

Implementasi Algoritme *Fuzzy K-Nearest Neighbor* Untuk Klasifikasi Urutan Metagenom

Deni Satria¹⁾, Harmaini²⁾

¹ Fakultas Hukum, Universitas Merangin

² FKIP, Universitas Merangin)

email: kkdeni.saviola87@gmail.com, email: harmaini83@gmail.com

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 menganalisis 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 fuzzy 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 fuzzy KNN maka semakin tinggi akurasi yang diperoleh. 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 94,57% dengan kesalahan sebesar 5,43%. Sedangkan dengan nilai K = 15 pada KNN menghasilkan akurasi yang lebih rendah yaitu sebesar 93,12% dengan kesalahan sebesar 6,88%. Hasil klasifikasi fragmen metagenom menggunakan algoritma fuzzy KNN berdasarkan ekstraksi fitur K-Mers dapat dilakukan dengan baik.

Keywords: klasifikasi, fuzzy k-nearest neighbor, metagenom, ekstraksi fitur, k-mers

PENDAHULUAN

Penelitian tentang analisis metagenom dalam lingkup bioinformatika terus berkembang. Secara umum, analisis materi genetik dilakukan dengan cara membudidayakannya di laboratorium, kemudian di-sequencing dan dilakukan perakitan. Proses ini dilakukan untuk menghasilkan urutan rantai DNA yang berisi informasi genetik suatu organisme. Akan tetapi, dari banyak mikroorganisme hanya 1% yang dapat dikulturkan[1]. Sisanya harus mengambil sampel langsung dari lingkungan. Ilmu yang mempelajari tentang

analisis metagenom dan materi genetiknya diperoleh langsung dari sampel lingkungan disebut metagenomika. Sampel ini ketika di-sequencing akan menghasilkan fragmen-fragmen. Fragmen-fragmen yang berasal dari berbagai organisme. Pada saat dilakukan perakitan fragmen-fragmen ini, akan menghasilkan chimeric contigs gabungan fragmen yang berasal dari organisme berbeda [2].

Pertama, data metagenomik dihasilkan dari komunitas mikroba yang diminati. Setelah melakukan prosedur kontrol kualitas, urutan metagenomik dapat dikenai berbagai analisis yang berpusat pada karakterisasi taksonomi dan fungsional komunitas. analisis gen penanda, binning, dan perakitan memberikan wawasan tentang keragaman taksonomi atau filogenetik komunitas dan dapat mengidentifikasi taksa atau genom baru. Metagenom juga dapat dapat memprediksi gen dan anotasi fungsional, yang dapat digunakan untuk mengkarakterisasi fungsi biologis yang terkait dengan komunitas dan mengidentifikasi gen baru[3].

Dengan kemajuan terbaru dalam teknologi pengurutan dan alat komputasi, tantangan ini telah dicoba untuk diatasi dengan membangun genom rakitan metagenom (MAG). MAG mengacu pada sekelompok perancah dengan karakteristik serupa dari rakitan metagenom yang mewakili genom mikroba . Dalam pendekatan ini, pembacaan sekuensing dirangkai menjadi perancah dan kemudian perancah dikelompokkan menjadi kandidat MAG berdasarkan frekuensi tetranukleotida (TNFs), kelimpahan, gen penanda pelengkap , penyelarasan taksonomi dan penggunaan kodon . MAG dengan kelengkapan tinggi dan tingkat kontaminasi rendah kemudian digunakan untuk anotasi taksonomi lebih lanjut dan prediksi gen[4].

Berdasarkan hasil studi literatur dari berbagai artikel jurnal ilmiah dan kajian yang telah dipublikasikan diketahui bahwa sebagian besar penelitian sebelumnya tentang mikroba sangat bergantung pada ketersediaan genom referensi. Mereka telah melibatkan penyelarasan langsung data sekuensing terhadap genom referensi, set gen penanda atau sekuens spesifik spesies untuk tugas taksonomi. Frekuensi k -mer dan kedalaman baca gen universal single-copy biasanya digunakan untuk memperkirakan kelimpahan taksonomi. Namun, karena genom referensi tidak lengkap, saat ini hanya mungkin untuk mengeksplorasi asosiasi mikroba dengan genom referensi berkualitas tinggi. Hal ini menjadi tantangan teknis untuk mengidentifikasi asosiasi yang melibatkan gen atau mikroba baru. Oleh karena itu, diperlukan kumpulan genom referensi yang dikarakterisasi dengan baik yang dihasilkan dari sekuensing metagenomik.

Untuk itu diperlukan metode yang membangun genom rakitan metagenom yang bertujuan untuk mengklasifikasikan fragmen-fragmen tersebut ke dalam tingkat taksonomi tertentu, yaitu metode algoritma Fuzzy K-Nearest Neighbor.

Proses binning (pengelompokan) dapat dilakukan dengan dua pendekatan yaitu berdasarkan komposisi dan homologi. . 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*. Pada proses binning (pengelompokan) berdasarkan komposisi melakukan perhitungan frekuensi ciri yang muncul dari pasangan basa (base pair) yang membentuk sekuens metagenom. Ciri komposisi digunakan sebagai masukkan pembelajaran dengan contoh (*supervised learning*) atau pembelajaran secara observasi (*unsupervised learning*). proses binning (pengelompokan) berdasarkan homologi melakukan pencarian penjajaran sekuens dengan membandingkan fragmen metagenom dengan sekuens referensi yang terdapat

pada basis data yang digunakan, misal National Centre for Biotechnology Information (NCBI). Hasilnya disimpulkan pada tiap level taksonomi. Hal tersebut menyebabkan pendekatan dengan homologi membutuhkan banyak waktu dalam proses pengelompokan [5].

Sedangkan metode fuzzy merupakan system yang didasarkan pada aturan ataupun pengetahuan yang meliputi keseluruhan proses pengolahan input tegas menjadi output tegas menggunakan fuzzifier, fuzzy rule base, fuzzy inference engine dan defuzzifier.

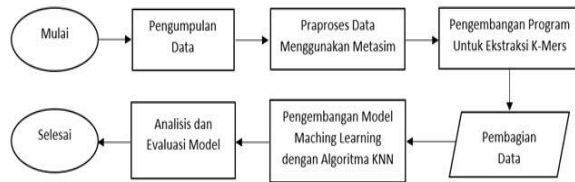
Beberapa hal yang perlu diketahui dalam system fuzzy adalah variable fuzzy, himpunan fuzzy, semesta pembicaraan dan domain. Variable fuzzy adalah variable yang akan dibahas dalam suatu system fuzzy. Himpunan fuzzy merupakan suatu grup yang mewakili suatu kondisi tertentu dalam suatu variable fuzzy. Semesta pembicaraan adalah interval nilai yang diperbolehkan untuk dioperasikan dalam suatu variable fuzzy. Sedangkan domain adalah interval nilai yang diperbolehkan untuk dioperasikan dalam himpunan fuzzy [6].

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 ekstraksi fitur *K-Mers*. Besarnya jumlah parameter n pada metode ekstraksi fitur *K-Mers* akan mengakibatkan dimensi fitur yang tinggi. Metode *K-Nearest Neighbor* (K-NN) merupakan metode klasifikasi data yang cara kerjanya relatif lebih sederhana bila dibandingkan dengan metode klasifikasi lainnya. Namun pada metode *K-Nearest Neighbor* kurang sesuai untuk pengklasifikasian yang cenderung mengatur jumlah data latih yang memiliki nilai mayoritas. Karena prinsip mayoritas *K-Nearest Neighbor* mengklasifikasikan data baru berdasarkan data mayoritas kelas yang ada pada sejumlah K tetangga terdekat dari data baru tersebut. Oleh James M. Keller ditemukan suatu metode baru untuk mengatasi permasalahan ini, yaitu digunakannya metode *Fuzzy K-Nearest Neighbor* yang merupakan gabungan dari metode fuzzy dengan *K-Nearest Neighbor* (Wisdarianto, Ridok dan Rahman, 2013). Namun pada Metode *Fuzzy K-Nearest Neighbor* memiliki kelemahan yaitu antara sejumlah k tetangga dianggap sama pentingnya padahal belum tentu, hal tersebut diatasi dengan perhitungan derajat keanggotaan data pada tiap kelas pada persamaan 2-2. Selain itu pada data latih belum dapat diketahui kekuatan keanggotaannya pada sebuah kelas dan untuk mengatasi hal tersebut menggunakan inisialisasi fuzzy (Keller, Gray dan Givens, 1985).

Adapun tujuan penelitian ini adalah untuk mengetahui dan mendeskripsikan implementasi klasifikasi urutan metagenom menggunakan metode *Fuzzy K-Nearest Neighbor* dan untuk mengetahui tingkat akurasi hasil klasifikasi fragmen metagenom menggunakan metode *Fuzzy K-Nearest Neighbor*

METODE PENELITIAN

Penelitian ini dilakukan dengan mengikuti beberapa tahapan proses, yaitu pengumpulan data fragmen metagenom, pembagian data (data latih dan data uji), pra-proses data, ekstraksi fitur dengan metode *k-mers*, reduksi dimensi dengan *PCA*, *k-fold cross validation*, pelatihan dan pengujian dengan metode algoritma fuzzy *k-nearest neighbor*, serta tahap analisis. Adapun tahapan penelitian dapat dilihat pada gambar 1 di bawah ini



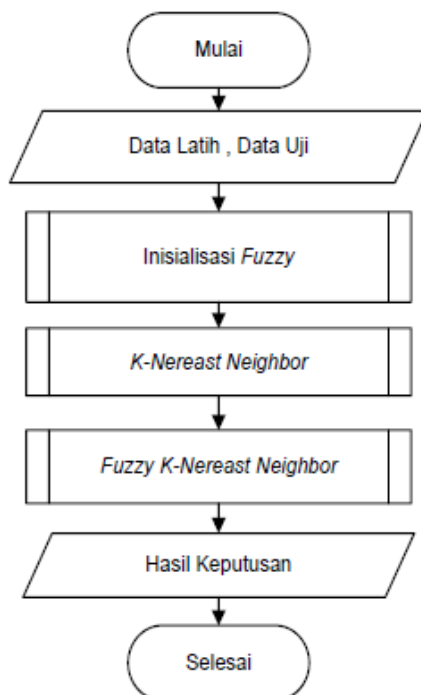
Gambar 1. Tahapan Penelitian

Data yang digunakan pada penelitian ini adalah data metagenome yang diunduh dari situs *National Centre for Biotechnology Information (NCBI)*. *NCBI* merupakan salah satu institusi yang menjadi rujukan atau sumber informasi perkembangan biologi molekuler. Data metagenom yang digunakan ini merupakan *simulated data* berupa fragmen-fragmen DNA yang dibangkitkan dengan menggunakan aplikasi *MetaSim* dari *sekuens genom lengkap bakteri yang diunduh dari NCBI*. *MetaSim* mensimulasikan kinerja *sequencer* untuk menghasilkan *fragmen metagenom yang direpresentasikan sebagai string dengan format data berupa FNA (FASTA Nucleic Acid)*.

Aplikasi yang akan dibuat merupakan aplikasi yang mengimplementasikan metode *fuzzy k-nearest neighbor* untuk mengklasifikasi urutan metagenom. Berikut langkah-langkah yang dilakukan dalam pengklasifikasian urutan metagenom:

1. Masukkan data latih dan nilai k .
2. Proses *k-nearest neighbor* untuk mencari nilai jarak antara data latih terhadap data uji menggunakan *Euclidean Distance*.
3. Proses *fuzzy k-nearest neighbor* dengan mencari nilai keanggotaan data.
4. *Output* berupa kelas tepat atau tidak tepat.

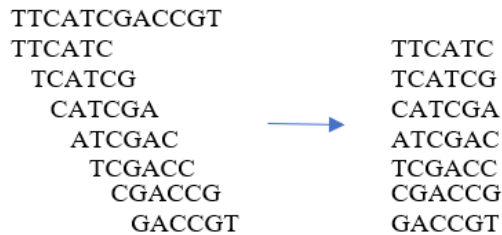
Tahapan-tahapan diatas digambarkan dalam Gambar 2:



Gambar 2. Alur proses klasifikasi

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 3. Ilustrasi dan Ekstraksi Fitur

Sebagai contoh untuk fragmen seperti berikut:

Fragmen1: AAAATCGACCCTTTTGAAG

Fragmen2: GAGAATCGACCCTTTTGAAT

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	<i>Agrobacterium</i>
F2	0	2	0	...	0	<i>Agrobacterium</i>
...	...	...	...	...	...	...
F90000	...	...	...	...	...	<i>Staphylococcus</i>

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, Misdrum, 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).

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).

n

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

Keterangan:

d = jarak antara x dan y

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

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*.

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*

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

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

Pengujian pada penelitian ini menggunakan pengujian akurasi. Proses pengujian dilakukan sebanyak empat tahap pengujian, yaitu :

1. pengujian nilai k
2. pengujian variabel m (bobot pangkat)
3. pengujian jumlah variasi data
4. pengujian perbandingan metode *Fuzzy K-Nearest Neighbor* dan *K-Nearest Neighbor*

1. 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
CCTAGTTGCCCCGAATCATAGAAATCGGAGTACTTCTATCTACGAAATGCCGTTAGTACCGATTCT
GCGTAAATAAAGAAACAGTCAACAAACCAACCATAAAGGCGCTTAGTAACTGGCCAAAATGGGAG
AAAACCATGACCCGACATCTGAGCAGCTCGCTTCTATCGAGACCTTTTCTGCAAGCTGGATACCGAT
TGGCAATCTCAGCTAGCCAGTTTCCCGCGATGCCGCGCAGCATGCGCAAGTTCACTTCCCCCGA
GGTTGCCGCTGCTTAACGTCACGGAAGCTATATTAGGAGGTTTCCCTGAAGGAGCAGGGGCGACG
CCGGAAGTCGCAATGGAGCGCTCTATACGCTCAGCAGATCTCAGTTGCGTATGCACTTGGCGC
AGAAGCGCCGAAGAAGTGGATGAACCGCGCAGGGTAGGCGAGGAGCAATGCCAGATCTGGCAATCAC
CAATTTCAAGGGCGGCTCAGCAAGACCTCCACCACGATCTCTCGGTCAATTATCTGGCCTTGAGAGGA
TACCGCTGCTGGCGTTGACCTCGACCCCAAGGACATCGCTGACGAGCTGCATGGTAGTCTGCCGATT
TCGATTATCTGAGCGGATACGCTTATTCGCGCATCCGCTCGAGGATCCGATCCCGACCGAGCAGGT
CATTCATCAGACCCACATCGCCGCTCGATGCTCATCTGCGCGCGCTGGATTGACGGGTTGCAAAAC
GCGCTTGCCTTGAATGCGCAAGCTCGCGGACGAGCTTCTCTTGGCGTATCGCAGGCACTGGAGC
AGGTGGCTGATCGTCAGCAGATCATCTGATGGATTGCGGCCGCTGCTGAACCTTCTCACTTGTCTC
CCTGACCGCGCGACAGGCGTCTGATACCGGTTCCGGCACATATGCTCGATGGATTCCACAGCGAAG
TTCTTGAACCTCGCGGATCTACATCGAGATCTCGAAGGTCGCGACGTCGCGCAATGGGATTTCCG
CCAAGTTCTGATCACCAAGTTCGAACCGAACGACCATCCCGAGGCAATATGACGGCGTGTATGCGCA
GGTTTTCGGCGAAGACCTGTTGCTTAACCTCGGTGATCAATCGACCGGTTGCCGATGCTTGACCTGG
AAACGAGGCTTTACGAGGTGACAGGCTCGAGATTTTCCGCGCCGAAACATATGACAGGGCGATTGAGT
CGATCAATCGGCAATGCCAAATCGAAGAGCTCTTTTGAAGGATCGGGACGCGAATGAGCAGGAAA
GATTCCAAAGGATGTTTGGCAATGCTGATCGGTTCTCGGCGAGGACCTCATCTCGTACCGCGCGCA
CGACATCGCGCTCGCGCATCTCTGAGGTCGCGCGCGGCTTCCGAGTTGCAAGGACGTGGCGAGCT
TGCGCAAGCTGCTGAGGTTGCGGATCACATCATCGAGCTCGACGTGGACGCGGCTTCCGCTTCTCC
ATTGCGGATCGATTGACGGAGCTTATGACGCGCAGGCTTCAAGGATCTCGTGCAGTCCATCGCGCAAC
ACGGGCAACGACGCTGCTGCTTCCGCGCATGCGCGGCGAGGGCAGCGGTTTCAAGATCGCTTCCG
```

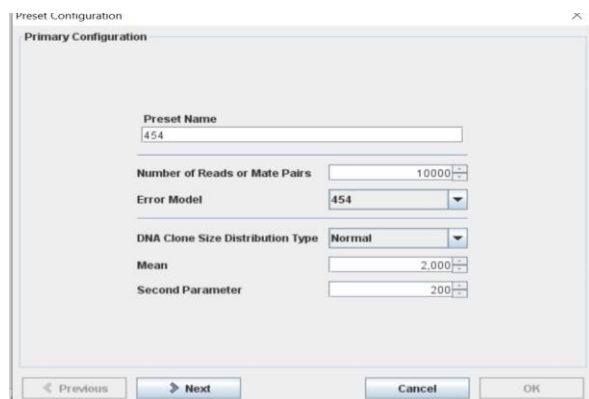
Gambar 3. Sampel Data *Agrobacterium Radiobacter K84 Chromosome 2*

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.



Gambar 4. Konfigurasi data pada MetaSim

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

```
TTTCATCGACCGTTTTGAAACCGGCTTTCGCGGATGCGGACGTGACTCACGGATCAGAAAAAAT
CATAGTTACCCCTGACTCCCAACTTACTTTACAACGACGTAGACCGTAACGTTGCCGGCATCGA
ACAGGTTTCAAGGGGTAAATTACGATGTTGCGATTACGGGTGTACGTTCCGGGCGACGCG
ACATTGTAGGCTCGCTGCTGAATCGGCAATGAATTAAGCATGTGCCGTGCTCATTACGGCCCG
CGCTTGGCGATCGCTCGCCACTGATGCGCGTCTCAACACGCTGCGCGGCAATTTCTGCGCGT
CAACCGGATGGGACCATTCCTCGCACGAGGATGACGGGCGAGTGGTCTTACGAATCCATGGC
CATCAATCGTCTATCTCGCCGCAAAATATGGGCGGCGCTCGACCCCGTTCGATCTGAAGGAAGA
TGGCGGATGCTGCGAATGCTCTTCTCGTGCCTGAAATCGAAACGAA
GATACCTGCTGAGCGCCACCTGATGGGCGACATCCAGCGTTGCCGCGATAATAGGGCGACCCAT
TCGGCACGAGGAGTTTACCGCCCGCTTTCGGCGAGCGTCTCGCACGCGCCATGCTGTTCCA
GATTTCTCGCAGATCGGAATGCCGATCGCGACGCGACGGAATATACGGGCGCGGAATCGAC
CCTTCTCGAAACACGTTTTTCTGTCGAACTCGCCGTAAGTTTCGGCAGATCGATCTTGTACGCA
ACGCGATGATCTTGGCACCATCCAGAAGCGCCACCGAATTGTCGCGGCTGTCTCGCCCTGACGC
GGAAACCGGATGACGACGCGCCCGCGCATCGGAGTCTCGCGCGCAGTTCTTCCACAGCCTT
CAGCGAGCCTTGAAGAGCGGCTTAAAGCGACAGATCTCGCGGAGTACGCTGAAATGAAC
AGCTCCGTCAGGAGCAGAAGGTGCGCACCTTCGCTCGCGCATCGCTCCGTCG
TCTTGAGCATGGACCTGCTTGCATTTTCCGGAATCTGGCGTATGCAAGTTCTTGTATCCCAT
TACGTGATCCGCTTGGCGAGGGGCTGAAAGGCTGTGCGGAGGCACTGGCGGATGATTGTA
TCGCGCGTGGGCGGAGAAAGCCTTGTAGGTTTTCGCGTGTCTGCTGCTACAATGAATAACG
CCTCGCGGCTCGGTTCTGCTCGCGGCGGTTTAAAAACAAAAACGAACCTTCTGTTAGGGGCG
CGAGAACCGGTCATTCCCATGAGGAAAGTGTGGAAGAGTTTATAAGGTCGGGCGTCTCGG
CAATACGCTTTCGAACAGGTCAACGTTTTGAAAGCGAGCGCGAGCGCGCGGGGCGGATATCA
TCGATCTCGGATGGGTAATCCGACCTTCCGACCCCAAGGCCATCGTCGACAAAGCTCTAGCG
AAGTCTTCAAGACCGCGCACCATCGTTATTCCTCGTCAAGGGGATTCCGGGCTTCCGCGG
CACAGGCCGCGTATTATGCCGCGTTTCGGCGTG
```

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 100.000 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 75.000 fragmen, dan data uji sebanyak 25.000 fragmen.

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

Fitur Fragmen	AAA	AAT	AAG	...	CCC	Genus
F1	7	1	6	...	7	<i>Agrobacterium</i>
F2	2	6	6		5	<i>Agrobacterium</i>
..	...	...	...	...	...	...
F1000000	14	12	4	...	3	<i>Staphylococcus</i>

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 <i>K</i>	Akurasi
3	94,57%
5	94,33%
7	94,20%
9	94,11%
11	93,85%
13	93,37%
15	93,12%

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 94,57%. 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 $K = 3$

<i>Prediction Actual</i>	<i>Agrobacterium</i>	<i>Bacillus</i>	<i>Staphylococcus</i>
<i>Agrobacterium</i>	8.406	30	15
<i>Bacillus</i>	325	7.370	757
<i>Staphylococcus</i>	5	335	7.599

Dari 25.000 fragmen data uji diperoleh 23.735 fragmen yang berhasil diprediksi dengan benar dengan rincian 8.406 fragmen Genus *Agrobacterium*, 7.370 fragmen Genus *Bacillus*, dan 7.599 fragmen Genus *Staphylococcus*. Kesalahan prediksi banyak terjadi pada Genus *Bacillus* sebesar 757 fragmen yang terprediksikan ke Genus *Staphylococcus*.

KESIMPULAN

Berdasarkan hasil yang diperoleh, klasifikasi fragmen metagenom menggunakan algoritma Fuzzy 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 94,57% dengan kesalahan sebesar 5,43%. Sedangkan dengan nilai $K = 15$ pada KNN menghasilkan akurasi yang lebih rendah yaitu sebesar 93,12% dengan kesalahan sebesar 6,88%

DAFTAR PUSTAKA

1. Maguire F, Jia B, Gray KL, Lau WYV, Beiko RG, Brinkman FSL. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb Genomics*. 2020;6(10):1–12.
2. Bagus I, Manuaba P, Kom S. Sistem penjadwalan produksi menggunakan metode fuzzy support vector machines dan algoritma evolusi fuzzy. *Smart-Green Technol Electr Inf Syst*. 2013;(November):14–5.
3. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci*. 2014;5(JUN):1–14.
4. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, Bian Z, et al. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J [Internet]*. 2021;19:6301–14. Available from: <https://doi.org/10.1016/j.csbj.2021.11.028>
5. Utami DK, Kusuma WA, Buono A. Klasifikasi Metagenom dengan Metode Naïve Bayes Classifier Metagenome Classification Using Naïve Bayes Classifier Method. *J Ilmu Komput Agri-Informatika*. 2014;3(1):9–18.
6. Surianti S. Classification Fragmen Metagenom Menggunakan Principal Component Analysis Neighbor. *J Ilm Matrik*. 2020;22(2):170–6.

7. 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>
8. 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.
9. 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>
10. 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>
11. 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.