-paru dengan baik dan perbedaan hasil yang diperoleh tidak terlalu jauh. Hal tersebut menunjukkan bahwa, algoritma SVM dan KNN sama-sama baik digunakan dalam melakukan deteksi dini pada penyakit kanker paru-paru. Namun, secara lebih rinci algoritma SVM memiliki performa yang lebih baik dibandingkan algoritma KNN pada klasifikasi kanker paru-paru.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan penelitian yang dilakukan, algoritma SVM dan KNN sama-sama menghasilkan nilai akurasi di atas 90% yang menunjukkan bahwa kedua algoritma tersebut memiliki performa yang baik dalam melakukan prediksi kanker paru-paru pada penelitian ini. Dengan membandingkan kedua algoritma dan kedua metode pembagian data yang digunakan, diperoleh bahwa algoritma SVM memiliki nilai akurasi, presisi, dan *recall* yang lebih tinggi daripada algoritma KNN baik menggunakan metode *percentage split* maupun metode *k-fold cross validation* walaupun hasil dari kedua algoritma hampir mirip dan tidak jauh berbeda. Hal tersebut menunjukkan bahwa algoritma SVM lebih baik dalam melakukan klasifikasi kanker paru-paru dibandingkan algoritma KNN.

5.2. Saran

Saran untuk penelitian berikutnya adalah dapat melakukan mengembangkan deteksi kanker paru-paru yang telah dibuat pada penelitian ini sehingga menghasilkan akurasi, presisi, dan *recall* yang lebih baik lagi.

DAFTAR PUSTAKA

[1] N. M. Aljamali, W. K. N. Al-Qraawy, and T. A. Helal, "Review on Carcinogens Materials in Chemical Laboratories," *Int. J. Mol. Biol. Biochem.*, vol. 4, no. 1, pp. 17–25, 2022.

[2] J. A. Barta, C. A. Powell, and J. P. Wisnivesky, "Global Epidemiology of Lung Cancer," *Ann. Glob. Heal.*, vol. 85, no. 1, p. 8, Jan. 2019, doi: 10.5334/aogh.2419.

[3] A. Desiani, Erwin, B. Suprihatin, S. Yahdin, A. I. Putri, and F. R. Husein, "Bi-Path Architecture of CNN Segmentation and Classification Method for Cervical Cancer Disorders Based on Pap-smear Images," *Int. J. Comput. Sci.*, vol. 48, no. 3, 2021.

[4] Ş. Yaşar, A. K. Arslan, C. Çolak, and S. Yoloğlu, "A Developed Web Based Software Can Easily Fulfill the Assumptions of Correlation, Classification and Regression Tasks in Data Processing," in *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1–5. doi: 10.1109/IDAP.2019.8875914.

[5] M. Onel, C. A. Kieslich, Y. A. Guzman, C. A. Floudas, and E. N. Pistikopoulos, "Big Data Approach to Batch Process Monitoring: Simultaneous Fault Detection and Diagnosis using Nonlinear Support Vector Machine based Feature Selection," *Comput. Chem. Eng.*, vol. 115, pp. 46–63, 2018, doi: https://doi.org/10.1016/j.compchemeng.2018.03.025.

[6] R. I. Borman, F. Rossi, Y. Jusman, A. A. A. Rahni, S. D. Putra, and A. Herdiansah, "Identification of Herbal Leaf Types Based on Their Image Using First Order Feature Extraction and Multiclass SVM Algorithm," in *2021 1st International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2021, pp. 12–17. doi: 10.1109/ICE3IS54102.2021.9649677.

[7] S. R. A. Ahmed, I. Al-Barazanchi, A. Mhana, and H. R. Abdulshaheed, "Lung Cancer Classification using Data Mining and Supervised Learning Algorithms on Multi-Dimensional Data Set," *Period. Eng. Nat. Sci.*, vol. 7, no. 2, pp. 438–447, 2019, doi: 10.21533/pen.v7i2.483.

[8] B. K. Francis and S. S. Babu, "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach," *J. Med. Syst.*, vol. 43, no. 6, 2019, doi: 10.1007/s10916-019-1295-4.

[9] F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," in *International Conference on Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 414–418. doi: 10.1109/ICOEI.2018.8553959.

[10] Y. R. Nugraha, A. P. Wibawa, and I. A. E. Zaeni, "Particle Swarm Optimization-Support Vector Machine (PSO-SVM) Algorithm for Journal Rank Classification," in *2019 2nd International Conference of Computer and Informatics Engineering (IC2IE)*, 2019, pp. 69–73. doi: 10.1109/IC2IE47452.2019.8940822.

[11] S. Widaningsih and S. Yusuf, "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: 10.35957/jatisi.v9i3.859.

[12] Y. Wang, Z. Pan, and Y. Pan, "A Training Data Set Cleaning Method by Classification Ability Ranking for the K-Nearest Neighbor Classifier," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 5, pp. 1544–1556, 2020, doi: 10.1109/TNNLS.2019.2920864.

[13] S. Bharati, P. Podder, R. Mondal, A. Mahmood, and M. Raihan-Al-Masud, "Comparative Performance Analysis of Different Classification Algorithm for the Purpose of Prediction of Lung Cancer," in *International Conference on Intelligent Systems Design and Applications*, 2020, vol. 941, pp. 447–457. doi: 10.1007/978-3-030-16660-1_44.

[14] R. Devika, S. V. Avilala, and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease Prediction using Naive Bayes, KNN and Random Forest," in *International Conference on Computing Methodologies and Communication (ICCMC)*, 2019, pp. 679–684. doi: 10.1109/ICCMC.2019.8819654.

[15] K. Taunk, S. De, S. Verma, and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.

[16] S. A. Taher, K. A. Akhter, and K. M. A. Hasan, "N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine," in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1–5. doi: 10.1109/ICBSLP.2018.8554716.

[17] S. Huang, C. A. I. Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and X. U. Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics*, vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[18] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," in *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, 2019, pp. 24–28. doi: 10.1109/ISS1.2019.8908018.

[19] L. Yahaya, N. D. Oye, and E. J. Garba, "A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques," *Am. J. Artif. Intell.*, vol. 4, no. 1, pp. 20–29, 2020, doi: 10.11648/j.ajai.20200401.12.

[20] D. M. Abdullah, A. M. Abdulazeez, and A. B. Sallow, "Lung cancer Prediction and Classification based on Correlation Selection method Using Machine Learning Techniques," *Qubahan Acad. J.* , vol. 1, no. 2, pp. 141–149, 2021, doi: 10.48161/Issn.2709-8206.

[21] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," *2018 3rd Int. Conf. Emerg. Trends Eng. Sci. Technol. ICEEST 2018*, pp. 1–4, 2019, doi: 10.1109/ICEEST.2018.8643311.

[22] A. Goel and S. K. Srivastava, "Role of kernel parameters in performance evaluation of SVM," *Proc. - 2016 2nd Int. Conf. Comput. Intell. Commun. Technol. CICT 2016*, pp. 166–169, 2016, doi: 10.1109/CICT.2016.40.

[23] E. Sathiyapriya and S. Venila, "A Study on Classification Algorithms and Performance Analysis of Data Mining using Cancer Data to Predict Lung Cancer Disease," *Int. J. New Technol. Res.*, vol. 3, no. 8, pp. 88–93, 2017.

[24] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," *Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020*, pp. 1285–1292, 2020, doi: 10.1109/ICECA49313.2020.9297576.

[25] H. F. Kareem, M. S. AL-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 21, no. 3, pp. 1731–1738, 2021, doi: 10.11591/ijeecs.v21.i3.pp1731-1738.

[26] R. R. A. Siregar, Z. U. Siregar, and R. Arianto, "Klasifikasi Sentiment Analysis Pada Komentar Peserta Diklat Menggunakan Metode K-Nearest Neighbor," *Kilat*, vol. 8, no. 1, pp. 81–92, 2019, doi: 10.33322/kilat.v8i1.421.

[27] J. Riany, M. Fajar, and M. P. Lukman, "Penerapan Deep Sentiment Analysis pada Angket Penilaian Terbuka Menggunakan K-Nearest Neighbor," *Sisfo*, vol. 6, no. 1, pp. 147–156, 2016, doi: 10.24089/j.sisfo.2016.09.011.
[28] N. Maleki, Y. Zeinali, and S. T. A. Niaki, "A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection," *Expert Syst. Appl.*, vol. 164, no. July 2019, p. 113981, 2021, doi: 10.1016/j.eswa.2020.113981.
[29] R. Patra, *Prediction of lung cancer using machine learning classifier*, vol. 1235 CCIS. Springer Singapore, 2020. doi: 10.1007/978-981-15-6648-6_11.
[30] F. Adams, R. A. D. Anggoro, M. B. Satria, A. W. Oktavia, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma Naïve Bayes, Decision Tree, dan Support Vector Machine," in *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, 2021, pp. 260–268.
[31] R. A. Wijayanti, M. T. Furqon, and S. Adinugroho, "Penerapan Algoritma Support Vector Machine Terhadap Klasifikasi Tingkat Risiko Pasien Gagal Ginjal," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, pp. 3500–3507, 2018.
[32] A. Septiarini, R. Saputra, A. Tejawati, and M. Wati, "Deteksi Sarung Samarinda Menggunakan Metode Naive Bayes Berbasis Pengolahan Citra," *J. Rekayasa Sist. dan Teknol. Inf.*, vol. 5, no. 5, pp. 927–935, 2021.
[33] S. A. Naufal, Adiwijaya, and W. Astuti, "Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, pp. 162–168, 2020, doi: 10.30865/jurikom.v7i1.2014.
[34] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis Performa Metode KNN pada Dataset Pasien Pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.
[35] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 5, no. 12, p. 12, 2020, doi: 10.1007/s41133-020-00032-0.
[36] R. Novendri, A. S. Callista, D. N. Pratama, and C. E. Puspita, "Sentiment Analysis of YouTube Movie Trailer Comments Using Naïve Bayes," *Bull. Comput. Sci. Electr. Eng.*, vol. 1, no. 1, pp. 26–32, 2020, doi: 10.25008/bcsee.v1i1.5.