KLASIFIKASI DNA MENGGUNAKAN FITUR N-MERS DENGAN INTEGRASI SELEKSI DATA DAN ELM (IDELM) SEBAGAI *CLASSIFIER*

Umi Mahdiyah[1], Lilia Sinta Wahyuniar[2], Siti Rochana[3]
[1, 2, 3)] Teknik Informatika, Fakultas Teknik, Universitas Nusantara PGRI Kediri
Jl. Ahmad Dahlan, Mojoroto Gg I, Kota Kediri,
E-mail: umimahdiyah@gmail.com

ABSTRAKS

DNA merupakan unsur yang sangat penting dan mendasar pada setiap organisme. Sekuensing DNA dapat dimanfaatkan untuk menentukan identitas suatu organisme dengan cara membandingkan urutan DNA nya dengan DNA lain yang sudah diketahui. Integrasi Seleksi data dan Extreme Learning Machine (IDELM) ini dipilih karena data DNA merupakan data yang besar serta karakteristik datanya yang kebanyakan adalah data yang imbalance. Pada proses penelitian data yang akan diolah terlebih dahulu diuraikan fragmennya dengan simulator MetaSim, selanjutnya dilakukan ekstraksi fitur dengan menggunakan n-mers, kemudian dilakukan proses klasifikasi dengan IDELM. Hasil dari pengklasifikasian tersebut memiliki performa yang baik, karena dengan ekstraksi fitur 3-mers maupun 4-mers performanya di atas 80%.

Kata Kunci: DNA, n-mers, IDELM

ABSTRACTS

DNA is a very important and fundamental element in every organism. DNA sequencing can be used to determine the identity of an organism by comparing its DNA sequence with other known DNA. The data selection and Extreme Learning Machine (IDELM) integration was chosen because DNA data is big data and the characteristics of the data are mostly imbalance data. In the research process, the fragments with MetaSim simulator are then broken down, then the feature extraction is done using n-mers, then the classification process is done with IDELM. The results of the classification have good performance, because with the extraction of 3-mers and 4-mers features, the performance is above 80%.

Keywords: DNA, n-mers, IDELM

1. PENDAHULUAN

1.1 Latar Belakang

DNA adalah bagian yang sangat penting pada makhluk hidup. DNA merupakan sebuah polimer yang terdiri dari satuan-satuan berulang yang disebut nukleutida Tiap-tiap nukleotida terdiri dari tiga komponen utama, yakni gugus fosfat, gula deoksiribosa, dan basa nitrogen(nukleubasa). Pada DNA, nukleobasa yang ditemukan adalah Adenina(A), Guanina (G), Sitosina (C) dan Timina(T).

Sekuensing atau pengurutan DNA merupakan proses atau teknik penentuan urutan basa nukleutida pada suatu molekul DNA. Urutan tersebut dikenal sebagai sekuens DNA, yang merupakan informasi paling mendasar suatu gen karena mengandung instruksi yang dibutuhkan untuk pembentukan tubuh makhluk hidup. Sekuensing DNA dapat dimanfaatkan untuk menentukan identitas maupun fungsi gen atau fragmen DNA lainnya dengan cara membandingkan sekuens-nya dengan sekuens DNA lain yang sudah diketahui.

Pengklasifikasian DNA akhir-akhir ini sudah mulai dilaksanakan dengan pengaplikasian *machine learning*. Dalam penelitian ini digunakan integrasi seleksi data dan *Extreme Learning Machine*, karena data DNA kebanyakan merupakan data yang berkarakter *imbalance*. Sehingga diperlukan perlakuan khusus untuk data tersebut.

1.2 Referensi

1.2.1 *Sequence* DNA

Deoxyribonucleic acid (DNA) adalah polimer asam nukleat yang tersusun secara sistematis dan merupakan pembawa informasi genetik yang diturunkan kepada keturunannya. Informasi genetik disusun dalam bentuk kodon yang berupa tiga pasang basa nukelotida. DNA merupakan molekul paling terkenal saat ini, karena molekul ini merupakan substansi penurunan sifat. DNA merupakan suatu polimer heliks ganda yang terdiri dari nukleotida, setiap nukleotida terdiri dari tiga komponen; satu basa nitrogen, satu gula pentosa yang disebut deoksiribosa, dan satu gugus fosfat. Basa nitrogennya bisa adenin (A), timin (T), guanin (G), atau sitosin (S). Adenin dan dan guanin adalah purin, basa nitrogen dengan dua cincin organik. Sebaliknya, sitosin dan timin adalah anggota famili basa nitrogen yang dikenal sebagai pirimidin, yang mempunyai satu cincin tunggal (Rapley, 2015).

DNA merupakan materi genetik sel, sebelum mengalami mitosis, sel eukariotik dengan tepat menggandakan kandungan DNA-nya, dan selama mitosis, DNA ini akan terdistribusi tepat sama ke dua sel anaknya. Selain itu kromosom diploid mempunyai DNA dua kali lebih banyak daripada kromosom haploid yang ditemukan di dalam gamet organisme yang sama. Komposisi DNA pada setiap spesies berbeda-beda. (Saefudin,2017)

1.2.2 *Ekstraksi Fitur dengan n-mers*

Salah satu tahapan dalam supervised learning adalah tahap ekstraksi fitur. Tahap ini merupakan tahap pengambilan data ciri dari suatu objek. Dalam penelitian ini adalah ciri dari DNA yang dinyatakan dalam bentuk numerik. Metode *n-mers* ini digunakan untuk mengetahui banyaknya kemunculan *substring* tertentu pada sebuah *string*. Intensitas kemunculan *string* tersebut dapat dijadikan sebagai penciri dari suatu kelompok *string*. Data sekuens DNA termasuk dalam data string, sehingga ekstraksi ciri yang digunakan (Haryanto, 2018).

1.2.3 Integrasi Seleksi Data dan *Extreme Learning Machine*

Extreme Learning Machine pertama kali di usulkan oleh Huang pada tahun 2004, yang merupakan algoritma learning sederhana untuk *Single Hidden Layer Feedforward Networks* (SLFN) dengan kecepatan learning dapat sangat cepat dibandingkan *algoritma learning feedforward network* biasa.

Algoritma ELM merupakan algoritma yang *diperoleh* dari solusi *minimum norm least square* SLFNs. Konsep utama dari ELM seperti yang disajikan dalam *paper* Huang (2015) adalah sebagai berikut:

Diberikan *training* set

$$\aleph = \{(x_j, t_j) | x_j \in \boldsymbol{R}^{n \times m}, t_j \in \boldsymbol{R}^{n},\ j \in [1, ,$$

fungsi aktivasi g, dan bilangan *hidden node*

Step 1: masukkan secara random bobot dan bias ,

$$i \in [1$$

Step 2: hitung *output* matriks *hidden layer*

Step 3: hitung bobot *output*

$$\beta = H \qquad (1)$$

dengan

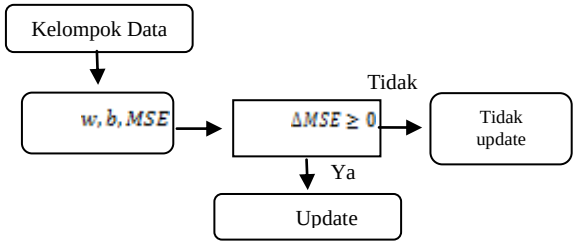
$$T = [t_1, \dots \dots, t,$$

adalah *Generalized Inverse*

$$H^{\dagger} = (H^T H)^{-1}$$ atau,

$$H^{\dagger} = H^T (H H^T$$

Pada ELM biasa, seluruh data langsung masuk dalam proses training secara bersama, akan tetapi dalam penelitian ini proses *training* dilakukan secara *sequential*, artinya data masuk proses *training* per kelompok data. Pada proses sequential tersebut terjadi proses seleksi data secara kelompok dengan melihat kemiripan datanya. Alur dari Integrasi seleksi data dan Extreme Learning Machine dapat dilihat pada Gambar 2.



Gambar 2 Proses *Training* ELM terintegrasi (Mahdiyah, 2017)

Keterangan:

$$\Delta MSE = MSE_{t+1} - MSE_t$$

1.3 Metode Penelitian

1.3.1 Pengambilan Data

Data yang digunakan dalam penelitian ini merupakan data eksperimental, data dalam bentuk FASTA yang diambil dari https://www.ncbi.nlm.nih.gov/. Selanjutnya FASTA tersebut dibagi menjadi beberapa fragmen dengan menggunakan MetaSim (Richter, 2008).

1.3.2 Ekstraksi Fitur

Ektraksi fitur dalam penelitian ini dilakukan dengan menggunakan *n-mers*. Nilai n yang digunakan adalah *3-mers* dan *4-mers*. Pola kemunculan dalam sekuens dihitung menggunakan empat basa utama (A, T, G, dan C) dipangkat dengan rangkaian pasangan basa yang ingin digunakan (pola kemunculan : 4^n, dengan n>= 1) (Kusuma, 2011).

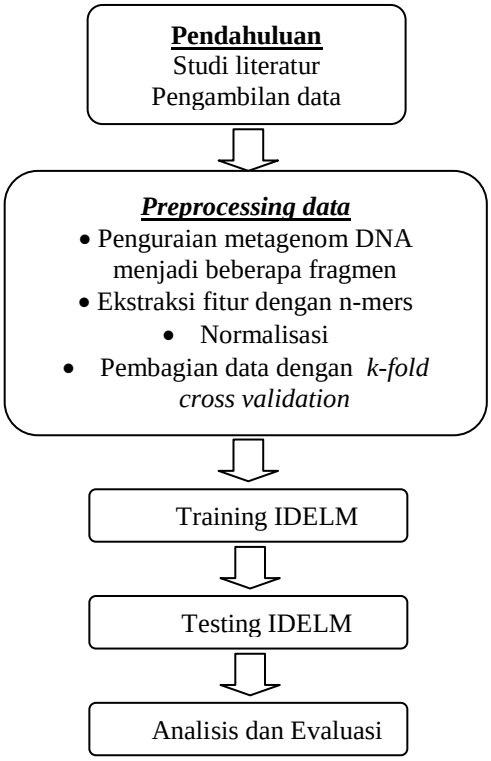
1.3.3 Data dan Pembagian Data

Data yang diambil diambil dari 2 genus yaitu genus “Bacillus” dan “Clostridium”. Masing-masing genus terdapat 10 data. Data tersebut selanjutnya dibagi fragmennya, kemudian dilakukan ekstraksi fitur dengan metode *n-mers*. Kemudian, data yang telah dinormalisasi tadi dibagi menjadi data training dan data testing dengan menggunakan 5-*fold cross validation*.

1.3.4 Prosedur Penelitian

Prosedur penelitian dapat dilihat pada Gambar 1.

ISSN: 2503-071X
E-ISSN: 2621-511X



Gambar 1. Prosedur Penelitian

Pada tahap awal dilakukan studi literatur dan pengambilan data. Data diambil dari web https://www.ncbi.nlm.nih.gov/, data diambil dari 2 jenis genus yang berbeda yaitu *Bacillus* dan *Clostridium*, masing masing genus diambil 10 data. Artinya dalam penelitian ini ada 2 kelas genus dan 20 organisme.

Pada tahap *prerocesing data* data yang diperoeh berupa metagenom DNA dalam bentuk file FASTA diuraikan fragmennya dengan menggunakan simulator MetaSim. Data fragmen tersebut selanjutnya diekstraksi fiturnya dengan menggunakan *n-mers*. Setelah proses ektraksi fitur, maka data numerik yang diperoleh dinormalisasi dengan normalisasi minmax.

$$x_{norm} = \left(\frac{x - x_{min}}{x_{max} - x_{min}} \times (max_{new} - min_{new})\right) + mi \quad (1)$$

Keterangan:

x_n= data hasil normalisasi
x_1 = nilai minimum dari data per kolom
x= nilai maximum dari data per kolom
min_1 = adalah batas minimum yang kita berikan
max = adalah batas maximum yang kita berikan

Jika data sudah dinormalisasi artinya data sudah siap dipakai, sehingga tahap berikutnyaadalah membagi data menjadi data training dan testing. Pada tahap pembagian data digunakan metode *k-fold cross validation*.

Pada tahap selanjutnya dari data yang sudah dibagi tersebut dilakukan proses training untuk mendapatkan model. Kemudian dilakukan testing berdasar model yang diperoleh, sehingga dapat dilihat seberapa bagus performa IDELM pada klasifikasi DNA dengan proses tersebut. Setelah ketahui performansinya maka langkah berikutnya adalah analisis dan evaluasi dari hasil yang diperoleh dengan *confusion matrix* untuk mendapatkan nilai *precision* dan *recall*, *specificity*, dan *G-mean*(Sokolova, 2009).

Precision adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

$$precision = \frac{TP}{TP + FP}. \quad (2)$$

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi**.**

$$recall = \frac{TP}{TP + FN}. \quad (3)$$

$sensitify = recall$

Geometric mean telah digunakan beberapa peneliti untuk mengevaluasi *classifier* pada dataset yang *imbalanced* . *G-mean* mengindikasikan keseimbangan antara kinerja klasifikasi pada kelas mayoritas dan minoritas. Ukuran *G-mean* diambil berdasarkan s*ensitify* (akurasi dari data positif) dan s*pecificity* (akurasi data negatif). [3]

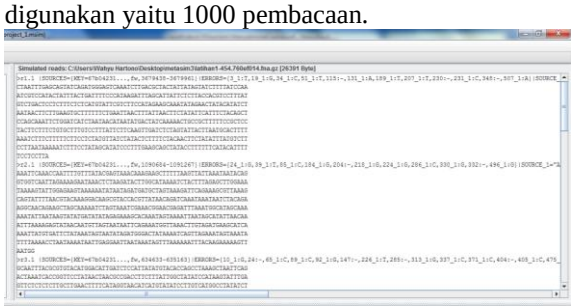
$$specificity = 1 - \frac{TP}{FP + TN} \quad (4)$$

$$G - mean = \sqrt{sensitivity \times specificity} \quad (5)$$

2. PEMBAHASAN

2.1 Penyiapan Data

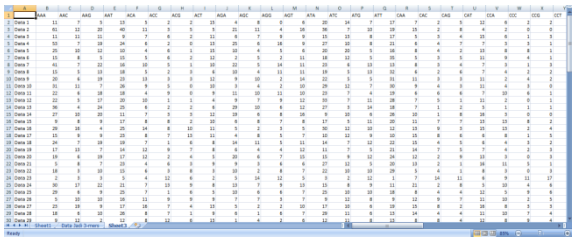
Data metagenom DNA yang diambil dari situs NCBI. Sequence DNA tersebut selanjutnya dibagi fragmennya dengan menggunakan simulator MetaSim. Data penelitian yang digunakan dalam penelitian ini berupa dataset kecil yang terdiri atas 20 organisme yang termasuk kedalam 2 genus. Data menggunakan panjang fragmen yang seragam, yaitu 400 bp, 800 bp, 1 kbp, 3 kbp, 5 kbp dan 10 kbp. Satuan bp (base pair) adalah banyaknya atau panjangnya unsur basa adenine (A), thymine (T), guanine (G) dan cytosine (C) suatu DNA. Banyaknya pembacaan dengan MetaSim yang digunakan yaitu 1000 pembacaan.



Gambar 2. *Screenshoot* pembagian fragmen menggunakan MetaSim

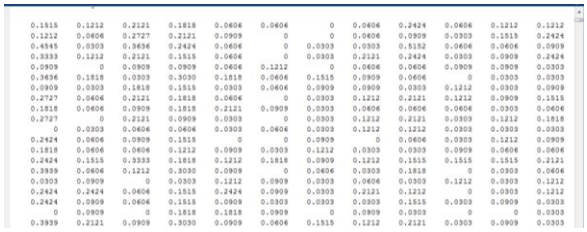
Selanjutnya dari data tersebut dilakukan ekstraksi fitur dengan menggunakan *n-mers*. Metode ini digunakan untuk mengetahui intensitas atau

banyaknya kemunculan *substring* tertentu pada sebuah *string*. Artinya untuk data DNA yang tersusun dari 4 huruf (A,C,G,T), sehingga jika n=3 makan akan ada 4^3 yaitu 64 *substring*, sedangkan jika n=5 maka akan ada 4^4 yaitu 256 *substring*. Pada setiap data dicari pula nilai rata-rata dan standar deviasinya serta dimasukkan pada data yang ada. Selanjutnya setiap data diberikan label seseuai dengan genusnya, *Bacillus* dilabeli dengan 1 dan *Clostridium* dilabeli dengan 0.



Gambar 3. Ekstraksi fitur dengan menggunakan n-*mers*

Setelah dilakukan proses ekstraksi fitur, perhitungan rata-rata dan standar deviasi, selanjutnya dilakukan normalisasi dengan menggunakan normalisasi minmax, sehingga diperoleh nilai diatara 0 sampai 1 sesuai *range* yang diinginkan.



Gambar 4. Contoh hasil normalisasi data

Data yang sudah dinormalisasi selanjutnya dibagi menjadi data training dan data testing menggunakan k-*fold cross validation* dalam penelitian k=5. 5-*fold cross validation* artinya ada 5 kelompok data jika 4 kelompok data sebagai data *training*, maka 1 kelompok sisanya sebagai data *testing*.

2.2 Hasil Penelitian

Hasil penelitian dapat dilihat pada Tabel 1.

Tabel 1. Hasil Penelitian

	3-mers	*4-mers*
Precision	0,90	0,8182
Recall	1	1
Specifivity	0,9091	0.8182
G-mean	0,95	0,90
akurasi	0,95	0,90

Dari Tabel 1 dapat dilihat bahwa IDELM dapat pengklasifikasikan nilai *precision* dari ekstraksi fitur dengan 3-mers dan 4-mers masing masing hasilnya adalah 0,9 dan 0,8182 artinya tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan oleh sistem yang menggunakan metode klasifikasi IDELM baik.

Hasil nilai *recall* pada 3-mers maupun 4-mers sama-sama sangat bagus yaitu 1, artinya data *bacillus* dapat dikenali semua oleh sistem. Sedangkan data *Clostridium* dapat dikenali sebesar 0,9091 atau 90,91% untuk *3-mers*, dan dikenali sebesar 0,8182 atau 81,82% untuk *4-mers.*

Geometric mean telah digunakan beberapa peneliti untuk mengevaluasi *classifier* pada dataset yang *imbalanced*. Sedangkan, akurasi biasa digunakan untuk evaluasi data biasa. Jika dilihat dari Tabel 1. Nilai akurasi dan G-mean bernilai sama yaitu 0,95 atau 95% untuk *3-mers* dan 0,90 atau 90% untuk *4-mers*

3. KESIMPULAN

Berdasarkan uraian yang telah disampaikan, klasifikasi metagenom DNA dengan ekstraksi fitur *n-mers*, serta Integrasi Seleksi Data dan *Extreme Learning Machine* sebagai *classifier* didapatkan hasil yang baik, jika dibandingkan *3-mers* dan *4-mers* performansinya lebih baik jika menggunakan ektraksi fitur *3-mers*

PUSTAKA

Haryanto, T, Bagus Guritno, H, Kustiyo, A dan Hermadi, I. 2018. *Optimasi Parameter pada Fast Correlation Based Fiter Menggunakan Algoritme Genetika untuk Klasifikasi Metagenom*. Vol 4. Jurnal Edukasi dan Penelitian Informatika (JEPIN).

Huang, G. et al., 2015. *Trends in extreme learning machines: A review*. Neural Networks, 61:32-48.

Huang, G., Zhu, Q. dan Siew, C., (2006a), "*Extreme Learning Machine*: Theory and Applications", *Neurocomputing*, Vol. 70, 489–501

Mahdiyah, Umi, Imah, E. M., Irawan, M. I.. 2017. *Integrating Data Selection And Extreme Learning Machine To Predict Protein-Ligand Binding Site*, Contemporary Enggineering Science, vol. 9.

Rapley, R & Whitehouse, D. 2015. *Molecular Biology & Biotechnology*. 6th edn. London: The Royal Society of Chemistry.

Richter, DC, Ott, F, Auch, AF, Schmid, R, dan Huson, DH. 2008. *MetaSim—A Sequencing Simulator for Genomics and Metagenomics*. PLoS ONE 3(10): e3373.

Saefudin. 2007. *Genetika*. Bandung (ID): Universitas Pendidikan Indonesia.

Sokolova, M. dan Lapalme, G.. 2009. *A systematic analysis of performance measures for classification tasks. Inf. Process. Manag.*, vol. 45, no. 4, hal. 427–437.