e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v17i1.5779

Penerapan Algoritma *K-Nearest Neighbors* Untuk Klasifikasi Fragmen Metagenom Berdasarkan Ekstraksi Fitur *K-Mers*

Ryan Ananda Nolly [1)], Amanda Fitria [2)] Kana Saputra S [3)]
Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Medan
Jl. Williem Iskandar Pasar V, Kota Medan, Sumatera Utara 20221
E-Mail : ryan_nolly@mhs.unimed.ac.id [1)]; amandafitria0127@gmail.com [2)]; kanasaputras@unimed.ac.id [3)]

ABSTRAK

Penelitian di bidang metagenomika menjadi salah satu bidang kajian bioinformatika yang terus berkembang. Metagenom merupakan sebuah teknik yang bertujuan untuk mengumpulkan gen-gen yang diambil secara langsung dari lingkungan dan mengenalisis informasi genetika di dalamnya. Data yang diambil langsung dari lingkungan memungkinkan fragmen yang dihasilkan mengandung berbagai mikroorganisme, sehingga akan berakibat pada terjadinya kesalahan perakitan terhadap fragmen metagenom. Proses binning (pengelompokan) dapat dilakukan dengan dua pendekatan, yaitu pendekatan homologi dan pendekatan komposisi. Pendekatan secara komposisi tidak perlu membandingkan dan menyimpulkan setiap hasil pencarian pada setiap level taksonomi sehingga waktu yang diperlukan untuk pengelompokan lebih cepat dibandingkan dengan pendekatan secara homologi. Pada proses binning (pengelompokan) dengan pendekatan komposisi, teknik yang dilakukan adalah dengan *supervised learning*. Tujuan dari penelitian ini adalah untuk mengklasifikasi fragmen metagenom menggunakan algoritma KNN dan *K-Mers* sebagai ekstraksi fitur. Selain itu, untuk menghitung tingkat akurasi klasifikasi fragmen metagenom menggunakan *confusion matrix*. Metode *K-Mers* yang digunakan sebagai ekstraksi fitur bertujuan untuk mempartisi data dan membentuk satu atau lebih kelompok yang memiliki kesamaan, sehingga perhitungan untuk mencari tingkat akurasi menjadi lebih mudah didapatkan. Berdasarkan hasil pengujian yang dilakukan menunjukkan bahwa semakin rendah nilai K yang digunakan pada KNN maka semakin tinggi akurasi yang diperoleh. Pada pengujian ini diperoleh perhitungan akurasi sebesar 94,37% dimana nilai K untuk KNN adalah 3 dan nilai K untuk *K-Mers* adalah 3. Hasil klasifikasi fragmen metagenom menggunakan algoritma KNN berdasarkan ekstraksi fitur K-Mers dapat dilakukan dengan baik.

Kata Kunci – klasifikasi, k-nearest neighbor, metagenom, ekstraksi fitur, k-mers.

1. PENDAHULUAN

Saat ini penelitian di bidang metagenomika menjadi salah satu bidang kajian bioinformatika yang terus berkembang. Metagenom merupakan suatu teknik yang secara khusus ditujukan untuk mengumpulkan gen-gen secara langsung dari suatu lingkungan, diikuti dengan menganalisis informasi genetika yang terkandung di dalamnya (Guritno, Haryanto, Kustiyo, & Hermadi, 2018). Tahapan utama dalam metagenom diawali dengan *DNA Sequencing* terhadap sampel metagenom. Data yang diambil langsung dari lingkungan memungkinkan fragmen yang dihasilkan mengandung berbagai mikroorganisme, sehingga akan berakibat pada terjadinya kesalahan perakitan terhadap fragmen metagenome (Pekuwali, Kusuma, & Buono, 2020).

Proses *Genom Sequencing* dari metagenom mengakibatkan percampuran antar organisme. Hal tersebut mengakibatkan kesulitan pada proses perakitan DNA. Sehingga, dibutuhkan proses pemilahan atau pengelompokan yang disebut dengan proses *binning*. Proses *binning* dapat dilakukan dengan dua pendekatan, yaitu pendekatan homologi dan pendekatan komposisi (Utami, Kusuma, & Buono, 2017). Pendekatan homologi, dilakukan dengan pencarian penjajaran sekuens dan membandingkan fragmen metagenom dengan basis data sekuens yang berasal dari *National Centre for Biotechnology Information* (NCBI). Tidak seperti pendekatan secara homologi, pendekatan secara komposisi tidak perlu membandingkan dan menyimpulkan setiap hasil pencarian pada tiap level taksonomi sehingga waktu yang diperlukan untuk pengelompokan lebih cepat dibandingkan dengan pendekatan secara homologi (Surianti, 2020).

Pada proses *binning* dengan pendekatan komposisi, teknik yang dilakukan adalah dengan *supervised learning* (Choiriyati, Arkeman, & Kusuma, 2020). Klasifikasi termasuk bagian dari prediksi, klasifikasi juga menggolongkan suatu data ke dalam kelompok data dimana kelompok data tersebut kelasnya telah terdefinisi. Penggunaan *K-Nearest Neighbor* (KNN) dalam hal pengklasifikasian data diterapkan pada data berjumlah besar serta memiliki banyak noise sehingga metode ini cukup mudah untuk diimplementasikan (Mutmainnah, Setiawan, & Dewi, 2019). Tahapan dalam supervised learning salah satunya ekstraksi fitur. Sehingga, penelitian ini menggunakan metode ektraksi fitur *K-Mers*. Besarnya jumlah parameter *n* pada metode ekstraksi fitur *K-Mers* akan mengakibatkan dimensi fitur yang tinggi.

Penelitian sebelumnya untuk mengklasifikasi metagenom pernah dilakukan menggunakan metode *Naïve Bayes Classifier* dengan hasil akurasi yang diperoleh dengan menggunakan fragmen pendek (400 bp) sebesar 49,34% untuk ekstraksi ciri 3-mer, dan 53,95% untuk ekstrasi ciri 4-mer. Untuk fragmen panjang (10 kbp), akurasi mengalami peningkatan sebesar 82,23% untuk ekstraksi ciri 3-mer dan 85,89% untuk esktraksi ciri 4-mer (Utami et al., 2017). Selain itu, algoritma KNN pernah diimplementasikan untuk

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v17i1.5779

analisis sentimen pada review objek wisata dunia fantasi dengan hasil akurasi yang terbesar dengan nilai k=7 dengan akurasi 77,01% (Sari, 2020).

Penelitian ini akan menerapkan algoritma KNN untuk mengklasifikasi fragmen metagenom dan *K-Mers* sebagai ekstraksi fitur. Selain itu, penelitian ini juga akan menghitung tingkat akurasi hasil klasifikasi fragmen metagenom berdasarkan *Confusion Matrix*.

2. TINJAUAN PUSAKA

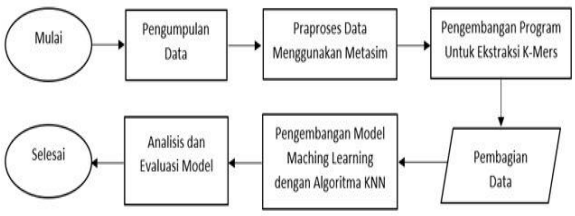
A. K-Nearest Neigbours (KNN)

Algoritma KNN merupakan algoritma klasifikasi *non-parametric* konvensional yang menghasilkan kinerja yang baik. Selain itu, algoritma ini juga mudah dipahami dan diimplementasikan (Prasetio, Rismayadi, & Anshori, 2018). Algoritma KNN merupakan algoritma *supervised learning*. *Supervised learning* bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang ada sebelumnya dengan data yang baru. Sedangkan pada *unsupervised learning*, data belum memiliki pola apapun, dan tujuan *unsupervised learning* untuk menemukan pola dalam sebuah data (Liantoni, 2015). Prinsip kerja KNN adalah mencari jarak terdekat antara data yang akan dievaluasi dengan k-tetangga (*neighbor*) terdekatnya dalam data pelatihan (Syaljumairi, Defit, Sumijan, & Elda, 2021). Tujuan dari algoritma KNN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan *training samples*, yang mana hasil dari sampel uji baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN.

B. Ekstraksi Fitur *K-Mers*

Ekstraksi fitur *K-Mers* dihasilkan dengan melakukan kombinasi dari 4 basa nukleotida, yaitu *Adenin* (A), *Cytosine* (C), *Guanine* (G) dan *Thymine* (T). Jumlah kombinasi fitur yang terbentuk adalah 4^k dengan nilai $k \geq 1$ (Utami et al., 2017).

3. METODE PENELITIAN



Gambar 1. Tahapan Penelitian

A. Pengumpulan Data

Data diperoleh dari situs NCBI yang dapat diakses melalui http://www.ncbi.nlm.nih.gov. Data tersebut meliputi 9 Spesies yang termasuk ke dalam 3 Genus, yaitu *Agrobacterium*, *Bacillus*, dan *Staphylococcus*. Untuk detailnya dapat dilihat pada Tabel 1.

Tabel 1. Dataset Spesies untuk Setiap Genus

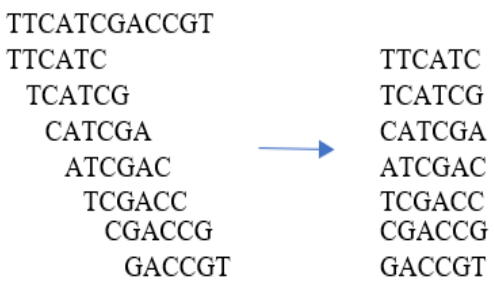
Spesies	Genus
Agrobacterium radiobacter K84 chromosome 2 *Agrobacterium tumefaciens str. C58 chromosome circular* *Agrobacterium vitis S4 chromosome 1*	*Agrobacterium*
Bacillus anthracis str. Ames Ancestor *Bacillus cereus 03BB102* *Bacillus pseudofarmus OF4 chromosome*	*Bacillus*
Staphylococcus aureus subsp. Aureus JH1 *Staphylococcus epidermidis ATCC 12228* *Staphylococcus haemolyticus JCSC1435*	*Staphylococcus*

B. Praproses Data

Data yang sudah dikumpulkan tersebut, selanjutnya disimulasikan menggunakan Aplikasi MetaSim sehingga diperoleh sekuens DNA dalam format FASTA (*.fna). Hasil simulasi tersebut menghasilkan 90000 fragmen data latih dengan komposisi 3000 fragmen *Agrobacterium*, 3000 fragmen *Bacillus*, dan 3000 fragmen *Staphylococcus*. Data uji yang digunakan adalah 4500 fragmen yang terdiri dari 1500 fragmen *Agrobacterium*, 1500 fragmen *Bacillus*, dan 1500 fragmen *Staphylococcus*. Setiap fragmen mempunyai panjang fragmen yang tetap, yaitu 500 bp (*basepair*).

C. Pengembangan Program Untuk Ekstraksi Fitur *K-Mers*

Ekstraksi fitur *K-Mers* dikembangkan menggunakan bahasa pemrograman Python untuk mengekstraksi fitur dari setiap sub string. Adapun algoritma dari ekstraksi tersebut adalah seperti berikut:



Gambar 2. Ilustrasi Proses dari Ekstraksi Fitur

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v17i1.5779

Sebagai contoh untuk fragmen seperti berikut:

Fragmen 1: AAAATCGACCCTTTTGAAAAG
Fragmen 2: GAGAATCGACCCTTTTGGAAT

Gambaran hasil ekstraksi fitur menggunakan *K-Mers* dengan nilai K = 3; maka banyak kombinasi fitur = 64 dapat dilihat pada Tabel 2.

Tabel 2. Gambaran Hasil Ekstraksi Fitur Menggunakan *K-Mers*

Fitur Fragmen	AAA	AAT	AAG	…	CCC	Genus
F1	4	1	1	…	1	*Agrobacterium*
F2	0	2	0		0	*Agrobacterium*
..	…	…	…	…	…	…
F90000	…	…	…	…	…	*Staphylococcus*

D. Pembagian Data

Dataset merupakan sebuah data yang diambil dari sumber data yang merepresentasikan data tabel dan relasinya dimana strukturnya mirip dengan data pada *database* (Syarifuddin, Misdram, Widodo, Informatika, & Pasuruan, 2020). Sebanyak 90.000 fragmen dari 3 Genus yang telah di ekstraksi akan dibagi menjadi 2 bagian yaitu data latih sebesar 75% dan data uji sebesar 25% (Ridok & Latifah, 2015).

E. Pengembangan Model Machine Learning dengan Algoritma Klasifikasi KNN

KNN merupakan algoritma klasifikasi yang menghitung jarak antar titik terdekat dengan titik yang mau diuji dan mencari kelas mana yang lebih banyak dekat. Adapun metode untuk menghitung jaraknya menggunakan *Minkowski* (Nishom, 2019).

$$d(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$$

Keterangan:
d = jarak antara x dan y
$x = data\ pusat\ klaster$
$y = data\ pada\ atribut$
$i = setiap\ data$
$n = jumlah\ data$
$x_i = data\ pada\ pusat\ klaster\ ke - i$
$y_i = data\ pada\ setiap\ data\ ke - i$
$p = power$

Model KNN akan dikembangkan menggunakan bahasa pemrograman *Python* dengan module management data seperti *Pandas*, *Numpy*, dan modul *Machine Learning scikit-learn*.

F. Analisis dan Evaluasi

Analisis dan evaluasi dilakukan untuk menghitung akurasi dari setiap perubahan nilai K pada KNN. Perhitungan akurasi menggunakan *confusion matrix* (Manik, Saputra S, & Br Ginting, 2020).

Tabel 3. *Confusion Matrix*

		Detected	
		Positive	***Negative***
Actual	***Positive***	*A: True Positive*	*C: False Negative*
	Negative	*B: False Positive*	*D: True Negative*

Semakin tinggi hasil perhitungan akurasi yang dihasilkan, maka modelnya akan semakin baik. Rumus untuk menghitung akurasi adalah sebagai berikut.

$$Akurasi = \frac{A + D}{A + B + C + D}$$

4. HASIL DAN PEMBAHASAN

Data yang digunakan pada penelitian ini meliputi 9 Spesies yang termasuk ke dalam 3 Genus, yaitu *Agrobacterium*, *Bacillus*, dan *Staphylococcus*. Sampel data yang diperoleh dari NCBI dengan format FASTA dapat dilihat pada Gambar 3.

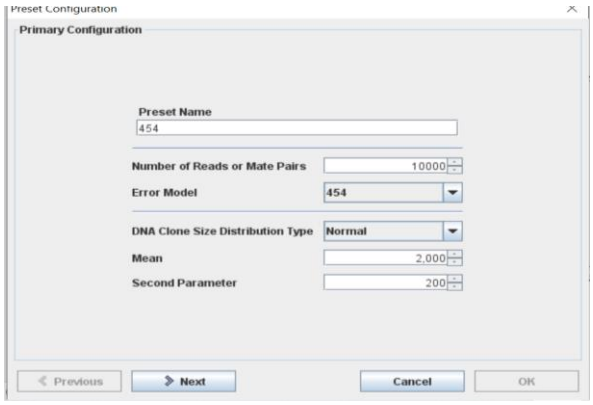

Agrobacterium radiobacter K84 chromosome 2 - Notepad
File Edit Format View Help
>NC_011983.1 Agrobacterium radiobacter K84 chromosome 2, complete sequence
CCTAGTTGCCCCGAATCATAGAATCGGAGTACTTCTTATCTACGGAAAATGCCGGTAGTACCGACTTTCT
GCGTAAATAAAGAAACCAGTCACAAAAACCCAACCATAAAGGCGCTTAGTAACGTGGCCAAAAATGGGAG
AAAACCATGACCCGACATCTGAGCAGCCTCGCCTTCATGCAGACCTTTTCTGCGAAGCTGGATACCGCAT
TGGGCAATCTCAGCCTAGCCCAGTTTCCCCCGCATGCCCGGCGCACCATGCGCAAGTTCACCTCCCCCGA
GGTTGCCGCGCTGCTTAACGTCACCGAAGCCTATATTAGGCAGGTTTCCCTGAAGGAGCAGGGGCCGACG
CCGGAAGTCGCCAATGGACGGCGTCTCTATACGCTCGAGCAGATCCTCGAGTTGCGTATGCACCTGGCGC
AGAACGGCCGCAAGAAGTGGATGAACCCGCGCAGGGTAGGCGAGGAGCAATGCCAGATCCTGGCAATCAC
CAATTTCAAGGGCGGCTCCAGCAAGACCTCCACCACGATCCATCTCGGTCATTATCTGGCCTTGAGAGGA
TACCGCGTGCTGGCCGTTGACCTCGACCCCCAGGCATCGCTGACCAGCCTGCATGGTAGTCTGCCGGATT
TCGATTATCGTGACGGCGATACGCTTTATTCCGCCATCCGCTTCGAGGATCCGATCCCGACCGAGCAGGT
CATTCATCAGACCCACATCGCCGGCTTCGATGTCATCTGCGCCGGCCTGGAATTGACGGAGTTCGAAACC
GCCGTTGCGCTTGAAATGCGCAAGTCTGGCGGCACGAGCTTCCTCTTGCGCGTATCGCAGGCACTGGAGC
AGGTGGCTGATCGCTACGACATCATCCTGATGGATTCGGCCCCGTCGCTGAACTTCCTCACCCTGTCATC
CCTGACCGCGGCGACAGGCGTCCTGATACCGGTTCCGGCACATATGCTCGATGTGGATTCCACAGCGAAG
TTCCTGGAACTCGCCGGATCCTACATGCAGATCCTCGAACAGGTCGGCACGTCGGCGCAATGGGATTTCG
CCAAGTTCCTGATCACCAAGTTCGAACCGAACGACCATCCCCAGGCCAATATGACGGCGCTGATGCGGCA
GGTTTTCGGCGAAGACCTGTTGCTTAACTCCGTGATCAAATCGACCGCGGTTGCCGATGCCTTGACCTGG
AAACAGAGCCTTTACGAGGTGCAGCGGTCGAGATTTTCCGCCCCGAAAACCTATGACAGGGCGATTGAGT
CGATCAATGCGGCAAATGCCGAAATCGAAGAGCTTCTTTCGAGAGCATGGGGACGCGAATGAGCAGGAAA
GATTCCAAAGGCATGTTTGCCAATGTCCTGATCGGTCTCGGCCAGGACACTCCATCCGTACCCGCCGCAA
CGACATCGCCGTCGCCGCATCTTCTGAAGGTCGCCGCCGGCGTTCGGCAGTTGCAGGAACGTGGCGAGCT
TGCCGACAAGCTGCTGCAGGGTGCCGATCACATCATCCAGCTCGACGTGGACGCGGTCCTTCCGTCTTCC
ATTCGGGATCGATTTGACGGAGCTTATGACGCCGACAGGCTTCAGGATCTCGTGCAGTCCATGCGCGAAC
ACGGGCAAACGACGCCTGGTCTCGTTCGGCCCATGCCGGGCGAGGGCAGCCGGTTTCAGATCGTCTTCGG

Gambar 3. Sampel Data *Agrobacterium Radiobacter K84 Chromosome 2*

Sebelum dilakukan ekstraksi fitur, data akan di praproses terlebih dahulu dengan membagikan keseluruhan dari *string* menjadi beberapa *substring* yang akan diekstraksi secara individual. Data spesies yang diperoleh dari NCBI selanjutnya disimulasikan menggunakan MetaSim dengan output berupa fragmen. MetaSim mensimulasikan kinerja *sequencer* untuk menghasilkan fragmen dari spesies yang direpresentasikan sebagai *string* dengan format data berupa FNA (*FASTA Nucleic Acid*) (Richter, Ott, Auch, Schmid, & Huson, 2008).

Panjang fragmen ditentukan sebanyak 10.000 bp dengan jenis distribusi ukuran Klon DNA Normal yang berarti bahwa parameter ke dua sebanyak 200. Konfigurasi MetaSim dapat dilihat pada Gambar 4.

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v17i1.5779



Gambar 4. Konfigurasi data pada MetaSim

Setelah itu selanjutnya lakukan *run with selected configuration*, sehingga diperoleh hasil frgamen seperti yang terlihat pada Gambar 5.


TTCATCGACCGTTTTGAAAACCGGCTTTCGCCGGATGCGGACGTGACTCACGGATCAGAAAAAAT
CATAGTTACCCCTGACTCCACAACTTACTTTACAACGACGTAGACCGTAACGTGCCGGCGATCGA
ACAGGTTTCAAGGGGTAATTTACGATGTTGCGATTTACGGTGTGTACCGTTCCCGCGCCACGCGC
ACATTGTAGGCTCGCTGCTGAACTCGGCATAGAATTCAAGCATGTGCCCTGTCATTCAGGCCCGC
CGCCTTGCCGATCCGCTCGCCACTGATGCGCCGTCTCAACACGCTGCGCCGGCATTTCTCGCCGT
CAACCCGATGGGCACCATTCCCTCGCACGAGGATGACGGGCGATGGTGCTTTACGAATCCATGGC
CATCAATCGTCTATCTCGCCCGCAAATATGGCGGGCCGCTCGACCCCGGTCGATCTGAAGGAAGA
TGCGGCGATGCTGCGAATGGTCCTTCTTCGCTGCCACTGAAATCGAAACGAA
GATACCTGCCTGAGCGCCACCTGATGGCGCACATCCAGCGTTGCCGCGATAATAGGGCGACCCAT
TCGGCACCAGCAGGATTTACCGCCCCGCTTTCGGCCAGCGTCTCGCACACGCCCATGTCGTTCCA
GATTTCCTCGCAGATCGGAATGCCGATCCGCACGCCACGGAAATTATACGGGCCCGGAAATCGAC
CCTTCTTCGAAAACACGTTTTTCGTCGAACTCGCCGTAAGTTCGGCAGATCGATCTTGTCACGCA
ACGCGATGATCTTGCCACCATCCAGAAGCGCCACCGAATTGTGCCGGCCTGTCTCGCCCTGACGC
GGAAAACCGATGACGACGCCCGGCCCGCCATCGGCAGTCTCCGCCGCCAGTTCTTCCACAGCCTT
CAGGCAGGCCTTCGAAAAGCCGGCTTTAAGCGACCAGATCCTCCGGCGGATAGCCTGAAATGAAC
AGCTCCGTCAGGAGCAGAAGGTCGGCACCCTCGCGTCGCCGCATCGGTCCCGTGCC
TCTTGAGCATGGACCTGCCTTGCATTTTCCCAGAATCTGGCGTTATGCAAGTTCCTTGATCCCAT
TACGTGATCCGCCTTGCCGAGGGGGCTGAAAAGGCTGTGCCGGAGGCACTGGCGGATGATTGTGA
TCGCCGCGTGGGCCCGAGAAAAGCCTTGTAGGTTTTCGCGTGCTTTGCGTCTACAATGAATAACG
CCTGCCGGGTCGGTTCGTCGTCGGCGGCGTTTAAAAACAAAAAACGAACCTTCGTTATGGGGCCG
GCAGAACCGGTCCATTCCCATGAGGAAAGTGATGGAAGAGTTTCATAAGGTCCGGCGTCTGTCCG
CAATACGTCTTCGAACAGGTCAACCGTTTTGAAAGCGAGCGCGCGAGCGGCCGGGGCCGATATCA
TCGATCTCGGCATGGGGTAATCCCGACCTTCCGACCCCCAAGGCCATCGTCGACAAGCTCTAGCG
AAGTCGTTCAGGACCCGCGCACCATCGTTATTCCTCGTCCAAGGGCATTCCCGGCCTTCGCCGCG
CACAGGCCCGGCTATTATGCCCGCCGTTTCGGCGTG

Gambar 5. Sampel Hasil Data Praproses Setelah Disimulasikan Menggunakan MetaSim

Setelah mendapatkan hasil berupa fragmen untuk setiap spesies, maka selanjutnya dilakukan ekstraksi menggunakan *K-Mers*. Ekstraksi fitur dilakukan untuk setiap fragmen dan dihitung frekuensi kemunculan dari setiap *substring* sehingga dari tahapan ini menghasilkan matriks komposisi yang berisi frekuensi kemunculan. Adapun nilai K yang digunakan pada *K-Mers* adalah 3.

Hasil fragmen dari ketiga Genus sebanyak 90.000 fragmen seperti yang terlihat pada Tabel 4. Selanjutnya data tersebut dibagi menjadi data latih sebesar 75% dan data uji sebesar 25% atau dengan kata lain data latih sebanyak 67.500 fragmen, dan data uji sebanyak 22.500 fragmen.

Tabel 4. Hasil Ekstraksi Fitur menggunakan *K-Mers*

Fitur Fragmen	**AAA**	**AAT**	**AAG**	**…**	**CCC**	**Genus**
F1	7	1	6	…	7	*Agrobacterium*
F2	2	6	6		5	*Agrobacterium*
..	…	…	…	…	…	…
F90000	14	12	4	…	3	*Staphyloccocus*

Setelah pembagian data, selanjutnya mengembangkan model klasifikasi KNN menggunakan bahasa pemrograman python. Klasifikasi dilakukan berdasarkan *voting* terbanyak diantara klasifikasi dari k objek. Algoritma KNN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari *query instance* yang baru (Salim & Mayary, 2020). Parameter KNN yang dikembangkan menggunakan nilai K = 3 sampai K = 15 dan rumus jarak *minkowski*. Hasil perhitungan akurasi untuk masing-masing nilai K dapat dilihat pada Tabel 5.

Tabel 5. Perhitungan Akurasi untuk Setiap Nilai K

Nilai K	**Akurasi**
3	94,37%
5	94,23%
7	94,17%
9	94,13%
11	93,95%
13	93,27%
15	93,07%

Berdasarkan hasil perhitungan akurasi tersebut, dapat diketahui bahwa semakin tinggi nilai K maka akan menyebabkan penurunan akurasi. Perhitungan akurasi tertinggi diperoleh pada nilai K = 3 dengan akurasi sebesar 94,37%. Hal ini menunjukkan bahwa algoritma KNN dapat digunakan untuk melakukan klasifikasi fragmen metagenom dengan baik dan memiliki akurasi yang cukup tinggi. Hasil *confusion matrix* untuk nilai K = 3 dapat dilihat pada Tabel 6.

Tabel 6. *Confusion Matrix* untuk Nilai K = 3

Prediction Actual	***Agrobacterium***	***Bacillus***	***Staphyloccocus***
Agrobacterium	7.516	23	11
Bacillus	269	6.532	674
Staphyloccocus	9	280	7.186

Dari 22.500 fragmen data uji diperoleh 21.234 fragmen yang berhasil diprediksi dengan benar dengan rincian 7.516 fragmen Genus *Agrobacterium*, 6.532 fragmen Genus *Bacillus*, dan 7.186 fragmen Genus *Staphyloccocus*. Kesalahan prediksi banyak terjadi pada Genus *Bacillus* sebesar 674 fragmen yang terprediksikan ke Genus *Staphyloccocus*.

5. KESIMPULAN

Berdasarkan hasil yang diperoleh, klasifikasi fragmen metagenom menggunakan algoritma KNN berdasarkan ekstraksi fitur *K-Mers* dapat dilakukan dengan baik. Nilai K pada *K-Mers* yang digunakan adalah 3, sehingga menghasilkan sebanyak 64 fitur. Dari hasil pengujian yang dilakukan diketahui bahwa semakin rendah nilai K pada KNN yang digunakan maka semakin tinggi akurasi yang didapatkan. Akurasi tertinggi diperoleh untuk nilai K = 3 pada KNN, yaitu 94,37% dengan kesalahan sebesar 5,63%. Sedangkan dengan nilai K = 15 pada KNN menghasilkan akurasi yang lebih rendah yaitu sebesar 93,07% dengan kesalahan sebesar 6,93%.

6. DAFTAR PUSTAKA

Choiriyati, N., Arkeman, Y., & Kusuma, W. A. (2020). Deep learning model for metagenome fragment classification using spaced k-mers feature extraction. *Jurnal Teknologi Dan Sistem Komputer*, *8*(3), 234–238. https://doi.org/10.14710/jtsiskom.2020.13407

e-ISSN 2597-4963 dan p-ISSN 1858-4853 DOI : https://dx.doi.org/10.30872/jim.v17i1.5779

Guritno, H. B., Haryanto, T., Kustiyo, A., & Hermadi, I. (2018). Optimasi Parameter pada Fast Correlation Based Filter Menggunakan Algoritma Genetika untuk Klasifikasi Metagenome. *Jurnal Edukasi Dan Penelitian Informatika*, *4*(2), 76–83.

Liantoni, F. (2015). Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor. *Jurnal ULTIMATICS*, *7*(2), 98–104. https://doi.org/10.31937/ti.v7i2.356

Manik, F. Y., Saputra S, K., & Br Ginting, D. S. (2020). Plant Classification Based on Extraction Feature Gray Level Co-Occurrence Matrix Using k-Nearest Neighbour. In *Journal of Physics: Conference Series* (Vol. 1566, pp. 1–9). https://doi.org/10.1088/1742-6596/1566/1/012107

Mutmainnah, U., Setiawan, B. D., & Dewi, C. (2019). Pengaruh Seleksi Fitur Information Gain pada K-Nearest Neighbor untuk Klasifikasi Tingkat Kelancaran Pembayaran Kredit Kendaraan. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *3*(9), 8882–8888.

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Pekuwali, A. A., Kusuma, W. A., & Buono, A. (2020). Seleksi Fitur Yang Berpengaruh Menggunakan Nilai Mean Pada Klasifikasi Fragmen Metagenome. *Jurnal Komputer Dan Informatika*, *8*(1), 9–17. https://doi.org/10.35508/jicon.v8i1.2188

Prasetio, R. T., Rismayadi, A. A., & Anshori, I. F. (2018). Implementasi Algoritma Genetika pada k-nearest neighbours untuk Klasifikasi Kerusakan Tulang Belakang. *Jurnal Informatika*, *5*(2), 186–194. https://doi.org/10.31311/ji.v5i2.4123

Richter, D. C., Ott, F., Auch, A. F., Schmid, R., & Huson, D. H. (2008). MetaSim - A sequencing simulator for genomics and metagenomics. *PLoS ONE*, *3*(10). https://doi.org/10.1371/journal.pone.0003373

Ridok, A., & Latifah, R. (2015). Klasifikasi Teks Bahasa Indonesia Pada Corpus Tak Seimbang Menggunakan NWKNN. *Konferensi Nasional Sistem Dan Informatika 2015*, (Oktober), 222–227.

Salim, S. S., & Mayary, J. (2020). Analisis Sentimen Pengguna Twitter Terhadap Dompet Elektronik Dengan Metode Lexicon Based Dan K – Nearest Neighbor. *Jurnal Ilmiah Informatika Komputer*, *25*(1), 1–17. https://doi.org/10.35760/ik.2020.v25i1.2411

Sari, R. (2020). Analisis Sentimen Pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma K-Nearest Neighbor (K-NN). *EVOLUSI : Jurnal Sains Dan Manajemen*, *8*(1), 10–17. https://doi.org/10.31294/evolusi.v8i1.7371

Surianti, S. (2020). Classification Fragmen Metagenom Menggunakan Principal Component Analysis Neighbor. *Jurnal Ilmiah Matrik*, *22*(2), 170–176. https://doi.org/10.33557/jurnalmatrik.v22i2.921

Syaljumairi, R., Defit, S., Sumijan, S., & Elda, Y. (2021). Akurasi Klasifikasi Pengguna terhadap Hotspot WiFi dengan Menggunakan Metode K-Nearest Neighbour. *Jurnal Sistim Informasi Dan Teknologi*, *3*. https://doi.org/10.37034/jsisfotek.v3i3.152

Syarifuddin, F., Misdram, M., Widodo, A. A., Informatika, P. S., & Pasuruan, U. M. (2020). Klasifikasi Data Set Virus Corona Menggunakan Metode Naïve Bayes Classifier. *Jurnal SPIRIT*, *12*(2), 46–52.

Utami, D. K., Kusuma, W. A., & Buono, A. (2017). Klasifikasi Metagenom dengan Metode Naïve Bayes Classifier. *Jurnal Ilmu Komputer Dan Agri-Informatika*, *3*(1), 9–18. https://doi.org/10.29244/jika.3.1.9-17