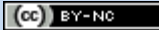



Analisis penerapan metode *K-Nearest Neighbor* (K-NN) pada dataset citra penyakit malaria

Aisyah^{a,1}, Siska Anraeni^{a,2}

^a Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.05, Makassar dan 90231, Indonesia

¹ 13020160081@umi.ac.id; ² siska.anraeni@umi.ac.id;

INFORMASI ARTIKEL	ABSTRAK
Diterima : 12 – 01 – 2022 Direvisi : 20 – 02 – 2022 Diterbitkan : 31 – 03 – 2022	Penyakit malaria yang disebabkan oleh parasit protozoa dari <i>genus plasmodium</i> dapat menginfeksi sel darah merah. karena penyakit ini dapat menimbulkan gangguan kesehatan dan menjadi masalah kesehatan masyarakat khususnya wilayah tropis serta dapat menurunkan produktivitas kerja bahkan dapat mengakibatkan kematian. Penelitian ini menggunakan sebanyak 1000 dataset citra penyakit malaria terinfeksi dan malaria tidak terinfeksi. Dataset tersebut dikelola oleh U.S National Library of Medicine College Hospital hingga tahun 2018. Tujuan dari penelitian ini untuk menganalisis performa (akurasi, presisi, <i>recall</i> dan <i>f-measure</i>) pada dataset citra penyakit malaria terinfeksi dan malaria tidak terinfeksi. Tahapan yang dilakukan yaitu: deteksi tepi <i>canny</i> , ekstraksi fitur hu moment invariant, klasifikasi metode K-NN, serta menggunakan berbagai simulasi rasio dataset dan nilai K=2 sampai K=10. Hasil penelitian dari simulasi rasio dataset 80:20 memperoleh nilai akurasi 84%, presisi 88%, recall 92% dan <i>f-measure</i> 90% dengan menggunakan nilai K=6.
Kata Kunci: deteksi tepi <i>canny</i> ekstraksi fitur hu <i>moment invariant</i> <i>k-nearest neighbor</i> penyakit malaria analisis performa	 

I. Pendahuluan

Malaria adalah salah satu penyakit menular yang menginfeksi sel darah merah ini disebabkan oleh parasit protozoa dari genus *plasmodium* dan ditularkan melalui gigitan nyamuk *Anopheles betina*. Diantara 170 jenis *plasmodium* ada empat penyebab malaria pada manusia, yaitu: *plasmodium falciparum*, *plasmodium malaria*, *plasmodium ovale* dan *plasmodium vivax*. Di Indonesia, spesies *plasmodium* yang sering ditemui adalah *plasmodium falciparum* (malaria tropika) dan *plasmodium vivax* (malaria tertiana). Malaria termasuk masalah kesehatan masyarakat yang penting dan endemis hampir semua wilayah luar Jawa dan Bali.[1] Penyakit malaria menjadi masalah kesehatan masyarakat yang penting, karena penyakit ini dapat menimbulkan gangguan Kesehatan dan dapat menurunkan produktivitas kerja bahkan dapat mengakibatkan kematian.[2]

Menurut Kementerian Kesehatan RI, “Penyakit malaria tergolong suatu masalah kesehatan yang besar bagi penduduk wilayah negara tropis termasuk Indonesia. Angka kesakitan malaria secara nasional selama tahun 2009-2018 menurun yaitu dari 1,8 per 1.000 penduduk pada tahun 2009 menjadi 0,84 per 1.000 penduduk pada tahun 2018”. [3]

Papua merupakan provinsi dengan API tertinggi, yaitu 52,99 per 1.000 penduduk. Angka ini sangat tinggi, jika dibandingkan dengan provinsi lainnya. Tiga provinsi dengan API per 1.000 penduduk tertinggi lainnya, Papua Barat (8,49), Nusa Tenggara Timur (3,42) dan Maluku (1,16). Sebanyak 66% kasus berasal dari Papua, Papua Barat, Nusa Tenggara Timur dan Maluku.

K-Nearest Neighbor (K-NN) dilakukan dengan mencari kelompok objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing [4]–[6]. Performa dari klasifikasi dapat dievaluasi dengan menghitung nilai akurasi, pada penelitian sebelumnya telah melakukan pengukuran nilai akurasi pada objek 120 data penyakit malaria, dimana data latihnya 80 dan data ujinya sebanyak 40.[7] Kemudian memperoleh nilai akurasi sebesar 95% ketika nilai k adalah 5 dan 7.[8]

Penelitian ini mencoba untuk mencari nilai performa terbaik dari berbagai kombinasi dataset serta beragam nilai K [5], [7], [9]–[11], yang terdiri dari rasio perbandingan 20:80 (200 citra penyakit malaria terinfeksi, 800 citra penyakit malaria tidak terinfeksi), 80:20 (800 citra penyakit malaria terinfeksi, 200 citra penyakit malaria tidak terinfeksi), 50:50 (500 citra penyakit malaria terinfeksi, 500 citra penyakit malaria tidak terinfeksi).

II. Metode

A. Dataset

Terdapat 3000 dataset citra penyakit malaria menggunakan metode K-Nearest Neighbor (K-NN) dengan nilai $K = 2, 3, 4, 5, 6, 7, 8, 9, 10$. Nilai inilah yang akan diuji dengan data terdekat dimulai dari nilai $K=2$ sampai 10 sehingga tingkat akurasi, presisi, *recall* dan *f-measure* bisa disimpulkan bahwa data uji tergolong sebagai penyakit malaria terinfeksi atau tidak terinfeksi. Didalam penelitian ini terdapat 3 macam kombinasi dataset, yaitu rasio perbandingan dataset 20:80, 80:20 dan 50:50. Penelitian ini mencoba mencari nilai performa terbaik dari berbagai kombinasi dataset serta beragam nilai K .

Dataset citra penyakit malaria terinfeksi dan tidak terinfeksi diperoleh dari Kaggle dataset library. Kaggle merupakan sebuah situs penyedia jasa repository dataset, dimana dataset citra penyakit malaria tersebut di kelola oleh U.S National Library of Medicine College Hospital. U.S National Library of Medicine College Hospital mengumpulkan dataset, kemudian dikelola hingga tahun 2018 kemudian diperbaharui di Kaggle pada tanggal 05 Desember 2018. Pada [Tabel 1](#). menunjukkan contoh dataset citra penyakit malaria.

Tabel 1. Dataset Penyakit Malaria

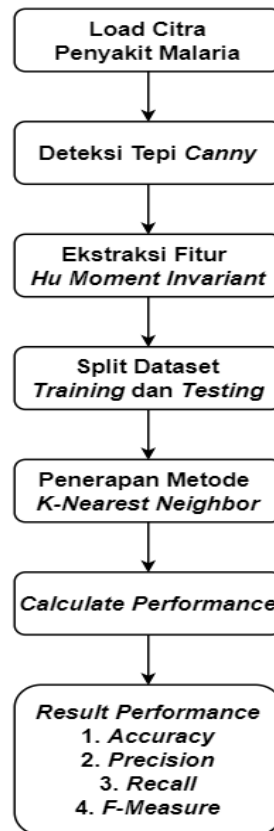
Id	Citra	Ket	Id	Citra	Ket
0		Terinfeksi	5		Terinfeksi
1		Tidak Terinfeksi	6		Tidak Terinfeksi
2		Terinfeksi	7		Terinfeksi
3		Terinfeksi	8		Terinfeksi
4		Tidak Terinfeksi	9		Terinfeksi
....
....
996		Tidak Terinfeksi	998		Terinfeksi
997		Terinfeksi	999		Tidak Terinfeksi

B. Perancangan Sistem

Dalam penelitian ini secara garis besar langkah-langkahnya terdiri dari load dataset, deteksi tepi *canny*, ekstraksi fitur hu *moment invariant*, memilah dataset menjadi data training dan data *testing* kemudian tahap persentasi data *training* dalam implementasi K-NN pada data *testing* kemudian menghitung nilai akurasi, *presisi*, *recall*, and *f-measure*.

1) Perancangan Proses

Adapun alur perancangan proses ang diawali pembacaan citra untuk penyakit malaria sampai dengan proses untuk mendapatkan hasil performance (*accuracy*, *precision*, *recall* and *F-measure*) yang akan dilakukan ditunjukkan pada [Gambar 1](#).



Gambar 1. Perancangan Proses

2) Tahapan Penelitian

a) Load Citra Penyakit Malaria

Load citra penyakit malaria terinfeksi dan tidak terinfeksi, dimana proses ini akan *load* citra penyakit malaria yang terbagi dalam 3 kombinasi dataset yaitu dengan rasio perbandingan dataset 20:80 (200 citra penyakit malaria terinfeksi dan 800 citra penyakit malaria tidak terinfeksi), 80:20 (800 citra penyakit malaria terinfeksi dan 200 citra penyakit malaria tidak terinfeksi), dan 50:50 (500 citra penyakit malaria terinfeksi dan 500 citra penyakit malaria tidak terinfeksi).

b) Deteksi Tepi Canny

Pada tahap ini deteksi tepi akan mengoptimalkan pendeteksian tepi pada citra yang memiliki *noise*. Algoritma deteksi tepi *canny* diantaranya: memaksimalkan nilai *Signal to Noise Ratio* (SNR) sehingga semua tepi dapat terdeteksi dengan baik atau tidak ada yang hilang, tepi yang terdeteksi berada pada posisi sebenarnya atau dengan kata lain bahwa jarak antara posisi sebenarnya adalah seminimum mungkin (idealnya adalah 0), hanya satu respon untuk sebuah tepi. Artinya deteksi tidak memberikan tepi yang bukan tepi sebenarnya. Dalam penelitian ini, *library* yang digunakan dalam deteksi tepi *canny* yaitu *library cv2*.

c) Ekstraksi Fitur Hu Moment Invariant

Pada tahap ekstraksi fitur hu moment invariant menggunakan *library cv2* yang dimana terjadinya proses perubahan data citra lalu dikonversi menjadi data numerik, kemudian menghasilkan 7 *array* nilai label mulai dari H1-H7, serta label target. Kemudian akan di *export* dalam *file* format *.csv* (*Comma Separated Values*) yang disimpan kedalam *google drive*.

d) Metode K-Nearest Neighbor (K-NN)

Dalam proses pengklasifikasian ini untuk mengetahui apakah termasuk citra penyakit malaria terinfeksi atau malaria tidak terinfeksi menggunakan rumus dari metode *K-Nearest Neighbor* (K-NN) yang menggunakan *library scikit learn* [12], [13].

e) Hasil Nilai Performa (Akurasi, Presisi, Recall dan F-Measure)

Sedangkan dalam proses ini, kita melakukan perhitungan hasil nilai performa yang terdiri dari nilai akurasi, presisi, *recall* dan *f-measure* pada citra penyakit malaria terinfeksi dan tidak terinfeksi [12], [14]–[17].

- Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual. rumus akurasi. Pada persamaan 2.
- Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. rumus presisi ditunjukkan pada persamaan 3.
- *Recall* didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Rumus *Recall* diuraikan pada persamaan 4.
- *F-measure* adalah *harmonic mean* antara nilai presisi dan recall, *F-measure* juga kadang disebut dengan nama *F-measure*. Rumus *F-measure* dijabarkan pada persamaan 5 [7], [18]–[21].

$$AKURASI = ((TP + TN))/((TP + TN + FP + FN)) \quad (1)$$

$$PRESISI = TP/(TP + FP) \quad (2)$$

$$RECALL = TP/(TP + FN) \quad (3)$$

$$F - measure = 2 (Presisi \times Recall)/(Presisi + Recall) \quad (4)$$

Keterangan :

TP : *True Positive*

TN : *True Negative*

FP : *False Positive*

FN : *False Negative*

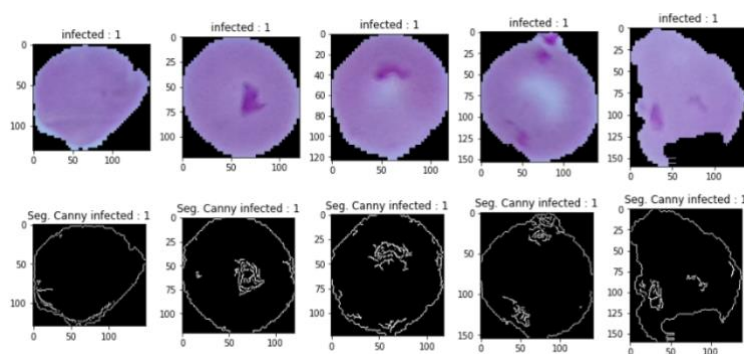
III. Hasil dan Pembahasan

. Bagian ini tempat menuliskan hasil penelitian yang dijabarkan secara detail, jelas dan terurut. Hasil penelitian disajikan dalam bentuk tabel, grafik atau ilustrasi lain dan disertai dengan pembahasan yang disajikan secara terstruktur dan sistematis. Uraian performansi, kelemahan, dan kelebihan dari hasil penelitian harus dijelaskan.

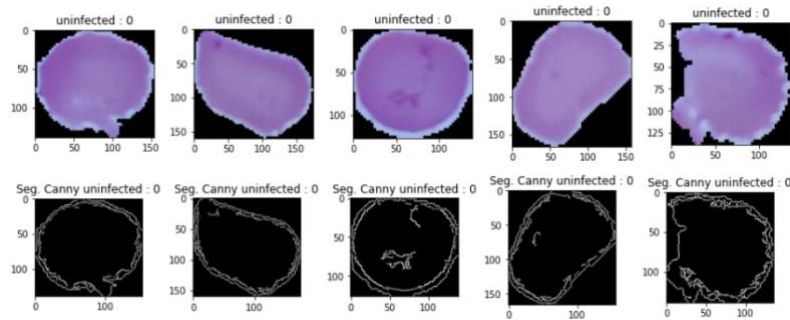
A. Implementasi

1) Implementasi Deteksi Tepi Canny

Implementasi deteksi tepi *canny* pada tahap ini merupakan contoh citra asli penyakit malaria terinfeksi dan malaria tidak terinfeksi diubah menggunakan deteksi tepi *canny*. Contoh implementasinya ditunjukkan pada Gambar 2 dan Gambar 3.



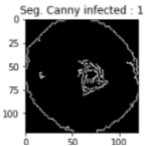
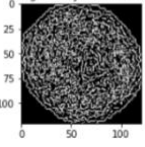
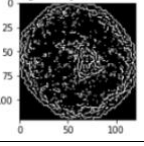
Gambar 2. Deteksi Tepi Citra Malaria Terinfeksi



Gambar 3. Deteksi Tepi Citra Malaria Tidak Terinfeksi

Pada deteksi tepi *canny* ini, peneliti menggunakan nilai *threshold* 80, 30. Nilai 80 merupakan nilai *threshold* I yang dimana letaknya dibagian deteksi tepi luar (garis pertama dari citra penyakit malaria). Sedangkan nilai 30 merupakan nilai *threshold* II yang dimana letaknya dibagian deteksi tepi bagian dalam (yang bisa membedakan mana penyakit malaria terinfeksi dan malaria tidak terinfeksi). Pada Tabel 2. menunjukkan perbandingan nilai *threshold* pada citra penyakit malaria:

Tabel 2. Prbandingan Nilai *Threshold* Citra Penyakit Malaria

No	Hasil Citra	Nilai T1	Nilai T2
1		80	30
2		0	11
3		25	25

Pada Tabel 2. dapat diberikan kesimpulan bahwa *thresholding* mengubah citra keabuan menjadi citra biner bergantung nilai *threshold* (T), sehingga dapat diketahui daerah mana yang termasuk objek dan latar belakang. Jika nilai *pixel* lebih besar dari *threshold* diatur menjadi 1. Sebaliknya jika kurang dari *threshold* diatur menjadi 0.

2) Implementasi Ekstraksi Fitur Hu Moment Invariant

Implementasi Ekstraksi Fitur pada tahap ini yang dimana terjadinya proses perubahan data citra lalu dikonversi menjadi data numerik, kemudian menghasilkan 7 array nilai label mulai dari H1-H7, serta label target dimana target merupakan hasil dari citra bahwa itu termasuk citra penyakit malaria terinfeksi atau malaria tidak terinfeksi. Setelah melakukan proses konversi maka data numerik tersebut akan di simpan dalam file format .csv (*Comma Separated Values*) di *google drive*. Pada Tabel 3. menunjukkan hasil ekstraksi fitur hu moment dari citra penyakit malaria terinfeksi:

Tabel 3. Hasil Ekstraksi Hu Moment Citra Malaria Terinfeksi

id	H1	H2	H3	H4	H5	H6	H7	Actual
1	0.013966	4.76x10 ⁻⁶	1.82x10 ⁻⁸	1.82x10 ⁻⁸	9.56x10 ⁻¹⁶	1.82X1010	3.14X1015	1
2	0.013095	3.96x10 ⁻⁶	5.86x10 ⁻⁸	5.86x10 ⁻⁸	1.70x10 ⁻¹⁵	7.93X1011	-9.15X1016	1
3	0.019049	5.22 x10 ⁻⁵	7.75x10 ⁻⁸	2.31x10 ⁻⁷	-2.85x10 ⁻¹⁴	1.27X1009	-1.20X1014	1
4	0.016203	8.70 x10 ⁻⁷	4.47x10 ⁻⁷	2.93x10 ⁻⁷	-1.01x10 ⁻¹³	1.70X1010	-3.23X1014	1
5	0.012697	4.71 x10 ⁻⁶	4.17x10 ⁻⁸	3.80x10 ⁻⁷	4.74x10 ⁻¹⁴	7.35X1010	-6.59X1015	1
6	0.016578	1.76 x10 ⁻⁵	8.56x10 ⁻⁷	1.16x10 ⁻⁷	-3.35x10 ⁻¹⁴	-3.56X1010	1.50X1014	1
7	0.008795	2.05 x10 ⁻⁷	2.04x10 ⁻⁸	7.96x10 ⁻⁹	-8.97x10 ⁻¹⁷	-3.55X1012	-4.73X1017	1
8	0.004598	3.72 x10 ⁻⁷	4.56x10 ⁻⁹	5.96x10 ⁻⁹	-2.65x10 ⁻¹⁷	-3.62X1012	1.62X1017	1
9	0.026308	7.43 x10 ⁻⁶	1.23x10 ⁻⁶	8.85x10 ⁻⁷	-7.15x10 ⁻¹³	2.08X1009	5.88X1013	1
10	0.015288	3.44 x10 ⁻⁵	1.36x10 ⁻⁶	2.03x10 ⁻⁷	5.04x10 ⁻¹⁴	-2.80X1010	9.42X1014	1

Pada Tabel 4. Menunjukkan hasil ekstraksi fitur hu *moment* dari citra penyakit malaria tidak terinfeksi.

Tabel 4. Hasil EKstraksi Hu *Moment* Citra Malaria Tidak Terinfeksi

id	H1	H2	H3	H4	H5	H6	H7	Actual
1	0.035080	5.11E-05	4.25E-07	1.41E-07	-2.99E-14	1.78E-10	-1.72E-14	0
2	0.034956	2.70E-05	1.43E-07	2.58E-07	1.85E-14	1.34E-09	4.59E-14	0
3	0.038674	0.000162	3.01E-06	1.99E-06	1.65E-12	1.46E-08	4.58E-12	0
4	0.033657	5.08E-06	4.39E-07	2.38E-07	-5.27E-14	2.25E-10	5.63E-14	0
5	0.031494	9.96E-06	2.91E-07	2.54E-07	6.48E-14	-7.08E-11	2.40E-14	0
6	0.020235	3.03E-05	8.21E-07	4.06E-08	-2.63E-15	2.04E-10	-6.93E-15	0
7	0.016468	2.21E-06	5.68E-09	3.13E-08	-3.43E-16	-4.46E-11	-2.38E-16	0
8	0.015832	2.18E-05	1.03E-07	6.57E-09	-1.29E-16	2.00E-11	-1.12E-16	0
9	0.017107	1.19E-05	2.61E-08	4.06E-08	1.26E-15	9.57E-11	-3.93E-16	0
10	0.016639	8.07E-06	2.61E-09	8.37E-08	5.50E-17	8.02E-11	1.24E-15	0

3) Implementasi Metode K-NN

Implementasi metode K-NN pada tahap ini merupakan contoh perhitungan manual mulai dari penetapan data *training* dan data *testing*, implementasi metode KNN hingga perhitungan performa. Pada Tabel 5. dan Tabel 6. ditunjukkan 12 *sample* data yang akan diterapkan untuk perhitungan manual.

Tabel 5. *Sample* Data Training

id	H1	H2	H3	H4	H5	H6	H7	Actual
1	0.0233951	1.40E-05	4.96E-08	4.96E-08	1.29E-13	2.61E-09	1.71E-13	1
2	0.0139667	4.76E-06	1.82E-08	1.82E-08	9.56E-16	1.82E-10	3.14E-15	1
3	0.0130955	3.96E-06	5.86E-08	5.86E-08	1.70E-15	7.93E-11	-9.15E-16	1
4	0.0190493	5.22E-05	7.75E-08	2.31E-07	-2.85E-14	1.27E-09	-1.20E-14	1
5	0.0162035	8.70E-07	4.47E-07	2.93E-07	-1.01E-13	1.70E-10	-3.23E-14	1
6	0.0202357	3.03E-02	8.21E-07	4.06E-08	-2.63E-15	2.04E-10	-6.93E-15	0
7	0.0164668	2.21E-06	5.68E-09	3.13E-08	-3.43E-16	-4.46E-11	-2.38E-15	0
8	0.016639	8.07E-06	2.61E-09	8.37E-08	5.50E-17	8.02E-11	1.24E-15	0
9	0.017107	1.19E-05	2.61E-08	4.06E-08	1.26E-15	9.57E-11	-3.93E-16	0
10	0.015832	2.18E-05	1.03E-07	6.57E-09	-1.29E-16	2.00E-11	-1.12E-16	0

Tabel 6. *Sample* Data Testing

id	H1	H2	H3	H4	H5	H6	H7	Actual
1	0.0206061	2.05E-06	5.30E-08	3.83E-07	-4.29E-14	2.81E-10	3.36E-14	1
2	0.0417400	4.09E-06	7.10E-08	6.44E-08	4.12E-15	-8.85E-11	-1.42E-15	0

Pada Tabel 5. menunjukkan dataset yang digunakan dimana 10 data digunakan sebagai data *training*. Sedangkan Tabel 6, dimana 2 data sebagai data *testing*. Setelah ditentukan data *testing* tahap selanjutnya yaitu proses implementasi metode K-NN dengan menggunakan persamaan 6.

Pada Tabel 8. menunjukkan hasil perhitungan data *testing* yang pertama. Dimana setiap satu data *testing* akan dihitung ke semua data *training*. Pada perhitungan dibawah ini merupakan setiap atribut yang akan disandingkan pada *atribut* data *training* sesuai pada rumus *Euclidian* K-NN, kemudian di urutkan sesuai jumlah K yang telah ditentukan dimana hasil dari urutan tersebut dilihat *class* yang paling dominan, dari hasil inilah yang menjadi acuan pengukuran performa dengan mencocokkan TP, TN, FP dan FN pada *Confusion Matrix*.

Tabel 7. *Sample* Data Testing I

id	H1	H2	H3	H4	H5	H6	H7	Actual
1	0.0206061	2.05E-06	5.30E-08	3.83E-07	-4.29E-14	2.81E-10	3.36E-14	1

Tabel 8. Perhitungan Data Testing I

	Hasil Data Testing 1	K=9	Urutan Hasil Perhitungan Jarak Terdekat	Actual
1	9,4284E+15	4	6,92713E+15	0
2	1,02996E+16	5	7,56222E+15	0
3	2,14902E+16	8	8,09978E+15	0
4	2,17748E+16	10	9,4284E+15	1
5	1,0698E+16	6	1,02996E+16	1
6	8,09978E+15	3	1,0698E+16	1
7	2,13716E+16	7	2,13716E+16	0
8	6,92713E+15	1	2,14902E+16	1

9	7,56222E+15	2	2,16844E+16	0
10	2,16844E+16	9	2,17748E+16	1

Jadi, kesimpulan pada Tabel 8. data testing I, untuk nilai K=9 mempunyai hasil klasifikasi yang tidak tepat.

Tabel 9. *Sample Data Testing II*

id	H1	H2	H3	H4	H5	H6	H7	Actual
2	0.0417400	4.09E-06	7.10E-08	6.44E-08	4.12E-15	-8.85E-11	-1.42E-15	0

Tabel 10. Perhitungan Data Testing II

Id	Hasil Data Testing 2	K=9	Urutan Hasil Perhitungan Jarak Terdekat	Actual
1	2,67292E+15	3	1,71658E+14	0
2	3,54417E+15	4	8,06749E+14	0
3	1,47348E+16	7	2,67292E+15	1
4	1,50194E+16	10	3,54417E+15	1
5	3,94249E+15	5	3,94249E+15	1
6	1,48553E+16	8	1,46161E+16	0
7	1,46161E+16	6	1,47348E+16	1
8	1,71658E+14	1	1,48553E+16	0
9	8,06749E+14	2	1,49289E+16	0
10	1,49289E+16	9	1,50194E+16	1

Pada Tabel 10. menunjukkan hasil perhitungan data *testing* yang kedua. Dimana setiap satu data *testing* akan dihitung kesemua data training pada data perhitungan merupakan hasil perhitungan dimana setiap atribut akan disandingkan pada atribut data *training* sesuai pada rumus *Euclidian* K-NN, kemudian diurutkan sesuai jumlah K yang telah ditentukan dimana hasil dari urutan tersebut dilihat *class* yang paling dominan. Jadi, kesimpulan pada Tabel 10. dari data testing II tersebut, untuk nilai K=9 memiliki hasil klasifikasi tepat yang dimana dominannya jatuh nilai 0 (tidak terinfeksi).

B. Pembahasan

Keseluruhan teknik pada penelitian ini menggunakan bahasa pemrograman *python* dan *library scikit learn* yang ada didalamnya. Di mulai dari tahap *Load Dataset* Citra Penyakit Malaria, Deteksi Tepi *Canny*, Ekstraksi Fitur Hu Moment Invariant, *Split Data Testing and Training*, Implementasi metode K-NN, hingga mendapatkan nilai performa.

1) Pemahaman DataSet

Berdasarkan apa yang telah diutarakan bahwa pengumpulan dataset yang digunakan pada penelitian ini berasal dari studi pustaka dan *dataset* dari Kaggle *dataset library*. Kaggle merupakan sebuah situs penyedia jasa *repository dataset*, dimana dataset citra tersebut diolah dari U.S National *Library of Medicine*. Dataset dikumpulkan dan dikelola hingga tahun 2018. Kemudian dirilis/diperbaharui di Kaggle pada tanggal 05 Desember 2018. Pada Tabel 11. menunjukkan simulasi pembagian data yang akan digunakan pada penelitian ini.

Tabel 11. Simulasi Pembagian Dataset

Dataset	Jumlah	Rasio Dataset	Split Dataset
Dataset 1	1000	Terinfeksi 500	90% Training
		Tidak Terinfeksi 500	10% Testing
Dataset 2	1000	Terinfeksi 800	90% Training
		Tidak Terinfeksi 200	10% Testing
Dataset 3	1000	Terinfeksi 200	90% Training
		Tidak Terinfeksi 800	10% Testing

Pada proses *split* (membagi) dataset dibagi menjadi 90% data *training* dan 10% data *testing* yang memiliki nilai *random state = 2*. *Random state* digunakan agar dataset yang dibagi tidak berubah-ubah, sehingga diberikan nilai ketetapanannya.

C. Pengujian

1) Pengujian Performa

Pengujian yang dilakukan pada penelitian ini yaitu pengujian performa metode K-NN pada dataset citra penyakit malaria yang terdiri dari 3000 data dimana 90% data sebagai *training* dan 10% data sebagai *testing*, pengujian performa terdiri dari perhitungan nilai akurasi, presisi, *recall* dan *f-measure* seperti yang telah dijelaskan pada persamaan 7, 8, 9 dan 10. Sebelum perhitungan performa akurasi, presisi, *recall* dan *f-measure*. Hasil implementasi K-NN ditunjukkan dalam bentuk *confusion matrix*. Setiap pengujian nilai K memiliki *confusion matrix* masing-masing.

Pada Tabel 12. menunjukkan *confusion matrix* untuk nilai K = 10 pada simulasi jumlah dataset malaria terinfeksi dan tidak terinfeksi dengan rasio 50:50 dari 1000 dataset.

Tabel 12. Confusion Matrix Nilai k = 10 Pada Rasio Dataset 50:50

	Actual Terinfeksi	Actual Tidak Terinfeksi
Classified Terinfeksi	TP = 47	FP = 7
Classified Tidak Terinfeksi	FN = 9	TN = 37

Setelah diperoleh nilai TP, FP, FN dan TN maka nilai akurasi, presisi, *recall* dan *f-measure* dapat dihitung seperti yang di ilustrasikan dibawah ini:

$$\text{Akurasi} = \frac{(47+37)}{(47+37+9+7)} = 0,84 = 84\%$$

$$\text{Presisi} = \frac{47}{(47+7)} = 0,87 = 87\%$$

$$\text{Recall} = \frac{47}{(47+9)} = 0,84 = 84\%$$

$$F - \text{Measure} = 2 \times \frac{0,87 \times 0,84}{(0,87+0,84)} = 0,85 = 85\%$$

Pada Tabel 13. menunjukkan *confusion matrix* untuk nilai K = 9 pada simulasi jumlah dataset malaria terinfeksi dan tidak terinfeksi dengan rasio 80:20 dari 1000 dataset.

Tabel 13. Confusion Matrix Nilai k = 9 Pada Rasio Dataset 80:20

	Actual Terinfeksi	Actual Tidak Terinfeksi
Classified Terinfeksi	TP = 9	FP = 3
Classified Tidak Terinfeksi	FN = 13	TN = 75

Setelah diperoleh nilai TP, FP, FN dan TN maka nilai akurasi, presisi, *recall* dan *F-measure* dapat di hitung seperti yang di ilustrasikan berikut.

$$\text{Akurasi} = \frac{(9+75)}{(9+75+13+3)} = 0,84 = 84\%$$

$$\text{Presisi} = \frac{9}{(9+3)} = 0,75 = 75\%$$

$$\text{Recall} = \frac{9}{(9+13)} = 0,41 = 41\%$$

$$F - \text{Measure} = 2 \times \frac{0,75 \times 0,41}{(0,75+0,41)} = 0,53 = 53\%$$

Pada Tabel 14. menunjukkan *confusion matrix* untuk nilai K = 5 pada simulasi jumlah dataset malaria terinfeksi dan tidak terinfeksi dengan rasio 20:80 dari 1000 dataset.

Tabel 14. Confusion Matrix Nilai k = 5 Pada Rasio Dataset 20:80

	Actual Terinfeksi	Actual Tidak Terinfeksi
Classified Terinfeksi	TP = 80	FP = 11
Classified Tidak Terinfeksi	FN = 0	TN = 9

Setelah diperoleh nilai TP, FP, FN dan TN maka nilai akurasi, presisi, *recall* dan *F-measure* dapat di hitung seperti yang di ilustrasikan berikut.

$$\text{Akurasi} = \frac{(80+9)}{(80+9+0+11)} = 0,89 = 89\%$$

$$\text{Presisi} = \frac{80}{(80+11)} = 0,88 = 88\%$$

$$\text{Recall} = \frac{80}{(80+0)} = 1,0 = 100\%$$

$$F - \text{Measure} = 2 \times \frac{0,88 \times 1,0}{(0,88 + 1,0)} = 0,94 = 94\%$$

2) Hasil Pengujian

Kesimpulan pengujian dilakukan setelah proses metode K-NN dilakukan kesimpulan yang dihasilkan berupa nilai performa yang terdiri dari akurasi, presisi, *recall* dan *f-measure*.

3) Hasil Pengujian Performa

Pengujian performa pada penelitian ini juga dilakukan pada nilai K=2 hingga nilai K=10 untuk melihat nilai performa dengan rasio data 50:50 untuk malaria terinfeksi dan tidak terinfeksi. Pada [Tabel 15](#). menunjukkan hasil performa dengan data:

Tabel 15. Hasil Uji Performa Metode K-NN Dataset Rasio 50:50

Nilai K	Akurasi	Presisi	Recall	F-Measure
2	79	76	91	83
3	79	84	77	80
4	82	82	88	84
5	84	95	75	84
6	81	84	82	83
7	83	91	77	84
8	82	85	82	84
9	81	91	73	81
10	84	87	84	85

Berdasarkan [Tabel 15](#). maka dapat disimpulkan bahwa dengan menguji nilai K=2 hingga K=10 maka diperoleh performa terbaik pada nilai K=10, dimana nilai K=10 mempunyai akurasi 84%, presisi 87%, *recall* 84% dan *f-measure* 85%. Pada [Tabel 16](#). pengujian performa pada penelitian dilakukan pada nilai K=2 hingga nilai K=10 untuk melihat nilai performa dengan rasio data 80:20 untuk malaria terinfeksi dan tidak terinfeksi. Hasil uji performa metode K-NN ditunjukkan pada [Tabel 16](#).

Tabel 16. Hasil Uji Performa Metode K-NN Dataset Rasio 80:20

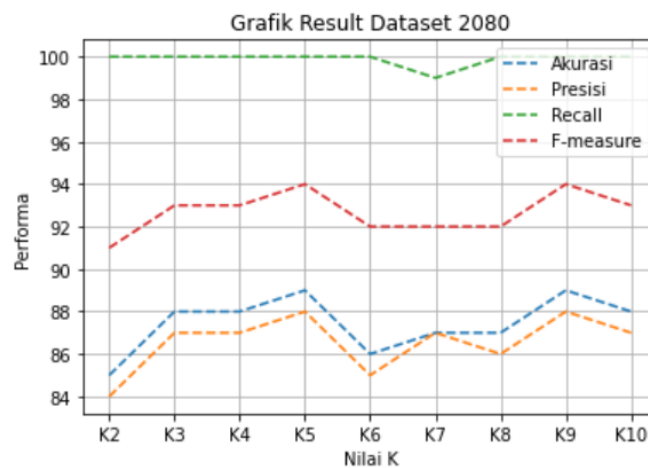
Nilai K	Akurasi	Presisi	Recall	F-Measure
2	75	45	59	51
3	79	54	32	40
4	83	65	50	56
5	82	64	41	50
6	84	67	55	60
7	82	63	45	53
8	82	63	45	53
9	84	75	41	53
10	84	69	50	58

Berdasarkan [Tabel 16](#). maka dapat disimpulkan bahwa dengan menguji nilai K=2 hingga K=10 maka diperoleh performa terbaik pada nilai K=6, dimana nilai K=6 mempunyai akurasi 84%, presisi 67%, *recall* 55% dan *f-measure* 60%. Pada [Tabel 17](#). pengujian performa pada nilai K=2 hingga nilai K=10 untuk melihat nilai performa dengan rasio data 20:80 untuk malaria terinfeksi dan tidak terinfeksi. Hasil uji performa metode K-NN ditunjukkan pada [Tabel 17](#).

Tabel 17. Hasil Uji Performa Metode K-NN Dataset Rasio 20:80

Nilai K	Akurasi	Presisi	Recall	F-Measure
2	85	84	100	91
3	88	87	100	93
4	88	87	100	93
5	89	88	100	94
6	86	85	100	92
7	87	87	99	92
8	87	86	100	92
9	89	88	100	94
10	88	87	100	93

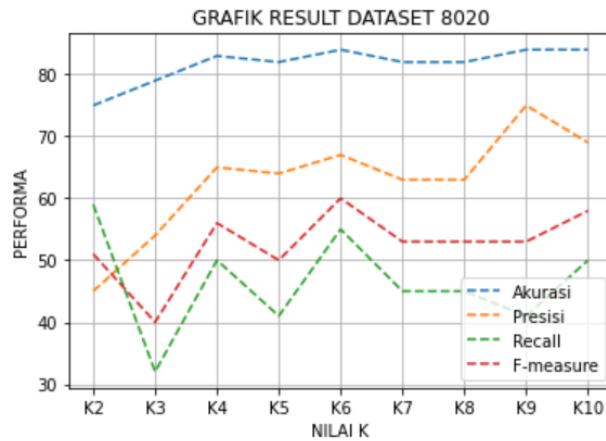
Berdasarkan Tabel 17. Maka dapat disimpulkan bahwa dengan menguji nilai K=2 hingga K=10 maka diperoleh performa terbaik pada nilai K=5, dimana nilai K=5 mempunyai akurasi 89%, presisi 88%, recall 100% dan *f-measure* 94%. Gambar 4. menunjukkan grafik perbandingan nilai K2 sampai K10 pada rasio dataset 20:80.



Gambar 4. Grafik Rasio Dataset Citra Penyakit Malaria 20:80

Kesimpulan dari Gambar 4. menunjukkan grafik rasio dataset citra 20:80 bahwa tanda garis berwarna biru diberikan keterangan untuk nilai akurasi, dimana nilai akurasi yang tertinggi dimiliki oleh nilai K5 dan K9 sebesar 89%. Sedangkan tanda garis berwarna orange diberikan keterangan untuk nilai presisi dapat dilihat bahwa presisi tertinggi dari nilai K3, K4, K7 dan K10 sebesar 87%, lalu K2 memiliki presisi terendah sebesar 84%.

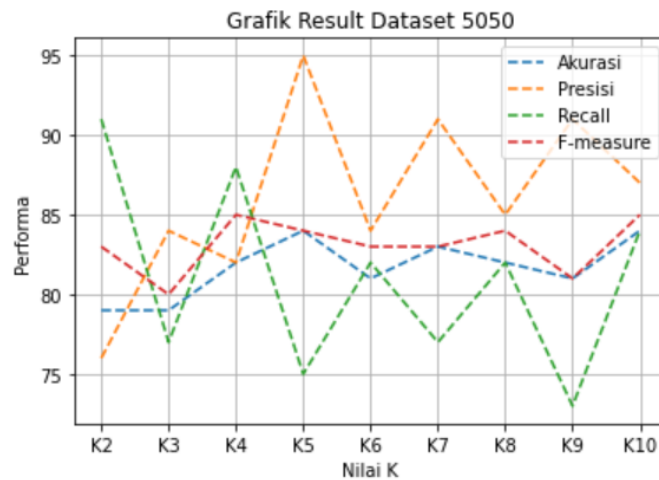
Tanda garis berwarna hijau diberikan keterangan untuk nilai recall dapat dilihat bahwa recall tertinggi dari nilai K2, K3, K4, K5, K6, K8 dan K9 sebesar 100%, lalu K7 memiliki recall terendah sebesar 99%. Kemudian pada tanda garis berwarna merah diberikan keterangan untuk nilai *f-measure* dapat dilihat bahwa *f-measure* tertinggi dari nilai K5 dan K9 sebesar 94%, lalu K2 memiliki *f-measure* terendah sebesar 91%. Gambar 5. menunjukkan grafik perbandingan nilai K2 sampai K10 pada rasio dataset 80:20.



Gambar 5. Grafik Rasio Dataset Citra Penyakit Malaria 80:20

Kesimpulan dari Gambar 5. yang menunjukkan grafik rasio dataset citra 80:20 bahwa tanda garis berwarna biru diberikan keterangan untuk nilai akurasi, dimana nilai akurasi yang tertinggi dimiliki oleh nilai K6, K9 dan K10 sebesar 84%. Sedangkan tanda garis berwarna orange diberikan keterangan untuk nilai presisi dapat dilihat bahwa presisi tertinggi dari nilai K9 sebesar 75%, lalu K2 memiliki presisi terendah sebesar 45%.

Tanda garis berwarna hijau diberikan keterangan untuk nilai recall dapat dilihat bahwa recall tertinggi dari nilai K2 sebesar 59%, lalu K3 memiliki recall terendah sebesar 32%. Kemudian pada tanda garis berwarna merah diberikan keterangan untuk nilai f-measure dapat dilihat bahwa f-measure tertinggi dari nilai K10 sebesar 58%, lalu K3 memiliki f-measure terendah sebesar 40%. Gambar 6. menunjukkan grafik perbandingan nilai K2 sampai K10 pada rasio dataset 50:50



Gambar 6. Grafik Rasio Dataset Citra Penyakit Malaria 50:50

Kesimpulan dari Gambar 6. yang menunjukkan grafik rasio dataset citra 50:50 bahwa tanda garis berwarna biru diberikan keterangan untuk nilai akurasi, dimana nilai akurasi yang tertinggi dimiliki oleh nilai K5 dan K10 sebesar 84%. Sedangkan tanda garis berwarna orange diberikan keterangan untuk nilai presisi dapat dilihat bahwa presisi tertinggi dari nilai K5 sebesar 95%, lalu K2 memiliki presisi terendah sebesar 76%.

Tanda garis berwarna hijau diberikan keterangan untuk nilai recall dapat dilihat bahwa recall tertinggi dari nilai K4 sebesar 88%, lalu K9 memiliki recall terendah sebesar 73%. Kemudian pada tanda garis berwarna merah diberikan keterangan untuk nilai f-measure dapat dilihat bahwa f-measure tertinggi dari nilai K10 sebesar 85%, lalu K3 memiliki f-measure terendah sebesar 80%.

Tabel 18. Perbandingan Hasil Penelitian

No	Peneliti	Metode	Rasio Dataset	Jumlah Data Training	Jumlah Data Testing	Akurasi	Presisi	Recall	F-Measure
1	Kartika	KNN	120	80	40	95%	-	-	-

(2013)		K = 5 dan 7								
2	Aisyah (2020)	KNN		50:50	900	100	84%	95%	75%	84%
		K = 5								
		K = 10		80:20	900	100	84%	75%	41%	53%
		KNN								
		K = 6								
		K = 9		20:80	900	100	89%	88%	100%	94%
		KNN								
		K = 5								
		K = 9								

Tabel 18. menunjukkan hasil dari penelitian ini serta perbandingan dengan penelitian sebelumnya, kesimpulan yang dapat ditarik yaitu:

- Penelitian sebelumnya memiliki nilai akurasi yang lebih baik. Namun, data yang digunakan belum cukup banyak serta tidak menunjukkan keseimbangan antara klasifikasi yang dilakukan condong yang tepat mengklasifikasikan pada class tertentu.
- Dari berbagai kombinasi rasio dataset serta percobaan berbagai nilai K, diperoleh nilai akurasi tertinggi pada rasio 80:20 yaitu sebesar 84% namun nilai *f-measure* terlihat rendah karena ketepatan klasifikasi lebih condong ke presisi dibanding recall artinya klasifikasi kurang tepat dalam menduga penderita positif malaria.

Nilai akurasi tertinggi pada penelitian ini terdapat pada kombinasi dataset 20:80 serta percobaan berbagai nilai K yang mana memperoleh akurasi sebesar 89%, dimana nilai *f-measure* menunjukkan nilai yg tinggi artinya keseimbangan proses klasifikasi dalam menduga positif malaria yang cukup seimbang.

IV. Kesimpulan

Berdasarkan hasil penelitian ini maka penulis dapat menarik kesimpulan dengan menguji nilai K=2 hingga K=10. Pada dataset 1 (dataset 50:50) diperoleh nilai performa terbaik pada K = 5 dan K=10, dimana performa terbaik terdapat pada nilai K=10 yang mempunyai akurasi 84%, presisi 87% yang artinya 87% di klasifikasikan sebagai positif (aktualnya bernilai negatif), recall 84% yang artinya 84% di klasifikasikan sebagai negatif (aktualnya bernilai positif) dan *f-measure* 85% yang artinya nilai dari perbandingan rata-rata presisi dan recall yang dibobotkan. Pada dataset 2 (dataset 80:20) diperoleh nilai performa terbaik pada nilai K=6, dimana nilai K=6 mempunyai akurasi 84%, presisi 67% artinya 67% di klasifikasikan sebagai positif (aktualnya bernilai negatif), recall 55% yang artinya 55% di klasifikasikan sebagai negatif (aktualnya bernilai positif) dan *f-measure* 60% yang artinya nilai dari perbandingan rata-rata presisi dan recall yang dibobotkan. Pada dataset 3 (dataset 20:80) diperoleh nilai performa terbaik pada nilai K=5 dan K=9, dimana mempunyai akurasi 89%, presisi 88% artinya 88% diklasifikasikan sebagai positif (aktualnya bernilai negatif), recall 100% yang artinya 100% di klasifikasikan sebagai negatif (aktualnya bernilai positif) dan *f-measure* 94% yang artinya nilai dari perbandingan rata-rata presisi dan recall yang dibobotkan. Penelitian sebelumnya memiliki nilai akurasi yang lebih baik. Namun, data yang digunakan belum cukup banyak serta tidak menunjukkan keseimbangan antara klasifikasi yang dilakukan condong yang tepat mengklasifikasikan pada class tertentu. Dari berbagai kombinasi rasio dataset serta percobaan berbagai nilai K, diperoleh nilai akurasi tertinggi pada rasio 80:20 yaitu sebesar 84% namun nilai *f-measure* terlihat rendah karena ketepatan klasifikasi lebih condong ke presisi dibanding recall artinya klasifikasi kurang tepat dalam menduga penderita positif malaria. Nilai akurasi tertinggi pada penelitian ini terdapat pada kombinasi dataset 20:80 serta percobaan berbagai nilai K yang mana memperoleh akurasi sebesar 89%, dimana nilai *f-measure* menunjukkan nilai yg tinggi artinya keseimbangan proses klasifikasi dalam menduga positif malaria yang cukup seimbang.

Daftar Pustaka

- D. Katiandagho and A. Donsu, "Analisis Faktor Risiko Kejadian Malaria pada Ibu Hamil di Puskesmas Manganitu Kabupaten Kepulauan Sangehe," *J. Hig.*, vol. 4, no. 2, pp. 109–120, 2018.
- S. Rostianingsih, A. W. Kitu, and I. Gunawan, "Perancangan Dan Pembuatan Sistem Informasi Geografis Untuk Penyebaran Penyakit Malaria," *Semin. Nas. Inform.*, vol. 1, no. 4, pp. 220–228, 2015.
- Kemenkes RI, "Malaria." 2013.
- F. T. Admojo and Ahsanawati, "Klasifikasi Aroma Alkohol Menggunakan Metode KNN," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 34–38, 2020.
- A. A. D. Halim and S. Anraeni, "Analisis Klasifikasi Dataset Citra Penyakit Pneumonia menggunakan Metode K-Nearest Neighbor (KNN)," *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 01–12, 2021, doi: 10.33096/ijodas.v2i1.23.
- I. P. Putri, "Analisis Performa Metode K- Nearest Neighbor (KNN) dan Crossvalidation pada Data

- Penyakit Cardiovascular,” *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 21–28, 2021, doi: 10.33096/ijodas.v2i1.25.
- [7] H. Azis, “Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor,” *Indones. J. Data Sci.*, pp. 1–4, 2018, doi: <https://doi.org/10.33096/ijodas.v1i1.3>.
- [8] N. K. I. SP and T. Sutojo, “Klasifikasi Plasmodium vivax dari Digitalisasi Mikroskopis Sediaan Darah Tebal Menggunakan Kombinasi Ekstraksi Ciri Statistik Orde Kedua dan Metode K-Nearest Neighbor (K-NN) Classifier,” *J. Univ. Dian Nuswantoro Semarang*, no. c, pp. 1–5, 2015.
- [9] S. Sahar, “Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Classifier Pada Dataset Penyakit Jantung,” *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.
- [10] A. Prasetya Wibawa, W. Lestar, A. Bella Putra Utama, I. Tri Saputra, and Z. Nabila Izdihar, “Multilayer Perceptron untuk Prediksi Sessions pada Sebuah Website Journal Elektronik,” *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 57–67, 2020, doi: 10.33096/ijodas.v1i3.15.
- [11] M. D. U. Ulhaq and Irawati, “Implementasi Metode Visekriterijumsko Kompromisno Rangiranje (VIKOR) Pada Seleksi Program Keluarga Harapan Komponen Pendidikan Berbasis Web,” *Indones. J. Data Sci.*, vol. 2, no. 1, pp. 38–49, 2021, doi: 10.33096/ijodas.v2i1.30.
- [12] M. M. Baharuddin, T. Hasanuddin, and H. Azis, “Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca,” *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.
- [13] Rosmasari *et al.*, “Usability Study of Student Academic Portal from a User’s Perspective,” *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, pp. 108–113, 2018, doi: 10.1109/EIConCIT.2018.8878618.
- [14] F. Tangguh and Y. Islami, “Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin,” *Indones. J. Data Sci.*, vol. 3, no. 1, pp. 1–8, 2022.
- [15] N. Litha and T. Hasanuddin, “Analisis Performa Metode Moving Average Model untuk Prediksi Jumlah Penderita Covid-19,” *Indones. J. Data Sci.*, vol. 1, no. 3, pp. 87–95, 2020.
- [16] H. Azis, F. T. Admojo, and E. Susanti, “Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah,” *Techno.Com*, vol. 19, no. 3, 2020.
- [17] L. Saiman and R. Satra, “Analisis performa metode Support Vector Machine untuk klasifikasi dataset aroma tahu berformalin,” *Indones. J. Data Sci.*, vol. 2, no. 2, pp. 50–61, 2021.
- [18] Hasran, “Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor,” *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.
- [19] A. Fitria and H. Azis, “Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier,” *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.
- [20] R. Bayu and Muslim, “Penerapan Metode Market Basket Analysis pada Minimarket Toko Baru,” *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–5, 2020.
- [21] A. Maulida, “Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes,” *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, 2020.