DISERTASI EE-186601

ANALISA TREN DAN PREDIKSI POLA PERUBAHAN HEPATITIS C VIRUS (HCV) PADA *ISOLATED* DNA BERBASIS *HYBRID CLUSTERING*

BERLIAN AL KINDHI

07111460012003

DOSEN PEMBIMBING

Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng.

Dr. Tri Arief Sardjono,S.T.,M.T.

PROGRAM DOKTOR

DEPARTEMEN TEKNIK ELEKTRO

FAKULTAS TEKNOLOGI ELEKTRO

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

SURABAYA

2019

LEMBAR PENGESAHAN
PREDIKSI POLA PERUBAHAN GENETIK PADA *SEQUENCE* DNA HEPATITIS C VIRUS (HCV) BERBASIS MUTASI CENTROID CLUSTER

Disertasi ini disusun untuk memenuhi salah satu syarat memperoleh gelar Doktor
di Institut Teknologi Sepuluh Nopember Surabaya

Oleh :
Berlian Al Kindhi (NRP. 07111460012003)
Tanggal Ujian : Periode Wisuda :

Prof. Dr. Ir. Mauridhi H. Purnomo, M.Eng. NIP. 19580916 198601 1 001	.....	(Pembimbing I)
Dr. Tri Arief Sardjono, S.T., M.T. NIP. 19700212 199512 1 001	.....	(Pembimbing II)
Dr. I Ketut Eddy Purnama, S.T., M.T. NIP. 19690730 199512 1 001	.....	(Penguji I)
Dr. Surya Sumpeno, S.T.,M.Sc. NIP. 19690613 199702 1 003	.....	(Penguji II)
Izzati Muhimah, S.T.,M.Sc.,Ph.D. NIP. 985240102	.....	(Penguji III)

Mengetahui,
Dekan Fakultas Teknologi Elektro

Dr. Tri Arief Sardjono, S.T., M.T.
NIP. 19700212 199512 1 001

i

Halaman ini sengaja dikosongkan

SURAT PERNYATAAN KEASLIAN DISERTASI

Yang bertandatangan di bawah ini:

Nama : Berlian Al Kindhi

Program Studi : Teknik Elektro

NRP : 07111460012003

dengan ini menyatakan bahwa isi sebagian maupun keseluruhan disertasi dengan judul:

ANALISA TREN DAN PREDIKSI POLA PERUBAHAN HEPATITIS C VIRUS (HCV) PADA *ISOLATED* DNA BERBASIS *HYBRID CLUSTERING*

adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diizinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri. Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya, 20 Januari 2019
Yang Membuat Pernyataan,

Berlian Al Kindhi
NRP. 07111460012003

Halaman ini sengaja dikosongkan

ANALISA TREN DAN PREDIKSI POLA PERUBAHAN HEPATITIS C VIRUS (HCV) PADA *ISOLATED* DNA BERBASIS *HYBRID CLUSTERING*

Nama	:	Berlian Al Kindhi
NRP	:	07111460012003
Promotor	:	Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng.
Co-Promotor	:	Dr. Tri Arief Sardjono,S.T., M.T.

ABSTRAK

Hepatitis C Virus (HCV) merupakan salah satu jenis penyakit yang peluang penularannya mayoritas di daerah tropis (penyakit tropis). Saat ini belum ada vaksin yang secara mutlak dapat digunakan untuk mencegah Hepatitis C karena virus ini secara genetik amat variatif (*subtype genome*) dan memiliki angka mutasi tinggi, sehingga memungkinkan generasi virus yang beraneka ragam. Menurut WHO, angka kematian akibat infeksi HCV cukup tinggi, yaitu mencapai 399 ribu jiwa per tahun. Indonesia merupakan salah satu negara yang memiliki jumlah pasien terinfeksi HCV tertinggi di Asia. Penyakit ini sebagian besar menjangkit di daerah tropis namun tidak menutup kemungkinan terdapat *carier agent* yang mampu menularkan penyakit hingga ke berbagai benua.

HCV adalah jenis virus RNA. Ketika terinfeksi virus, RNA akan berusaha beradaptasi dengan mengubah pola kode RNA sehingga DNA yang terbentuk dapat bertahan hidup. Akibatnya DNA yang terbentuk juga akan berubah, sehingga terus menghasilkan *subtype* baru pada HCV dan tidak semua primer mampu mengenali adanya HCV di dalam *isolated* DNA. Oleh karena itu dibutuhkan suatu metode yang mampu menganalisa primer yang menjadi tren. Serta dibutuhkan metode yang mampu memprediksi adanya HCV dengan pola mutasinya yang beragam.

Pada disertasi ini, diusulkan sebuah metode baru yaitu *Hybrid Clustering* yang mampu memberikan tiga analisa sekaligus *similarity, trend,* dan, *hierarchical*. Analisa tersebut adalah kecenderungan *isolated* DNA terhadap suatu primer, analisa tren primer HCV, dan runutan infeksi HCV terhadap *isolated* DNA. Selain itu, SVM juga dibutuhkan untuk memprediksi adanya HCV pada *isolated* DNA. SVM mampu memprediksi adanya/tidak adanya HCV pada 1000 *isolated* DNA yang diujikan, *isolated* DNA tersebut berasal dari berbagai negara di dunia. Namun sebelum melakukan proses *clustering* dan prediksi, perlu dilakukan normalisasi dengan *semantic similarity.* Hasil analisa *clustering* dan prediksi tersebut diharapkan dapat digunakan sebagai evaluasi di bidang kedokteran untuk selangkah lebih dekat dalam penemuan vaksin HCV.

Kata kunci : ***approximate string matching, hybrid clustering, SVM, DNA HCV***.

Halaman ini sengaja dikosongkan

TREND ANALYSIS AND PREDICTION OF HEPATITIS C VIRUS (HCV) CHANGE PATTERNS IN ISOLATED DNA BASED ON HYBRID CLUSTERING

Name	: Berlian Al Kindhi
Nrp	: 07111460012003
Promotor	: Prof. Dr.Ir. Mauridhi Hery Purnomo, M.Eng
Co-promotor	: Dr. Tri Arief Sardjono, S.T.,M.T.

ABSTRACT

Hepatitis C Virus (HCV) is one type of disease that has the highest chance of transmission in the tropics (tropical diseases). At present there is no vaccine that can absolutely be used to prevent Hepatitis C because this virus is genetically very varied (genome subtype) and has high mutation rates, thus enabling the generation of diverse viruses. According to WHO, the mortality rate due to HCV infection is quite high, reaching 399 thousand people per year. Indonesia is one of the countries with the highest number of patients infected with HCV in Asia. The disease is mostly infectious in the tropics but does not rule out the possibility of a carrier agent that is capable of transmitting the disease to various continents.

HCV is a type of RNA virus. When infected with a virus, RNA will try to adapt by changing the pattern of RNA codes so that the DNA formed can survive. As a result the DNA formed will also change. So that it continues to produce new subtypes in HCV and not all primers are able to recognize the presence of HCV in isolated DNA. Therefore a method that is able to analyze the primary which is a trend and a method that can predict the presence of HCV with a variety of mutation patterns is needed.

In this dissertation, a new method is proposed, namely Hybrid Clustering that is able to provide three analyzes as well as similarity, trend, and hierarchical. The analysis is the tendency of isolated DNA towards a primer, analysis of HCV primary trends, and trace of HCV infection to isolated DNA. In addition, SVM is also needed to predict the presence of HCV in isolated DNA. SVM was able to predict the presence or absence of HCV in 1000 isolated DNA tested, these isolated DNA originated from various countries in the world. But before doing the clustering and prediction process, semantic similarity needs to be normalized. The results of the clustering and prediction analysis are expected to be used as evaluations in the medical field to be one step closer to the discovery of the HCV vaccine.

Keywords : approximate string matching, hybrid clustering, SVM, DNA HCV

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Puji syukur Alhamdulillah senantiasa penulis panjatkan kehadirat Allah SWT yang telah memberikan Rahmat, karunia, berkah dan hidayah-Nya, dan shalawat serta salam senantiasa tercurahkan kepada junjungan kita Rasulullah SAW, hingga terselesaikannya penulisan disertasi yang berjudul "Analisa Tren dan Prediksi Pola Perubahan Hepatitis C Virus (HCV) pada *Isolated* DNA berbasis *Hybrid Clustering*". Disertasi ini disusun untuk memenuhi salah satu syarat akademik Program Doktor Departemen Teknik Elektro, Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember Surabaya.

Banyak tantangan dan hambatan yang penulis hadapi dalam penulisan disertasi ini. Alhamdulillah atas pertolongan Allah SWT dan bantuan dari berbagai pihak akhirnya penulisan ini dapat Penulis selesaikan. Pada kesempatan ini, penulis menyampaikan terimakasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng. selaku pembimbing dan Bapak. Dr. Tri Arief Sardjono,ST.,MT. selaku co-pembimbing yang telah dengan sabar memberikan arahan, pembimbingan dan semangat dalam menempuh studi Program Doktor di Teknik Elektro ITS Surabaya. Beliau-beliau telah membimbing penulis hingga dapat mempublikasikan hasil penelitian penulis di Seminar Internasional dan Jurnal Internasional terindeks Scopus dan memberikan berbagai solusi yang sangat bermanfaat dalam pengembangan hasil penelitian dan amanat Tri Dharma Perguruan Tinggi.
2. Dr. Marco Alexander Wiering selaku pembimbing di Universitas Groningen Belanda, terima kasih untuk semua diskusi, *sharing*, pengalaman, tempaan, bantuan, dan pengajaran dalam membuat paper yang baik dan benar. Terima kasih sudah mengajak saya dalam pertemuan peneliti *AI* Eropa, mendaftarkan saya dalam program *NVIDIA deep learning*, mengikutkan saya dalam *machine*

learning course, serta pelajaran berharga lainnya yang tidak dapat saya sebutkan satu-persatu.

3. Prof. (Bart) G.J. Verkerke, selaku pembimbing selama di Universitas Groningen. Terima kasih telah membantu memeriksa paper dan semua diskusi menyenangkan baik di Belanda maupun di Indonesia.
4. Bapak M. Amin selaku peneliti di Laboratorium Hepatitis, Institut Tropical Disease (ITD), Universitas Airlangga, atas bantuan dan bimbingannya selama ini.
5. Dr. I Ketut Eddy Purnama,ST.,MT., Dr. Surya Sumpeno,ST.,M.Sc., dan Izzati Muhimah,ST.,M.Sc.,Ph.D. selaku penguji terima kasih atas saran dan masukannya.
6. Seluruh dosen-dosen Fakultas Teknologi Elektro, Institut Teknologi Sepuluh Nopember yang tidak dapat penulis sebutkan satu persatu.
7. Seluruh dosen-dosen Prodi Teknik Informatika, UNTAG Surabaya yang tidak dapat penulis sebutkan satu persatu.
8. Seluruh dosen-dosen Prodi KP dan LB, Akademi Teknik Keselamatan Penerbangan - KEMENHUB yang tidak dapat penulis sebutkan satu persatu.
9. Seluruh dosen-dosen Dept. Teknik Elektro Otomasi, Institut Teknologi Sepuluh Nopember yang tidak dapat penulis sebutkan satu persatu.
10. Seluruh rekan-rekan mahasiswa S3 B 401, 2015: Bu Rima, Bu Irma, Bu Evi, Pak Yuli, dan Pak Fanani, terima kasih untuk kebersamaan dan diskusi yang luar biasa.
11. Seluruh rekan-rekan mahasiswa B 204, Bu Ros, Bu Rosi, Bu Yuana, Bu Diana, Bu Peni, Bu Nova, Bu Rikha, Bu Tita, Bu Yuni, Mas Wahyu, Pak Cucun, Pak Andi, Pak Imam, Pak Aryo, Pak Adri, Pak Alam, serta teman-teman lainnya yang tidak dapat saya sebutkan satu-persatu, terima kasih untuk diskusi, kekompakkan, dan saling supportnya.

12. Rekan-rekan di Ruang 305 Bernoulliborg, Pry, Sheng, Run, dan Maruf untuk dukungan, kesigapan bantuan, kebersamaan, dan semuanya hingga saat ini.
13. Seluruh rekan-rekan mahasiswa Ph.D. Departemen Artificial Intelligence di Universitas Groningen, Mixue, Yuri, Oscar, Stefan, Ega, Christina, dan Emanuel untuk diskusi yang menyenangkan, gurauan, dan motivasinya.
14. *My support system*, Masca Indra, Aozora Janeeta Masca, dan Aldric Anindyapraja Masca untuk waktu, tenaga, dukungan dan kasih sayangnya.
15. Dan seluruh teman-teman, kenalan dan saudara-saudara yang tidak dapat Penulis sebutkan satu persatu, atas doa dan dukungannya selama ini.

Penulis mengucapkan banyak terima kasih telah membantu selama melakukan kegiatan riset dan studi S3 ini. Semoga Allah SWT memberi balasan dengan pahala kebaikan yang sempurna. Demikian Laporan Disertasi ini disusun, segala masukan dan koreksi sangat diharapkan untuk penyempurnaan ke depan. Semoga bermanfaat.

Surabaya, 20 Januari 2019
Penulis

Halaman ini sengaja dikosongkan

DAFTAR ISI

Halaman ini sengaja dikosongkan

DAFTAR GAMBAR

DAFTAR TABEL

Halaman ini sengaja dikosongkan

DAFTAR ISTILAH

DNA : *DeoxyriboNucleic Acid*, adalah biomolekul yang menyimpan dan menyandi instruksi-instruksi genetika setiap organisme dan banyak jenis virus.

FASTA : adalah *file* hasil pengubahan *isolated* DNA yang mengambil urutan nukleotidanya saja. Dalam satu nukleotida diwakilkan dalam satu karakter huruf besar.

Hepatitis C : penyakit menular yang mempengaruhi hati, yang disebabkan oleh virus hepatitis C (HCV)

***Isolated* DNA** **:** *File* hasil ekstrasi dari DNA yang berisi origin (negara asal DNA tersebut), tahun, dan urutan nukleotida yang dipisahkan spasi setiap satu untingnya.

Molecular clock : teknik yang digunakan untuk mengukur mutasi biomolekul berdasarkan catatan sejarah, waktu, maupun fosil

Mutasi : perubahan yang terjadi pada bahan genetic (DNA maupun RNA), baik pada taraf urutan gen (disebut mutase titik) maupun pada taraf kromosom.

Nukleotida : molekul yang tersusun dari gugus basa heterosiklik, gula, dan satu atau lebih gugus fosfat

Penyakit tropis : penyakit yang hanya terjadi pada daerah tropis dan atau ekivalensi dengan peluang kemunculannya lebih besar terjadi pada daerah tropis

Primer : adalah urutan nukleotida yang terdiri dari 20 hingga 50 nukleotida yang berfungsi sebagai penentu apakah *isolated* DNA tersebut terinfeksi penyakit dari primer yang diujikan

RNA : *RiboNucleic Acid*, adalah satu dari tiga makromolekul utama (bersama dengan DNA dan protein) yang berperan penting dalam segala bentuk kehidupan.

Rantai heliks : Suatu rangka bergulung membentuk bagian dalam tangkai dan rantai sisi meluas keluar seperti suatu pilinan DNA

***Sequence* DNA** : urutan basa nukleotida pada suatu molekul DNA yang merupakan informasi paling mendasar suatu gen atau genom karena mengandung instruksi yang dibutuhkan untuk pembentukan tubuh makhluk hidup

Strand : Untaian DNA yang menyimpan protein dan basa

Vaksin : bahan antigenik yang digunakan untuk menghasilkan kekebalan aktif terhadap suatu penyakit yang disebabkan oleh bakteri atau virus, sehingga dapat mencegah atau mengurangi pengaruh infeksi oleh organisme alami

BAB 1
LATAR BELAKANG

1.1 Pendahuluan

Penyakit Tropis adalah penyakit yang hanya terjadi pada daerah tropis dan atau ekivalensi dengan peluang kemunculannya lebih besar terjadi pada daerah tropis (Airlangga U. , 2012). Bakteri pembawa penyakit tersebut mencakup agen infeksi yang *multi resistance* dan atau *transibility* (mudah menular). Semakin lama bakteri bermutasi dan berkembang biak semakin banyak. Pola resistensi bakteri terhadap suatu antibiotik atau zat kimia resisten lainnya pun mulai mengalami pergeseran kekebalan. Penyebaran infeksi yang semakin cepat diikuti dengan mutasi kekebalan yang cepat pula dapat menimbulkan dampak buruk bagi kelangsungan hidup manusia. Penyakit tropis merupakan salah satu epidemi yang paling merusak dan ancaman utama bagi penduduk dunia, yang mempengaruhi kesejahteraan sosial, ekonomi dan politik secara keseluruhan serta kesehatan individu (O.A. Akalu, A. Endale, N. Tesfaye, D. Woldemichael, 2010)

Hepatitis C Virus termasuk salah satu penyakit tropis yang mengalami siklus mutasi genetik (*molecular clock*) cepat dan mudah menular. Menurut WHO, jumlah penderita hepatitis C di dunia diperkirakan mencapai 71 juta jiwa per-tahunnya dan menyebabkan kematian pada sekitar 399 ribu penderitanya (WHO, 2017). Sementara di Asia Tenggara sendiri, jumlah penderita yang meninggal akibat komplikasi kanker hati (*sirosis*) akibat hepatitis C diperkirakan mencapai 90.000 jiwa tiap tahunnya. Indonesia merupakan salah satu negara dengan tingkat kasus hepatitis C tertinggi di Asia Tenggara.

Hepatitis C umumnya tidak menunjukkan gejala pada tahap-tahap awal. Karena itu, sekitar 75 persen penderita hepatitis C tidak menyadari bahwa dirinya sudah tertular sampai akhirnya mengalami kerusakan hati bertahun-tahun kemudian. Meski

ada gejala hepatitis C yang muncul, indikasinya mirip dengan penyakit lain sehingga sulit disadari. Hepatitis yang berlangsung kurang dari 6 bulan disebut “hepatitis akut”, hepatitis yang berlangsung lebih dari 6 bulan disebut “hepatitis kronis”.

Virus Hepatitis C adalah jenis virus RNA. RNA merupakan *molecular* dalam tubuh yang bertanggung jawab sebagai pembawa pesan kode untuk pembentukan protein DNA baru. Jika RNA terinfeksi virus, akibatnya DNA yang terbentuk juga akan berubah. Sifat alami sel adalah mempertahankan diri supaya tetap hidup dengan mengubah pola kode DNA yang baru melalui perubahan pola RNA. Probabilitas perubahan yang terjadi DNA juga dapat berubah-ubah tergantung adaptasi alami atau pengobatan yang dilakukan saat itu. Sebagai contoh virus sudah resisten dengan obat yang sama karena pola kode sudah berubah lagi (sebagai adaptasi bertahan hidup), membentuk *subtype genome* yang baru lagi dan seterusnya.

Pola penyebaran dan perkembang-biakan HCV yang semakin cepat inilah yang menjadi permasalahan dunia kedokteran. Hingga saat ini, belum ditemukan anti-HCV atau vaksin untuk HCV. Penelitian baik dengan pendekatan biologi kedokteran (Juniastuti et al., 2014) maupun teknik biomedika mengarah ke tujuan utama dari penelitian pada sub bidang HCV ini yaitu penemuan vaksin. Namun hingga saat ini, hasil capaian penelitian tersebut, khususnya dengan pendekatan teknik biomedika baru pada tahap prediksi tren mutasi pada DNA-nya (Bin Liu; Shanyi Wang; Qiwen Dong; Shumin Li; Xuan Liu, 2016). Pada penelitian ini diusulkan metode baru pada tahap analisa pengelompokan dan metode SVM dengan perumusan *matriks dataset* yang baru untuk menuju tahap penemuan vaksin.

Database Genomik adalah *database* yang terdiri dari data biologis DNA suatu individu (Cronkite, 2002). Data biologis yang terdapat dalam *database* genomik berbentuk rangkaian DNA, rangkaian RNA, dan rangkaian protein. *Sequence* DNA adalah urutan rangkaian basa purin dan pirimidin yang berkaitan satu sama lain menyusun informasi genetika. Melalui *sequence* tersebut, dunia kedokteran mampu

melakukan analisa *strand* DNA dengan tujuan untuk memperoleh sejarah, fungsi, peranan biokimia, struktur kimia, dan mutasi suatu genomik.

Metode *semantic similarity* dapat digunakan sebagai metode dalam meneliti sejarah, dan mutasi suatu genomik. Setiap *sequence* DNA memiliki informasi genetik tertentu yang dapat digali melalui uji kecocokan *sequence* dengan suatu *sequence* tertentu (Juniastuti et al., 2014). Hasilnya dapat berupa analisa kecocokan suatu *isolated* DNA dengan primer (rumusan *sequence* DNA suatu penyakit). Melalui analisa kecocokan struktur DNA tersebut dapat dianalisa apakah *isolated* DNA tersebut terinfeksi suatu penyakit atau tidak.

Metode *clustering* atau klasifikasi mampu mengelompokkan suatu kumpulan data sesuai dengan nilai kedekatannya. Hasil dari pencocokan *sequence* DNA dapat diolah untuk mengetahui tren mutasi genomik dengan cara melakukan *clustering*. Tren mutasi genomik adalah analisa kecenderungan perubahan genetik suatu virus atau penyakit. Pada penelitian pendahulu, metode *clustering* digunakan untuk mengelompokkan *sequence* DNA sesuai dengan nilai kedekatannya terhadap primer. Sehingga dapat dianalisa arah perubahan genetik suatu virus atau penyakit (S. Tapan and D. Wang, 2016).

Terdapat beberapa metode yang mampu mengakomodasi prediksi perubahan data tak terstruktur, (Marco Capó, Aritz Pérez, Jose A. Lozano, 2017), menganalisa metode K-Means sebagai prediksi DNA. Beberapa penelitian pendahulu menggunakan *Support Vector Machine* untuk mengklasifikasi data DNA berdasarkan kedekatannya dengan primer, hasilnya adalah prediksi kecenderungan perubahan genetiknya (Srinivasareddy Putluri, Md Zia Ur Rahman, Shaik Yasmeen Fathima, 2018) (Neelam Goel, Shailendra Singh, Trilok Chand Aseri, 2015)**.**

Pada disertasi ini diusulkan sebuah metode baru dalam *clustering* yang mampu memberikan tiga analisis dalam satu proses, yaitu analisa kecenderungan suatu *isolated* DNA terhadap suatu primer, tren mutasi primer yang di tandai dengan banyaknya jumlah anggota yang tergabung dalam kelompok *centroid* primer tersebut, serta yang

terakhir adalah untuk menelusuri urutan-urutan nukelotida yang mana sajakah yang positif terhadap suatu primer dan primer mana sajakah yang mampu mengenali adanya HCV pada *isolated* DNA tersebut. Pada penelitian ini diusulkan suatu metode baru pada tahap pengelompokan dan prediksi dengan performansi akurasi mencapai 95% untuk pengelompokan dan 99% untuk prediksi.

Pada disertasi ini data sampel yang digunakan adalah 1000 data *isolated* DNA yang terdiri dari 500 *isolated* DNA homo sapiens yang positif terinfeksi HCV dan 500 *isolated* DNA homo sapiens negatif HCV. 1000 data tersebut dihitung jarak kemiripannya menggunakan metode *Edit Levensthein Distance*. Hasil dari penghitungan *Edit Levensthein Distance* kemudian dimasukkan ke dalam matriks sebagai variabel. Matriks tersebut adalah input data pada proses prediksi menggunakan SVM. Dari hasil penelitian ini, diharapkan mampu memberikan analisa perubahan genetik DNA khususnya pada DNA yang terinfeksi HCV dan hasilnya dapat dimanfaatkan oleh dunia kedokteran sebagai evaluasi.

1.2 Perumusan Masalah

Semakin beragamnya pola perubahan suatu virus, merupakan salah satu kendala dalam mencari suatu pola pada jutaan urutan DNA yang ada. Oleh karena itu dibutuhkan suatu metode yang tepat baik dari sisi ketepatan (efektifias) dan kecepatan metode untuk menemukan pola tersebut (efisiensi).

Dunia kedokteran membutuhkan analisa trend perubahan genetik infeksi HCV dari data-data *isolated* DNA sebagai dasar evaluasi. *Clustering* genetik merupakan pengelompokan data dalam jumlah besar, sehingga analisa pengelompokan secara manual tentu akan membutuhkan waktu lama dan kemungkinan terjadinya *human error* cukup tinggi.

Data *isolated* DNA yang besar namun tidak di imbangi dengan infrastruktur yang baik tidak akan dapat memberikan hasil pembelajaran mesin yang baik pula. Oleh

karena itu dibutuhkan suatu rancangan infrastruktur sistem pakar analisis DNA yang mampu membantu tenaga ahli dalam melakukan penelitian.

Semakin beragamnya pola mutasi suatu virus, akan menghasilkan urutan DNA yang berbeda-beda pula. Sehingga tidak semua primer mampu mengenali adanya HCV di dalam suatu *isolated* DNA. Oleh karena itu dibutuhkan metode pembelajaran mesin yang mampu mempelajari pola mutasi tersebut sehingga mampu mengenali adanya HCV pada *isolated* DNA yang berasal dari berbagai negara di dunia.

1.3 Tujuan Penelitian

Penelitian ini dilakukan dengan tujuan untuk:

1. Menemukan pola suatu virus di dalam *isolated* DNA,
2. Melakukan Pengelompokan DNA untuk memberikan analisa kencenderungan *isolated* DNA terhadap primer tertentu,
3. Menganalisa tren mutasi suatu virus,
4. Menelusuri primer-primer yang mampu mengenali adanya HCV di dalam *isolated* DNA tersebut untuk mendapat keterkaitan virus *isolated* DNA tersebut dengan primer tertentu
5. Mengusulkan rancangan infrastruktur sistem pakar DNA analisis yang mampu menjawab kebutuhan penelitia dewasa ini,
6. Memprediksi adanya HCV di dalam *isolated* DNA melalui pola perubahan genetiknya

1.4 Manfaat Penelitian

Manfaat dari penelitian ini yaitu didapatnya analisa metode yang sesuai untuk “semantic similarity” dengan data set DNA yang bermutasi, didapatnya suatu metode pengelompokan yang mampu membantu ahli dalam melakukan proses analisa tren dan kecenderungan *isolated* DNA terhadap suatu primer, khususnya primer HCV, di dapatnya kernel SVM yang paling sesuai untuk memprediksi dan adanya HCV di

dalam *isolated* DNA, didapatnya suatu rancangan infrastruktur system pakar analisis DNA yang terintegrasi antar rumah sakit dan pemerintah yang berjalan di *cloud*.

1.5 Kontribusi dan Originalitas Penelitian

1.5.1 Kontribusi Penelitian

Hasil akhir penelitian ini memberikan terobosan baru pada dunia teknologi biomedika dengan mengusulkan *hybrid clustering* yang menggabungkan kelebihan dari tiga metode *clustering* sebagai sistem analis DNA yang mampu menghasilkan tiga analisa sekaligus dalam satu kali proses *clustering*. Hasil dari analisa tersebut diharapkan dapat dimanfaatkan oleh dunia kedokteran sebagai dukungan evaluasi dan penelitian DNA khususnya pada studi kasus HCV.

1.5.2 Originalitas Penelitian

Sistem yang dibangun pada penelitian ini mampu melakukan pengelompokan tren HCV dengan tiga pendekatan yang berbeda yang dapat memenuhi kebutuhan ahli dalam menganalisa DNA HCV serta melakukan prediksi adanya HCV dalam *isolated* DNA dari berbagai negara. Selain itu, penilitian ini mengusulkan sebuah rancangan sistem pakar analisis DNA yang terintegrasi antar rumah sakit dengan pemerintah turut serta memfasilitasi dan mengawasi penggunaan sistem pakar tersebut.

1.6 *Road Map* Penelitian

Pada sub bab 1.6 akan disampaikan *miles stone* sejauh mana pencapaian penelitian lain sebidang dan posisi kontribusi dari penelitian ini.

1.6.1 Penelitian Sebelumnya

Penelitian pada bidang teknik biomedika, khususnya pada data DNA sudah banyak dilakukan sebelumnya, Table 1.1. adalah beberapa studi literatur penelitian sebelumnya yang dikelompokkan sesuai dengan fokus penelitian dan *road map* dari disertasi ini.

Tabel 1.1. *State of the art* penelitian bidang teknik biomedika

Topik	No	Judul	Hasil
DNA HCV	1	Sandra Iurecia et. al., "Epitope-driven DNA vaccine design employing immunoinformatics against B-cell lymphoma: A biotech's challenge" ,2011	Menghasilkan model vaksin kanker berbasis epitope DNA yang dapat memungkinkan membangun plasmid dari beberapa epitope imunogenetik
	2	Hemiyanti Emmy, "Biologi Molekul Virus", Pasca Sarjana Universitas Padjajaran, 2012	Penjelasan mengenai pola mutasi dari molecular virus, tempat hidupnya dan pola perkembang biakannya.
	3	Grey Rebecca R., et al., "Evolutionary analysis of hepatitis C virus , 2013	Menganalisis dua urutan HCV subgenomic diperoleh dari individu yang terinfeksi di 1953, yang merupakan bukti genetik tertua infeksi HCV. Metodenya adalah dengan memasangkan keragaman genetik antara dua sekuens sehingga menunjukkan substansial periode penularan HCV sebelum tahun 1950-an, dan masuknya virus tersebut dalam evolusi analisis memberikan perkiraan baru dari nenek moyang HCV di Amerika Serikat. Memperkirakan bahwa saat awal mula munculnya HCV subtipe 1b di Amerika Serikat terjadi sekitar tahun 1901 (1874-1926), yang berarti perkiraan ini konsisten dengan perkiraan sebelumnya. Namun, analisis ini memberikan hasil CI yang tinggi daripada yang dilaporkan sebelumnya untuk subtipe 1b yang menggunakan dua wilayah subgenomik (1905-1965 dan 1806-1959;). Selain itu hasil penelitian ini mencerminkan informasi meningkat diperoleh dari menggunakan seluruh genom urutan referensi dan dari masuknya dua urutan primer yaitu pada tahun 1953 .
	4	Takayakagi Toshiaki, "Modeling chronic hepatitis B or C virus infection during antiviral therapy using an analogy to enzyme kinetics: Long-term viral	Model dasar untuk virus hepatitis B kronis (HBV) atau virus Hepatitis C (HCV) selama terapi memungkinkan kita untuk menganalisis kinetika virus jangka pendek. Namun, model ini tidak berguna untuk menganalisis jangka panjang

Topik	No	Judul	Hasil
		dynamics without rebound and oscillation"(2013)	kinetika virus. Oeh karena itu, pada penelitian ini diusulkan model baru yang diperoleh dengan memperkenalkan Michaelis-Menten kinetika ke dalam model dasar. Model baru dapat menunjukkan kinetika virus jangka panjang tanpa *Rebound* dan osilasi, tidak seperti model dasar. Nilai parameter K dalam model baru analog dengan Michaelis adalah konstan dan diprediksi menjadi kurang dari sekitar 1.010 / ml.
Infrastruktur Sistem Pakar DNA Analisis	1	Shabut et.al., "An intelligent mobile-enabledd expert system for tuberculosis disease in real time" (2018)	Suatu *expert system* untuk mendiagnosa penyakit tuberculosis dengan melakukan analisa gejala-gejala secara langsung berbasis aplikasi mobile.
DNA Semantic Similarity	1	Fredonnet Julie, "Dynamic PDMS inking for DNA patterning by soft lithography"(2013)	Pencetakan microcontact (LCP) digunakan sebagai teknik pola untuk menghasilkan DNA microarray sederhana, cepat dan biaya-efektif.
	2	Mika Göös,et al., "Search methods for tile sets in patterned DNA self-assembly"(2014)	Pattern self Assembly Tile set Synthesis (PATS), yang muncul dalam teori terstruktur DNA self-assembly, adalah untuk menentukan satu set *coloured tiles*, mulai dari struktur benih berbatasan, hingga merakit diri untuk pola warna persegi panjang yang diberikan. Tugas mencari minimum ukuran tile set dikenal NP-keras. Penelitian mengeksplorasi beberapa teknik pencarian yang lengkap dan tidak lengkap untuk menemukan minimal *tile set* dan juga menilai keandalan solusi yang diperoleh sesuai dengan Tile kinetik Assembly Model.
	3	Fernau Henning, et al.,, "Pattern matching with variables: A multivariate complexity analysis"(2015)	Dalam DNA *pattern matching* terdapat banyak parameter masalah antara lain: jumlah variabel, panjang w, panjang kata-kata menggantikan variabel, jumlah kejadian per variabel,

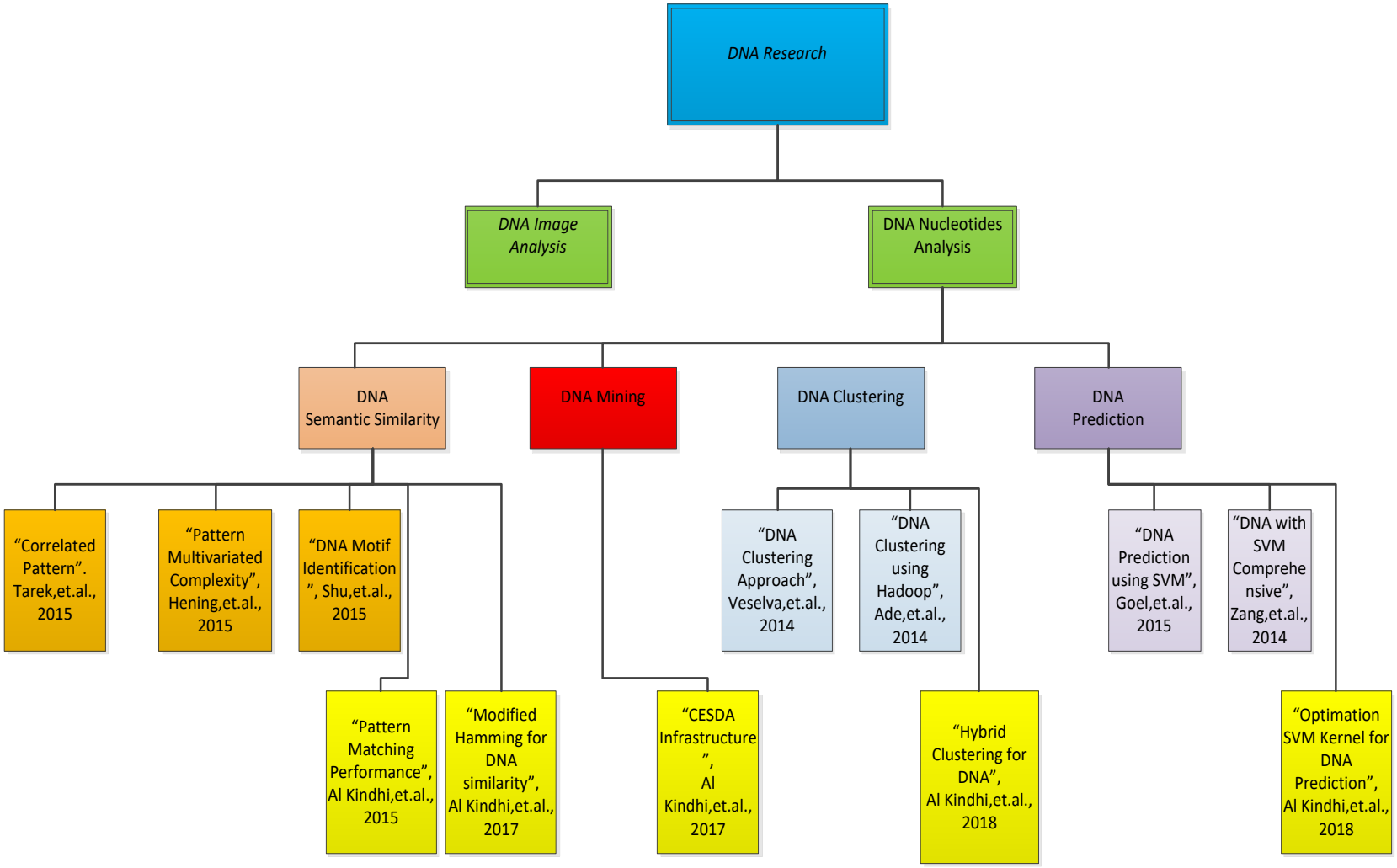
Topik	No	Judul	Hasil
			kardinalitas alfabet terminal dan untuk semua kemungkinan kombinasi dari parameter (dan varian yang dijelaskan sebelumnya), penelitian ini menjawab pertanyaan apakah ada masalah atau tidak pada NP-lengkap jika parameter ini dibatasi oleh konstanta. Hasil dari penelitian menunjukkan bahwa pemberian konstanta akan memudahkan analisis DNA namun dengan adanya konstanta juga akan menurunkan tingkat sensitivitas terhadap mutasi.
Pengelompokan DNA	1	Yilmas Kaya, Murat Uyar, "A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease"(2013)	Mengusulkan diagnosis penyakit hepatitis menggunakan metode Rough Set dan Extreme Learning Machine (RS-ELM) dalam sebuah kumpulan data diagnosa. Hasil penelitian menunjukkan bahwa model RS-ELM 100% telah cukup sukses dibandingkan dengan metode lainnya dalam literatur
	2	Boeka veselva," *Clustering* approaches for dealing with multiple DNA microarray datasets"(2014)	Menggabungkan empat algoritma *clusterring* untuk menangani *multiple gene expression* matrik pada DNA Microarray. Metode *cluestering* tersebut adalah dua *unsupervised technique* berbasis integrasi informasi dan dua *supervised technique* yaitu menggabungkan *Particle Swarm Optimization* dan *k-means.* Hasilnya Pendekatan MapReduce Clusterring melebihi tiga algoritma pengelompokan lainnya. Selain itu, versi FCA-ditingkatkan memungkinkan untuk menganalisis lebih lanjut partisi diproduksi dan untuk mengekstrak wawasan biologis yang berharga dari data.
	3	Abolfazl Doostparast Torshizi, "A new cluster validity measure based on general type-2 fuzzy sets:	Meneliti pendekatan baru di bidang General Type-2 Fuzzy Sets (GT2 FS) dan aplikasi yang dikembangkan. Pada penelitian ini telah dianalisis ukuran kesamaan berdasarkan jarak yang

Topik	No	Judul	Hasil
		Application in gene expression data *clustering*"(2014)	melebihi pendekatan yang ada dan mencakup sebagian besar kekurangan penelitian sebelumnya. Setelah pengujian pada beberapa dataset buatan dengan berbagai jumlah *outlier*, dengan menggunakan tiga gen nyata ekspresi dataset dan memverivikasi kualitas terhadap sejenis pendekatan baik secara visual dan komputasi. Percobaan ini terbukti akurasi dan presisi dari metode yang telah dikembangkan.
	4	Dios Fransisco.,et al., "DNA *clustering* and genome complexity"(2014)	Mengelompokkan DNA kompleks berdasarkan sepuluh elemen genome manusia.
	5	Jamal Ade, et al., "Scalability of DNA Sequence Database on Low-End Cluster using Hadoop(2014)	Skalabilitas data *sequence DNA* pada *world gen bank* untuk di akses dan di kelompokkan menggunakan hadoop. Data diambil dari NJBI kemudian di buat sebuah arsitektur jaringan untuk *clustering server data*
	6	Dzung Dinh Nguyen, "Towards hybrid *clustering* approach to data classification: Multiple kernels based interval-valued Fuzzy C-Means algorithms"(2015)	Kelemahan dari Fuzzy C-Means adalah pengelompokan dapat melibatkan berbagai fitur masukan menunjukkan dampak yang berbeda pada hasil yang diperoleh. Penelitian ini mengusulkan metode baru dari Fuzzy C-Means yaitu, komposit kernel dibangun dengan memetakan setiap fitur masukan ke ruang kernel individu dan linear menggabungkan kernel ini dengan bobot dioptimalkan dari kernel yang sesuai.
Prediksi DNA	1	Wang Hongfei, et.al., "Evaluation of an artificial neural network to ascertain why there is a high incidence of hepatitis B in the Chinese population after vaccination"(2013)	Menerapkan *artificial neural network* untuk menganalisa kenapa angka infeksi HBV tinggi setelah vaksin. Hasil dari neural network menunjukkan tidak ada hubungannya antara tingginya infeksi dengan vaksin.

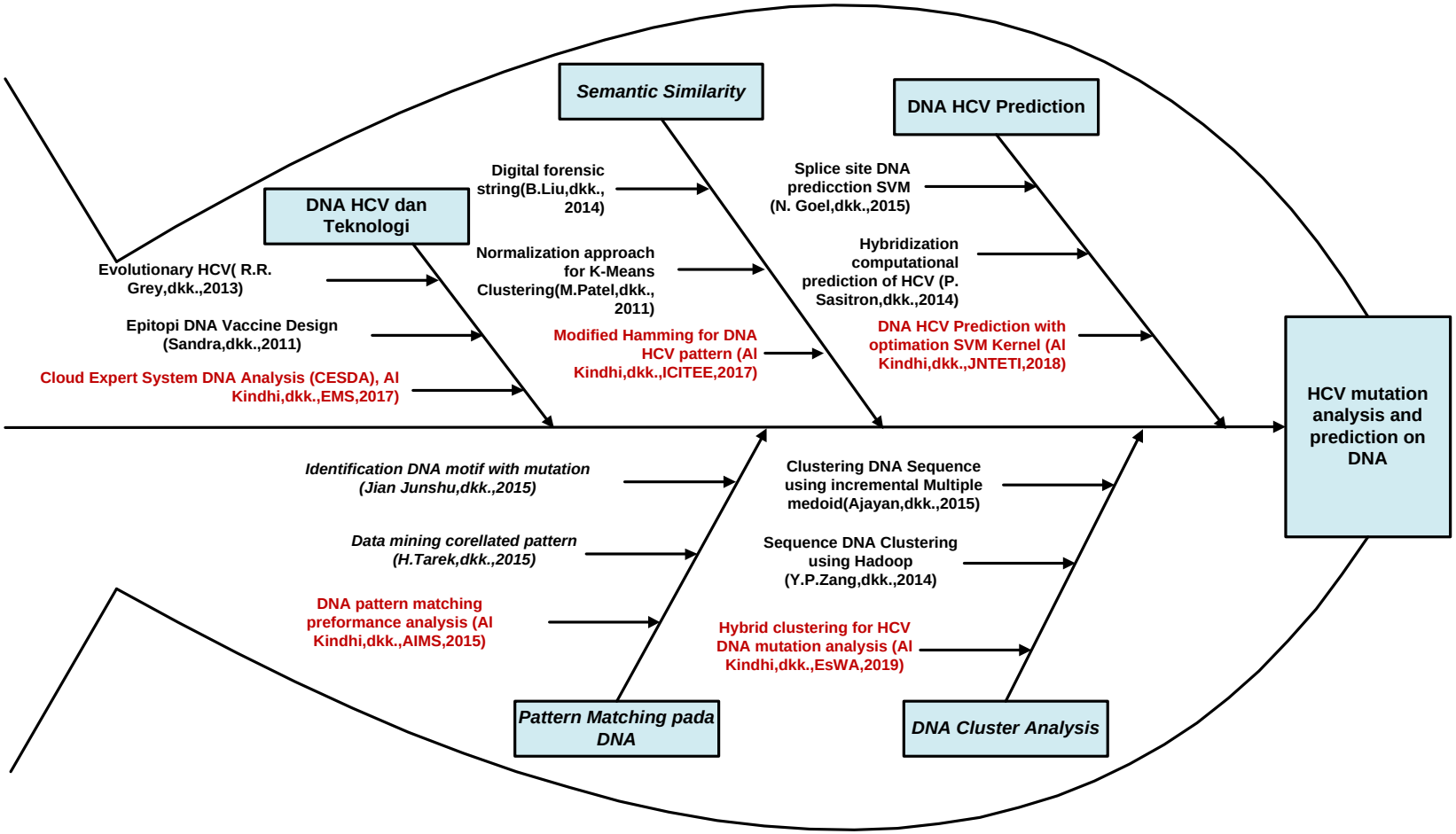
Topik	No	Judul	Hasil
	2	Sasitorn Plakumonthon, "Computational prediction of hybridization patterns between hepatitisC viral genome and human microRNAs"(2014)	Penelitian ini mengambil beberapa human RNA (MiRNA) untuk dibandingkan dengan beberapa primer dan di prediksi apakah RNA tersebut ada kemungkinan mengidap HCV (Sasitorn Plakunmonthon, Nattanan Panjaworayan T-Thienprasert, Kritsada Khongnomnana, Yong Poovorawanc, Sunchai Payungporna, 2014)
	3	T. Feng,et.al., "A medical cost estimation with fuzzy neural network of acute hepatitis patients in emergencyroom"(2015)	Menerapkan FNN (Fuzzy Neural Network) untuk memprediksi biaya seorang pasien hepatitis, dengan menggunakan neuron acak yang diambil berdasarkan pasien hepatitis yang ada sebanyak 110. Hasil penelitian ini menunjukkan bahwa akurasi prediksi total biaya yang dibutuhkan oleh pasien mencapai 90%. (T. Feng, T. S. Li , P. Kuo, 2015).
	4	Neelam Goel,et.al., "An improved method for splice site prediction in DNA sequences using support vector machines (2015)	Melakukan prediksi *pre-messenger-RNA (pre-mRNA),* untuk menentukan manakah splicing yang intron (dibuang) dan exon (bergabung) untuk berbagai tujuan ahli. Mengusulkan perbaikan, dengan menggabungkan dua metode yaitu SVM dan Markov Model (Neelam Goel, Shailendra Singh, Trilok Chand Aseri, 2015).

1.6.2 *Road Map* Penelitian

Berdasarkan telaah pustaka dari penelitian sebelumnya, penelitian DNA predictive modelling dapat dibagi menjadi beberapa topik penelitian yaitu (1) DNA HCV, (2) *DNA Semantic Similarity*, (3) Sistem Pakar DNA Analisis, (4) *DNA Cluster analysis* dan (5) DNA predictive Modelling. Peta jalan penelitian akan dijelaskan pada Gambar 1.1. dan diagram tulang ikan dijelaskan pada Gambar 1.2.



Gambar 1.1. Peta jalan penelitian

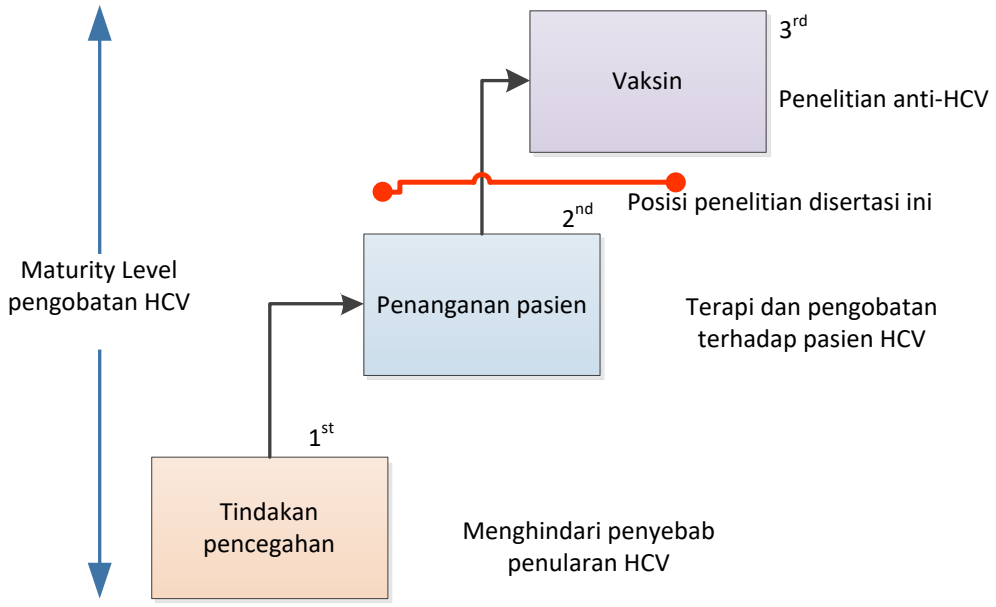


Gambar 1.2. Diagram tulang ikan penelitian

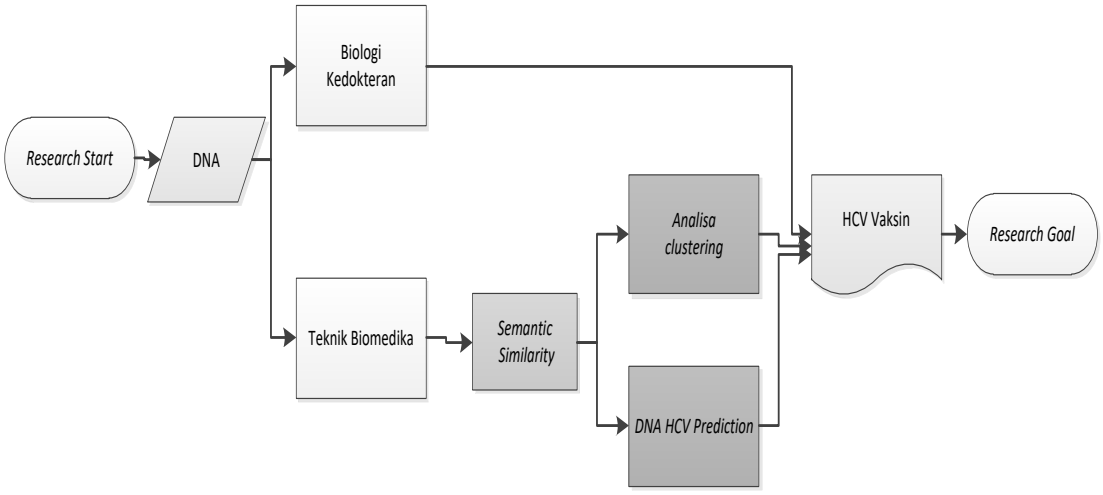
1.7 Posisi Penelitian

Seperti yang telah dijelaskan pada sub-bab sebelumnya, penelitian ini adalah salah satu cabang ilmu penelitian pada bidang DNA khususnya pada DNA HCV. Selain pencegahan dan terapi penanganan pada penderita HCV, tujuan utama dari penelitian di bidang HCV adalah dapat ditemukannya vaksin karena hingga saat ini belum ada vaksin untuk HCV, namun penelitian di bidang ini masih terus dilakukan untuk hasil yang terbaik.

Tingkatan penelitian pada bidang DNA HCV atau yang disebut dengan *maturity level* dapat diamati pada Gambar 1.3, dimana penelitian yang dilakukan pada disertasi ini menuju ke arah vaksin (*maturity level 3).* Untuk dapat menemukan vaksin beberapa tahap penelitian pendahuluan dilakukan yaitu dengan melakukan analisa *clustering* dan prediksi eksitensi HCV dalam suatu *isolated* DNA, dalam arti lain, penelitian disertasi ini adalah pendahuluan yang menuju ke arah *maturity level* 3.



Gambar 1.3. *Maturity level* di penelitian bidang DNA khususnya pada *DNA HCV*



Gambar 1.4. Posisi penelitian yang dilakukan pada disertasi ini terhadap perkembangan penelitian HCV

Melalui Gambar 1.4. dapat diamati bahwa, untuk mencapat *research goal* dapat ditempuh dengan du acara, yaitu secara biologi kedokteran dan secara teknik biomedika. Terdapat dua *output* yang berbeda dari pengolahan data sampel yang digunakan, yaitu analisa *clustering* dan *infected prediction.* Keduanya memiliki tujuan dan hasil penelitian yang berbeda. Namun dua metode tersebut tetap menuju ke arah penelitian vaksin HCV yang merupakan *research goal* dari penelitian di bidang HCV. Untuk mencapai *research goal*, salah satu pendekatan yang dapat dilakukan adalah dalam segi teknik biomedika dengan melakukan analisa dan penghitungan secara matematis.

Tahap *semantic similarity* memiliki peranan penting dalam penelitian di bidang DNA analisis. Algoritma kecerdasan buatan dilakukan dengan penghitungan secara matematis, oleh karena itu jika terdapat data berupa *string* maka langkah awal yang harus dilakukan adalah melakukan normalisasi data ke dalam bentuk numerik yang siap digunakan sebagai *input features* pada mesin pembelajaran. Proses perubahan data mentah menjadi data yang siap digunakan dapat dipelajari pada sub bab berikutnya, yaitu sub bab 1.8. Perubahan Data Sampel.

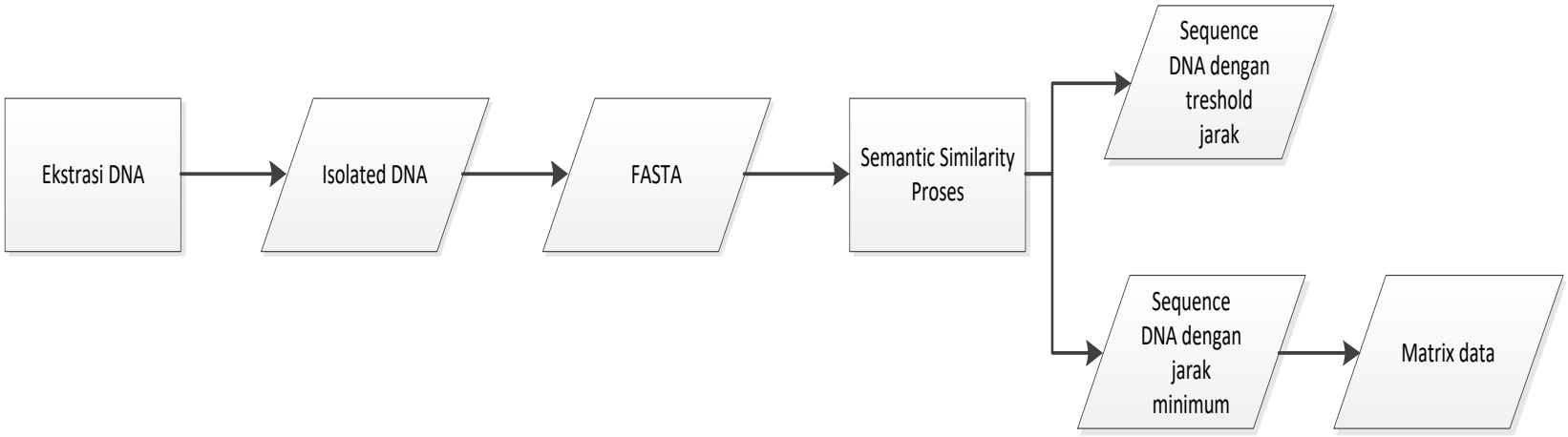
1.8 Perubahan Data Sampel

Normalisasi data mentah ke dalam bentuk data sampel yang diinginkan merupakan satu tahapan penting dalam penelitian ini. Tanpa adanya proses normalisasi, data mentah tersebut tidak dapat diolah ke dalam mesin pembelajaran. Data awal yang digunakan pada penelitian ini adalah *isolated* DNA yang berisi tahun di daftarkannya pada gen bank dunia, origin (negara asal), protein, peneliti yang mendaftarkannya, urutan nukleotida, dan sebagainya.

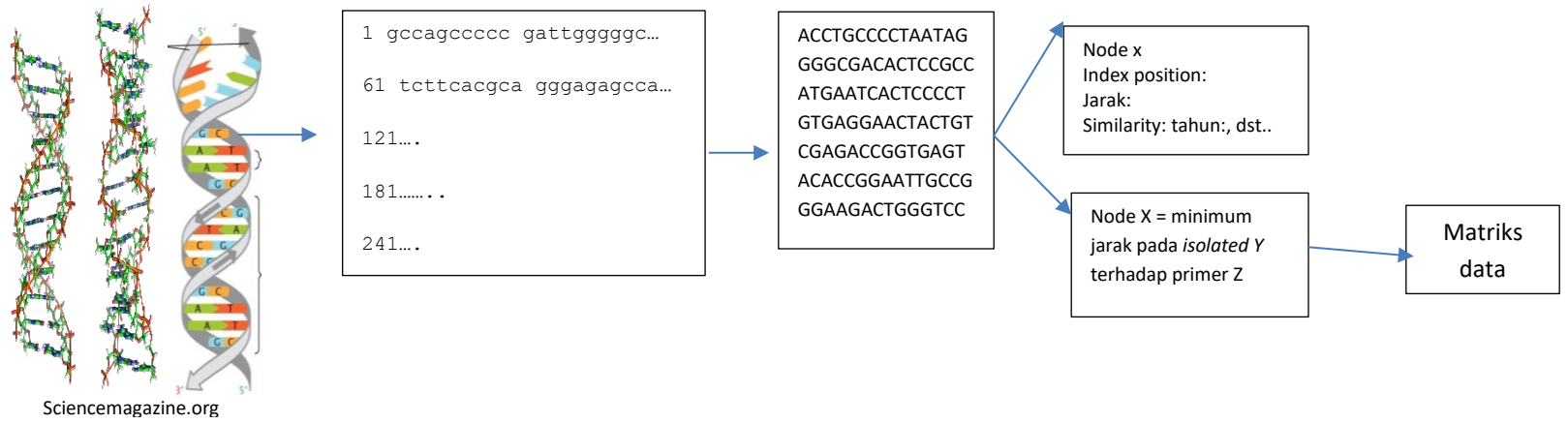
Pada tahap *semantic similarity, isolated* DNA akan diubah menjadi bentuk FASTA. FASTA adalah *string* yang berisi urutan nukleotida yang berasal dari *isolated* DNA yang ditulis dalam karakter huruf besar dan tanpa spasi hingga urutan terakhir dari nukleotida. Seribu data FASTA tersebut kemudian dipotong sepanjang pola primer HCV dan dibandingkan satu persatu dengan primer HCV. Untuk dicari jarak masing-masing urutan tersebut. Siklus perubahan data mentah menjadi data sampel yang siap diolah dapat diamati pada Gambar 1.5.

Gambar 1.6. menjelaskan alur perubahan data pada proses *hybrid clustering,* dimana proses pengelompokan antara data negatif dengan data positif dilakukan secara terpisah. Hal ini dimaksudkan agar hasil pengelompokan data positif tidak terpengaruh oleh data negatif dan mampu menghasilkan analisa yang sesuai dengan yang diharapkan. Selain itu juga memudahkan proses validasi dari hasil pengujian sampel data.

Alur perubahan data pada proses *DNA Prediction* dapat diamati pada Gambar 1.7. Seribu data *isolated* DNA baik positif dan negatif di masukkan secara bersamaan (tidak terpisah). Dari seluruh jarak yang dihasilkan *isolated* DNA terhadap primer, pada akhirnya hanya satu jarak terkecil terhadap masing-masing primer yang di pilih. Jarak terkecil tersebut akan menjadi satu variabel matriks, yaitu sebagai *input features* dari tiap *isolated* DNA.

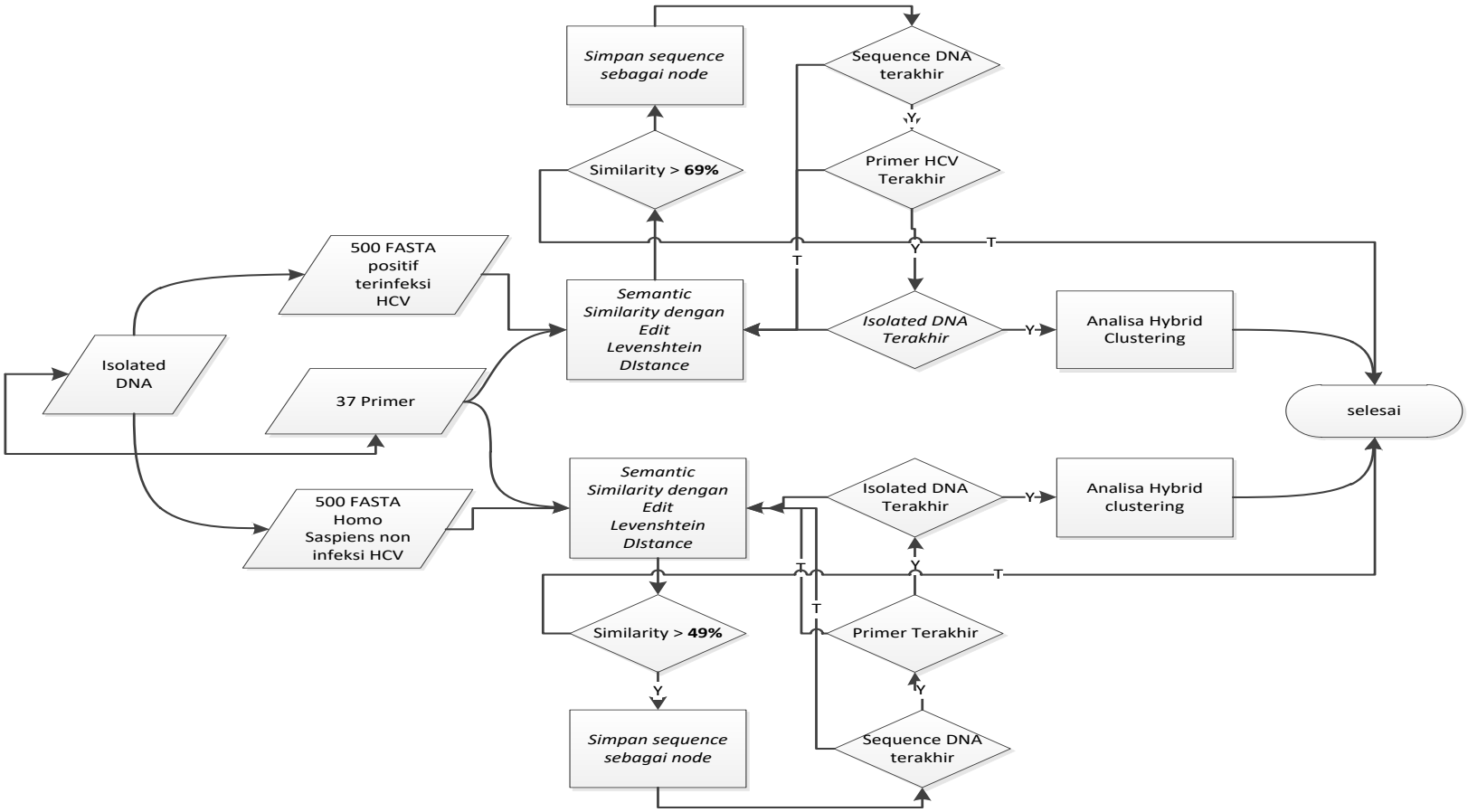


(a)

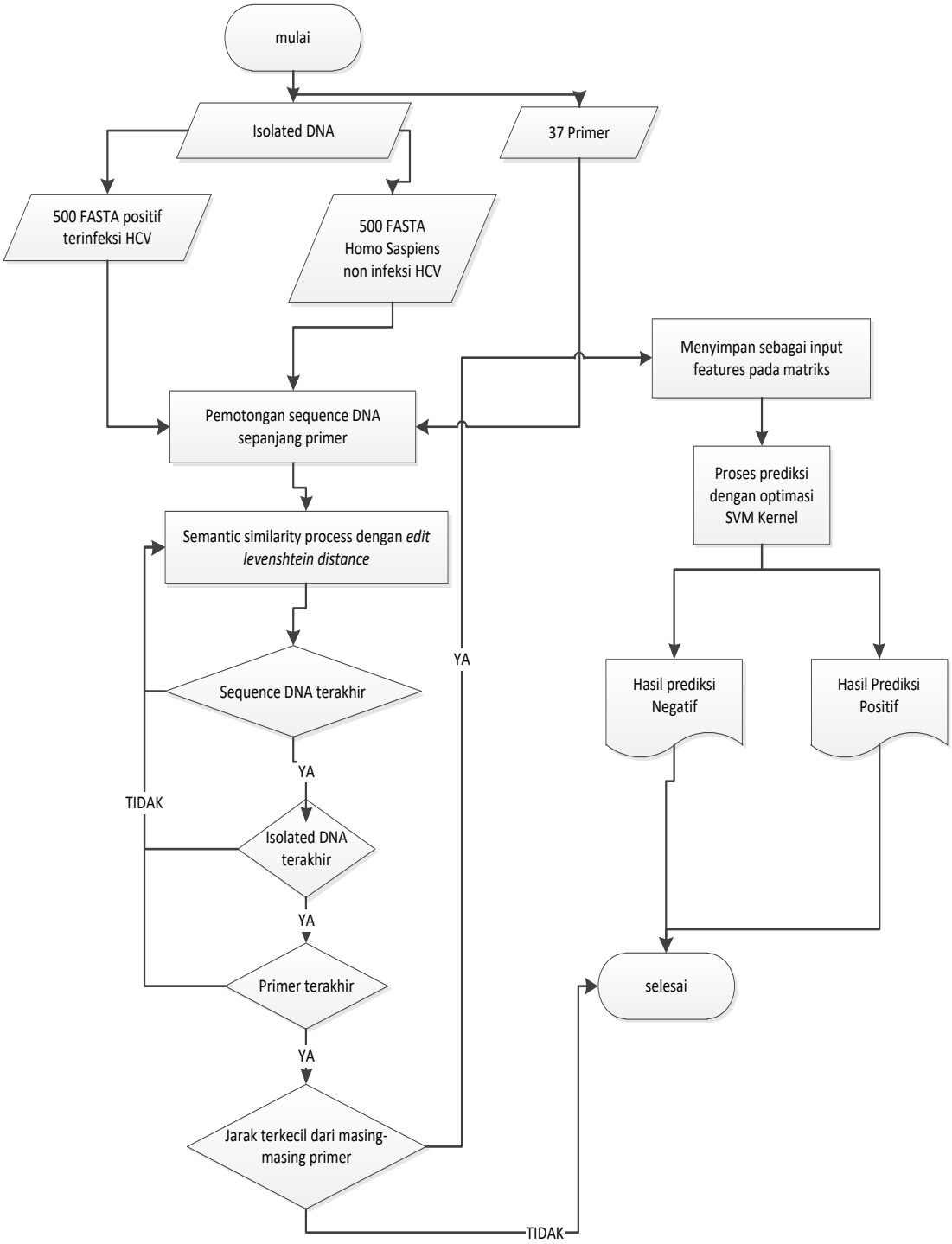


(b)

Gambar 1.5. Alur perubahan data mentah menjadi data set yang siap diproses, (a) alur secara teknis, (b) ilustrasi



Gambar 1.6. Alur perubahan data mentah yang dinormalisasi menjadi data sampel pada tahap pengelompokan (*hybrid clustering*)



Gambar1.7. Alur perubahan bentuk data mentah hingga data berupa hasil prediksi pada tahap DNA *prediction*

Tabel 1.2. Pembuatan matriks *input features* dengan contoh sepuluh data *isolated* DNA terhadap 37 primer HCV

PRIMER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	TARGET
ISOLATES																																							Target
1	12	1	6	7	4	6	9	7	9	4	7	6	6	7	8	5	5	11	5	12	3	12	2	11	4	11	9	10	9	10	11	5	4	7	7	6	10	1
2	12	2	7	7	4	5	9	9	9	4	7	5	6	6	9	5	4	10	4	9	0	10	2	11	4	10	7	11	6	11	11	4	7	6	7	6	11	1
3	12	2	7	7	4	5	9	9	9	4	7	5	6	6	9	5	4	10	4	9	0	10	2	11	4	10	7	11	6	11	11	4	7	5	7	6	11	1
4	11	1	7	7	5	6	9	8	8	4	6	6	7	6	9	6	4	10	4	10	3	11	2	10	5	11	7	11	10	11	11	8	7	6	6	5	7	1
5	13	3	7	7	5	6	9	8	9	8	6	7	7	6	8	6	11	11	12	11	11	10	11	9	11	11	11	10	9	11	11	16	5	7	4	5	10	1
6	13	1	7	7	6	7	9	8	8	7	8	7	6	7	9	6	12	11	12	11	11	11	9	10	11	11	11	10	9	11	11	16	5	7	6	5	9	0
7	12	1	6	7	4	6	10	8	9	4	7	6	7	7	8	5	4	10	4	10	3	11	2	10	4	10	7	11	6	11	10	6	6	6	7	4	11	1
8	11	1	6	7	4	6	10	8	9	4	7	6	7	6	8	5	4	10	4	10	3	11	2	10	4	10	7	11	6	11	10	6	6	6	7	4	11	1
9	13	1	7	8	5	5	10	8	9	4	6	7	6	7	9	5	11	10	11	11	10	12	9	10	4	11	8	11	8	10	11	5	6	6	7	5	10	0
10	13	1	7	8	5	5	10	8	9	4	6	7	6	7	9	5	11	10	11	11	10	12	9	10	4	11	8	11	8	10	11	5	6	6	7	5	10	1

1.9 Sistematika Penulisan Disertasi

Sistematika penulisan buku disertasi ini terdiri dari:

- Bab 1. Memuat hierarki pendahuluan yang terdiri dari latar belakang penelitian yang dilakukan, rumusan masalah, tujuan penelitian, manfaat penelitian, beberapa penelitian pendahulu yang melatar-belakangi penelitian ini, kontribusi dan originalitas penelitian yang dihasilkan dari kegiatan penelitian, serta posisi penelitian dalam bidangnya.
- Bab 2. Membahas tentang kajian literatur yang dipelajari sebagai dasar analisa dalam hal ini ilmu mengenai DNA dan virus, dan metode-metode yang dipelajari untuk diterapkan dalam penelitian.
- Bab 3. Membahas tentang uji coba dan analisa metode *string matching* untuk mencari persamaan primer HCV terhadap *sequence isolated* DNA menggunakan metode Knuth-Morris-Pratt, Boyer Moore, dan Brute Force. Kemudian diperbaiki menjadi uji coba dan analisa *semantic similarity* untuk mencari adanya mutasi pada *sequence isolated* DNA menggunakan metode Hamming dan Edit Levenshtein Distance. Serta tentang pembangunan *expert system* analisis DNA baik dari segi perangkat lunak dan perangkat keras. Yaitu, perancangan infrastruktur aplikasi pengolahan data DNA yang terintegrasi dengan seluruh rumah sakit di Indonesia di bawah pengawasan pemerintah sehingga aplikasi ini mampu memberikan manfaat baik bagi kesehatan maupun penelitian.
- Bab 4. Membahas perangkat lunak untuk pengelompokan *sequence* DNA berdasarkan hasil pengolahan *semantic similarity*. Pengelompokan ini bertujuan untuk mengetahui primer mana yang paling banyak memiliki anggota *sequence* positif, sehingga dapat disimpulkan bahwa primer tersebut adalah primer *trend* yang layak digunakan sebagai parameter pengujian HCV. Selain itu analisa kecenderungan suatu *isolated* DNA terhadap suatu primer dan peta infeksi dari primer hingga daftar *isolated* DNA mana saja yang positf terinfeksi prmer tersebut juga dapat ditampilkan pada hasil pengelompokan.

- Bab 5. Membahas tentang prediksi *isolated* DNA yang terinfeksi HCV. Prediksi ini adalah solusi untuk memprediksi adanya HCV dengan keberagaman pola primer dengan kebaragaman mutasi yang terjadi pada *isolated* DNA dari berbagai negara di dunia.
- Bab 6. Berisikan kesimpulan dari kegiatan hasil penelitian yang telah dicapai dan rencana penelitian selanjutnya.

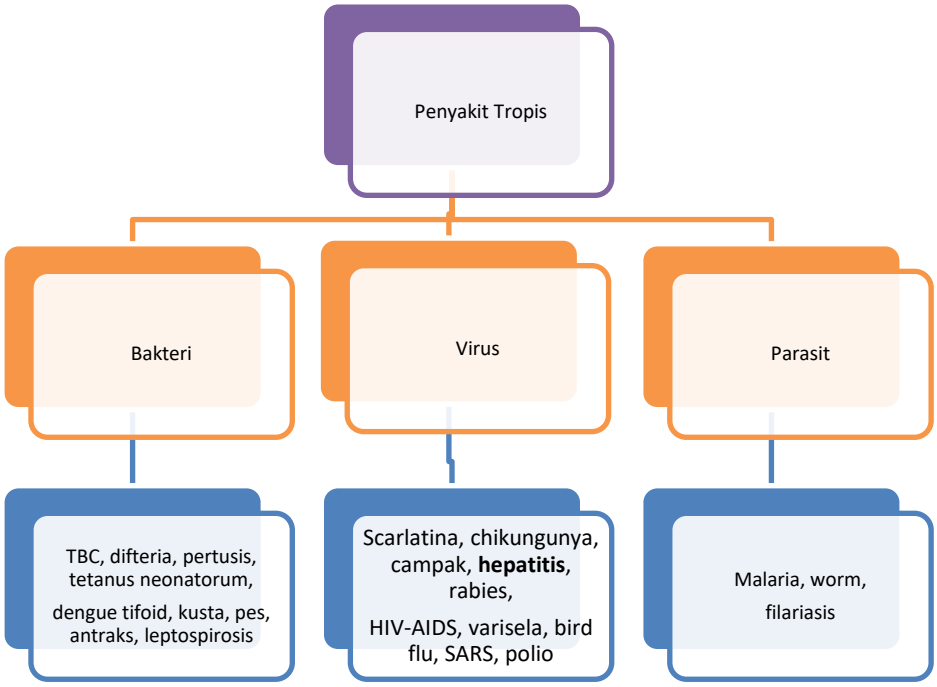
BAB 2
KAJIAN PUSTAKA

Pada bab ini akan dijelaskan beberapa lingkup pengetahuan yang yang digunakan sebagai bahan kajian penelitian untuk membantu dalam proses penyelesaian masalah.

2.1 Penyakit Tropis

Penyakit Tropis adalah penyakit yang hanya terjadi pada daerah tropis atau dengan kata lain peluang kemunculannya lebih besar terjadi pada daerah tropis dibandingkan dengan daerah non-tropis (Airlangga, Ilmu Ilmiah Lanjut Penyakit Tropis, 2012). Bakteri pembawa penyakit tersebut mencakup agen infeksi yang multi resisten dan atau *transibility* (mudah menular).Penularan beberapa penyakit menular sangat dipengaruhi oleh faktor iklim (Brisbois BW, Ali SH, 2010). Organisme pembawa penyakit tropis sangat peka terhadap faktor iklim, khususnya suhu, curah hujan, kelembaban, permukaan air, dan angin. Begitu juga dalam hal distribusi dan kelimpahan dari organisme vektor dan *host intermediate*. Penyakit yang tersebar melalui vektor infeksi perlu diwaspadai karena penularan penyakit seperti ini akan makin meningkat dengan perubahan iklim. Di banyak negara tropis penyakit ini merupakan penyebab kematian utama (C.D. Ramesh , P. Sharmila , G. P. S. Dhillon , P. D. Aditya , 2010).

Epidemiologi siklus mutasi penyakit tropis, khususnya HCV merupakan ancaman bagi kesehatan manusia. Menurut WHO, pada 2018, diperkirakan 71 juta orang terinfeksi HCV akut yang mengakibatkan kanker hati atau sirosis. Selain itu, diperkirakan juga bahwa setiap tahunnya terdapat 399.000 jiwadi seluruh dunia meninggal karena infeksi HCV. Walaupun penelitian yang fokus pada HCV sudah banyak dilakukan, namun hingga saat ini vaksin dari penyakit tersebut belum dapat ditemukan dikarenakan pola mutasinya yang cepat (WHO, 2017).



Gambar 2.1. Jenis-jenis penyakit tropis

Pada Gambar 2.1. dapat diamati bahwa penyakit tropis memiliki tiga jenis, yaitu bakteri, virus, dan parasit. Ketiga jenis penyakit tersebut merupakan organisme yang sangat kecil yang menular melalui infeksi. Pada jenis bakteri dan virus, organismenya hanya dapat diamati melalui bantuan peralatan penelitian. Sedangkan pada parasit, beberapa parasit akan tumbuh subur dan membesar pada makhluk hidup yang ditumpanginya.

2.2 DNA

Asam Deoksiribo Nukleat (DNA), adalah sebuah biomolekul yang berfungsi menyimpan dan menyandi instruksi-instruksi genetika setiap organisme dan banyak jenis virus. Instruksi-instruksi genetika ini berperan penting dalam pertumbuhan, perkembangan, fungsi dari organisme dan virus. DNA merupakan asam nukleat; bersamaan dengan protein dan karbohidrat, asam nukleat adalah makromolekul esensial bagi seluruh makhluk hidup. Molekul DNA terdiri dari dua unting biopolimer yang berpilin satu sama lainnya membentuk heliks ganda. Dua

unting DNA ini dikenal sebagai polinukleotida karena keduanya terdiri dari satuan-satuan molekul yang disebut nukleotida. Tiap-tiap nukleotida terdiri atas satu jenis basa nitrogen (guanina(G), adenina (A), timina(T), atau sitosina(C)), gula monosakarida yang disebut deoksiribosa, dan gugus fosfat. Nukleotida-nukleotida ini kemudian tersambung dalam satu rantai ikatan kovalen antara gula satu nukleotida dengan fosfat nukelotida lainnya. Hasilnya adalah rantai punggung gula-fosfat yang berselang-seling. Menurut kaidah pasangan basa (A dengan T dan C dengan G), ikatan hidrogen mengikat basa-basa dari kedua unting polinukleotida membentuk DNA unting ganda.

Dua unting DNA bersifat anti-paralel, yang berarti bahwa keduanya berpasangan secara berlawanan. Pada setiap gugus gula, terikat salah satu dari empat jenis nukleobasa. Urutan-urutan empat nukleobasa di sepanjang rantai punggung DNA inilah yang menyimpan kode informasi biologis. Melalui proses biokimia yang disebut transkripsi, unting DNA digunakan sebagai tempat untuk membuat unting RNA. Unting RNA ini kemudian ditranslasikan untuk menentukan urutan asam amino protein yang dibangun.

Struktur kimia DNA yang ada membuatnya sangat cocok untuk menyimpan informasi biologis setiap makhluk hidup. Rantai punggung DNA resisten terhadap pembelahan kimia, dan kedua-dua unting dalam struktur unting ganda DNA menyimpan informasi biologis yang sama. Karenanya, informasi biologis ini akan direplikasi ketika dua unting DNA dipisahkan. Sebagian besar DNA (lebih dari 98% pada manusia) bersifat non-kode, yang berarti bagian ini tidak berfungsi menyandikan protein. Dalam sel, DNA tersusun dalam kromosom. Semasa pembelahan sel, kromosom-kromosom ini diduplikasi dalam proses yang disebut replikasi DNA. Organisme Eukariotik (hewan, tumbuhan, fungi, dan protista) menyimpan kebanyakan DNA-nya dalam inti sel dan sebagian kecil sisanya dalam organel seperti mitokondira ataupun kloropas (Russel, 2001). Sebaliknya organisme prokariotik (bakteri dan arkaea) menyimpan DNA-nya hanya dalam sitoplasma. Dalam kromosom, protein kormatin seperti histon berperan dalam penyusunan DNA menjadi struktur kompak.

Struktur kompak inilah yang kemudian berinteraksi antara DNA dengan protein lainnya, sehingga membantu kontrol bagian-bagian DNA mana sajakah yang dapat ditranskipkan.

Para ilmuwan menggunakan DNA sebagai alat molekuler untuk menyingkap teori-teori dan hukum-hukum fisika, seperti misalnya teorema ergodik dan teori elastisitas. Sifat-sifat materi DNA yang khas membuatnya sangat menarik untuk diteliti bagi ilmuwan dan insinyur yang bekerja di bidang mikrofabrikasi dan nanofabrikasi material. Beberapa kemajuan di bidang material ini misalnya origami DNA dan material hibrida berbasis DNA (Mashagi,2013).

2.3 Perbedaan DNA dan RNA

Perbedaan DNA dan RNA secara umum dapat diketahui bahwa DNA mengandung polimer yang lebih panjang dari RNA. DNA (Deoxyribo Nucleic Acid) merupakan tempat penyimpanan informasi genetik. Pada tahun 1953, berhasil ditemukan DNA berstruktur heliks yaitu struktur DNA beruntai ganda (Crick,Watson,1953). Mereka menyebut bahwa DNA heliks mengandung makromolekul plinukleotida yang tersusun secara berulang dari polimer nukleotida. Susunan rangkap membentuk haliks ganda yang menghadap ke kanan. Terdapat tiga gugus molekul dalam setiap nukleotida. Ketiga gugus tersebut adalah gugus folat, gula 5 karbon, dan basa nitrogen atas purin serta adenine.

Letak struktur dapat dijadikan pembeda antara DNA dan RNA. DNA mempunyai letak struktur di mitokondria, sentriol, kloroplas dan inti sel. RNA mempunyai letak struktur di sitoplasma, ribosom, dan inti sel. Bentuk DNA adalah polinukleotida ganda dan terpilin ganda, sedangkan pada RNA berbentuk tunggal dan polinukleotida pendek. Di dalam DNA terdapat gula yang bernama deoxyribosa dan ribose dalam RNA. DNA tergolong dalam beberapa golongan yaitu purin (adenine dan guanine), serta pirimidin (timin dan cytosine). Sementara golongan RNA adalah guanine dan adenine dan juga pirimin (urasil dan cytosine). DNA dan RNA mempunyai fungsi yang berbeda. DNA

mempunyai fungsi sebagai sentesis RNA, yang dilanjutkan pada sintesis protein, serta sebagai pengontrol sifat yang mulai menurun. RNA mempunyai fungsi hanya untuk sintesis protein. Kadar yang ada dalam DNA tidak mendapat pengaruh dari sintesis protein. Kedua pita yang terdapat pada basa nitrogen saling berhadapan dan diikat oleh ikatan hydrogen. Hal ini berbeda dengan RNA yang mendapat pengaruh dari sintesis protein.

Struktur DNA terdiri dari makromolekul berstruktur primer yang dilengkapi rantai rangkap berpilin. Fosfodiester dipilih sebagai tempat penghubung para struktur DNA. DNA Heliks ganda mempunyai polaritas yang berlawanan dengan orientasi yang mempunyai tiga model yang digunakan untuk mengetahui peristiwa pergerakan DNA. Fungsi RNA sebagai penyalur informasi dan juga penyimpanan genetik yang merupakan proses translasi yang mempunyai tujuan untuk sistensis protein. Ketidak samaan struktur pada DNA dan RNA menjadikan perbedaan antara keduanya, namun keduanya sama-sama tersusun dari dari nukleotida.

Berikut adalah penjelasan singkat mengenai perbedaan antara DNA dengan RNA:

1. Deoksiribosa: adalah gula penyusun DNA, sedangkan gula RNA disusun oleh ribosa
2. Timin dimiliki oleh DNA dan uracyl adalah milik RNA. Basa nitrogen yang terkandung dalam DNA disusun oleh purin yang berasal dari susunan Guanin (G) dan Adenin(A) serta pirimidin yang berasal dari susunan Cytocine(C) dan Timin (T). Sedangkan basa nitrogen RNA disusun oleh purin Guanin (G) dan Adenin(A), serta Pirimidin Uracyl(U) dan Cytocine(C).
3. DNA mempunyai rantai panjang dan ganda berpilin (double helix), sedang rantai tunggal dan pendek dimiliki oleh RNA.
4. DNA dapat dijumpai di kloropas, mitokondira, dan nukleus. RNA dapat dijumpai di ribosom (r-RNA), sitoplasama (t-RNA), dan nukleus (m-RNA).

5. DNA mempunyai peranan mewariskan sifat serta mensintesis protein. Sedang RNA mempunyai peranan hanya untuk mensintesis protein.
6. Sel DNA mempunyai sel tetap, sedangkan kadar RNA berubah-ubah. Hal ini disesuaikan dengan jumlah protein yang dibutuhkan

2.4 *Sense* dan *Anti-sense*

Sebuah urutan sekuens DNA disebut sebagai "sense" apabila urutan basa DNA-nya sama dengan urutan kopi RNA duta yang ditranslasikan menjadi protein (Gregory,et.al.,2006). Urutan pada unting komplementernya disebut sebagai urutan "antisense". Baik urutan sense dan antisense dapat ditemukan pada berbagai bagian unting DNA yang sama (kedua unting DNA dapat mengandung baik urutan sense maupun antisense). Pada prokariota dan eukariota, urutan RNA antisense juga diproduksi, namun fungsi RNA antisense ini tidaklah diketahui dengan jelas. RNA antisense diajukan terlibat dalam regulasi ekspresi gen melalui pemasangan basa RNA-RNA (Munroe,2004).

Pada sebagian kecil urutan DNA prokariota dan eukariota, dan sebagian besar urutan DNA plasmid dan virus, perbedaan antara unting sense dan antisense menjadi kabur dikarenakan terdapatnya gen yang tumpang tindih (Makalowska,2004). Dalam hal ini, beberapa urutan DNA memiliki tugas ganda, yakni menyandikan protein pertama ketika dibaca melalui salah satu unting, dan menyandikan protein kedua ketika dibaca dengan arah berlawanan melalui unting komplementernya. Pada bakteri, ketumpang-tindihan ini kemungkinan terlibat dalam regulasi transkripsi gen. Sedangkan pada virus, gen yang tumpang tindih ini meningkatkan jumlah informasi yang dapat disandikan dalam genom virus yang berukuran kecil (Johnson, Chisholm,2004).

2.5 Hepatitis C Virus (HCV) dan DNA

HCV pertama kali ditemukan pada tahun 1989 oleh Michael Hougton, salah seorang profesor di Fakultas Kedokteran Universitas Alberta, Canada. Hingga saat ini beliau dan tim laboratoriumnya fokus mengembangkan vaksin untuk menyembuhkan pasien infeksi HCV. Hepatitis C adalah salah satu penyakit yang dapat menyerang hati. Penyakit yang disebabkan oleh virus ini dapat memicu infeksi dan inflamasi pada hati.

Hepatitis C umumnya tidak menunjukkan gejala pada tahap-tahap awal. Karena itu, sekitar 75 persen penderita hepatitis C tidak menyadari bahwa dirinya sudah tertular sampai akhirnya mengalami kerusakan hati bertahun-tahun kemudian. Meski ada gejala hepatitis C yang muncul, indikasinya mirip dengan penyakit lain sehingga sulit disadari. Beberapa di antaranya meliputi selalu merasa lelah, pegal-pegal, serta tidak bernafsu makan.

Virus Hepatitis C adalah jenis virus RNA yaitu virus yang bertanggung jawab pembawa pesan kode untuk pembentukan DNA/protein baru. Jika RNA terinfeksi virus, akibatnya DNA yang terbentuk juga akan berubah. Sifat alami sel adalah mempertahankan diri supaya tetap hidup atau malah mati dengan mengubah pola kode DNA yang baru melalui perubahan pola RNA.

RNA yang terinfeksi HCV, juga akan membentuk pola kode DNA yang berbeda dengan aslinya. Pada mahluk hidup secara alami dan otomatis, DNA abnormal seharusnya akan mati, tetapi sifat lainnya adalah berusaha beradaptasi dengan mengubah pola kode RNA sehingga DNA yang terbentuk dapat bertahan hidup. Sifat RNA yang dapat beradaptasi berakibat pola kode DNA baru juga berubah. Siklus ini berjalan berantai terus menerus, DNA kode baru akan membuat sinyal RNA baru, RNA yang baru terinfeksi lagi dengan HCV akan membentuk DNA baru lagi, kedepannya RNA yang baru beradaptasi membentuk pola kode baru untuk DNA yang akan dibentuk dan berulang lagi.

Sehingga dengan kondisi yang seperti ini, pada HCV pattern DNA sel selalu dapat berubah-ubah (sebagian mungkin tetap dengan pola kode yang sama, sebagian

mengalami perubahan). Pola-pola RNA inilah yang disebut subtype genome dari HCV. Oleh karena itu semakin lama jumlah subtype akan bertambah seiring dengan semakin beragamnya virus mempertahankan diri dengan pola RNA sehingga mengubah susunan nukleotida pada DNA pula.

2.6 Penambangan Data

Nama sistem pakar dan penambangan data sebenarnya mulai dikenal sejak tahun 1990, ketika pekerjaan pemanfaatan data menjadi sesuatu yang penting dalam berbagai bidang, mulai dari bidang akademik, bisnis, hingga medis (Mohammed J. Zaki, Wagner Meira, 2014). Penambangan data dapat diterapkan pada berbagai bidang yang mempunyai sejumlah data, tetapi karena wilayah penelitian dengan sejarah yang belum lama, maka data mining masih diperdebatkan posisi bidang pengetahuan yang memilikinya. Maka, J. Zaki menyatakan bahwa "penambangan data adalah campuran dari statistik, kecerdasan buatan, dan riset basis data" yang masih berkembang (Mohammed J. Zaki, Wagner Meira, 2014).

2.6.1 Akar Ilmu Penambangan Data

Knowledge-Discovery in Database (KDD) adalah makna lain untuk penambangan data, dimana tujuannya sama yaitu untuk memanfaatkan data dalam basis data dan mengolahnya menjadi suatu informasi yang berguna. Pada gambar 2.2. dapat dianalisa akar keilmuan yang menyangkut penambangan data adalah sebagai berikut:

1. Statistik

   Statistik mampu mengolah data menjadi sebuah *exploratoty data analysis (EDA). EDA* berguna untuk mengidentifikasi hubungan sistematis antar variabel/fitur ketika tidak ada cukup informasi alami yang dibawanya. Penghitungan statistik dapat memberikan dua jenis analisa yaitu analisa mundur (mengevaluasi hal yang telah lampau) dan analisa maju (memprediksi hal yang akan terjadi nanti).

2. Kecerdasan Buatan (*Artificial Intelligence*)

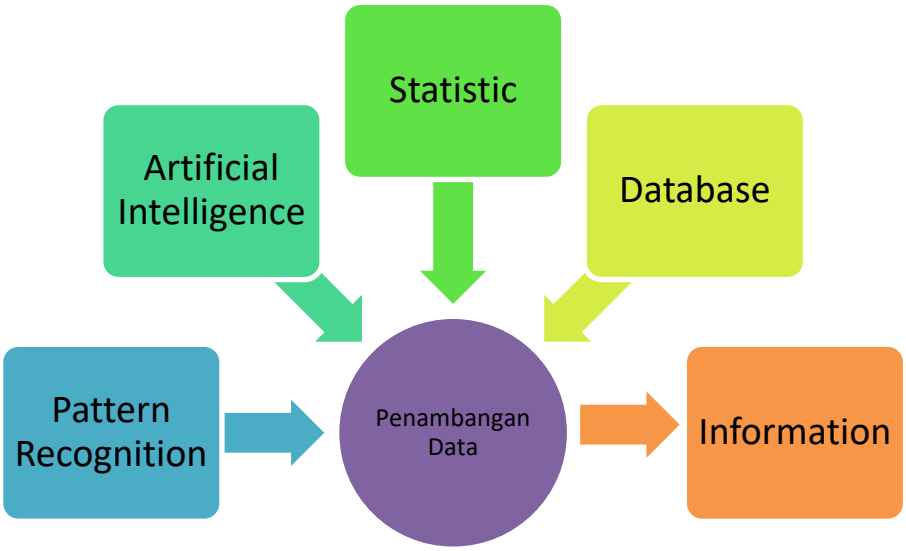
Teori kecerdasan buatan (AI) dibangun berdasarkan teknik heuristik sehingga AI berkontribusi terhadap teknik pengolahan informasi bedasarkan pada model penalaran manusia. Salah satu cabang dari AI, yaitu pembelajaran mesin atau *machine learning*, merupakan disiplin ilmu yang paling penting yang direpresentasikan dalam pembangunan data mining, menggunakan teknik pembelajaran dan pelatihan pada sistem komputer (*machine learning)*.

3. Pengenalan Pola

Data mining dalam pengenalan pola berfungsi untuk mengolah data dari *database*. Data yang diolah bukan dalam bentuk relasi, melainkan dalam bentuk normal sehingga set data dibentuk menjadi data normal pertama. Kemudian dari data normal pertama tersebut akan dicari kecocokannya dengan suatu pola yang diberikan.

4. Basis Data

Basis data menyediakan informasi berupa data-data mentah yang siap digali menggunakan metode-metode yang disebutkan pada nomer 1-4.



Gambar 2.2. Akar ilmu penambangan data

2.6.2 Proses Penambangan Data

Secara sistematis, ada tiga langkah dalam pemrosesan penambanngan data yaitu: (Mohammed J. Zaki, Wagner Meira, 2014)

1. Eksplorasi/pemrosesan awal data

   Eksplorasi/pemrosesan awal terdiri dari normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.
2. Membangun model dan melakukan validasi terhadapnya

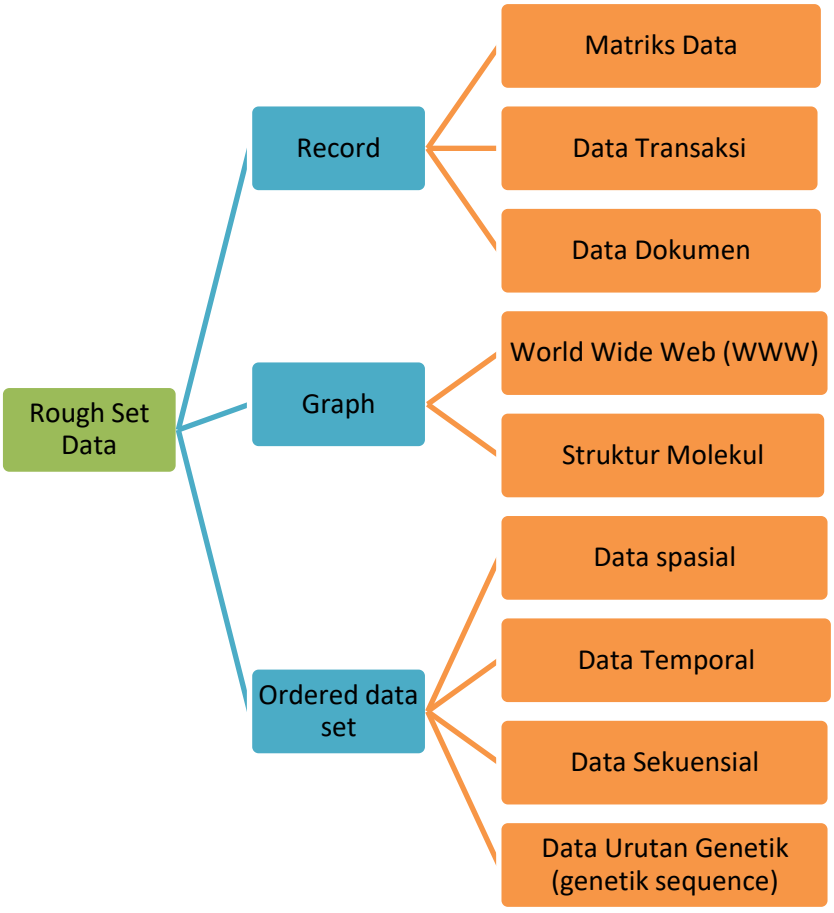
   Melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik.
3. Penerapan

   Menerapkan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang diinvestigasi.

2.6.3 Rough Set Data

"Data" dalam terminologi statistik adalah kumpulan objek dengan atribut-atribut tertentu, objek tersebut adalah individu berupa data di mana setiap data memiliki sejumlah atribut. Atribut tersebut berpengaruh pada dimensi dari data. Semakin banyak atribut/fitur maka semakin besar dimensi data sehingga membentuk suatu himpunan data.

Seperti yang dijelaskan pada gambar 2.2., terdapat tiga kategori dalam *data set*, yaitu *record*, *graph*, dan *ordered data set.* Pada penelitian ini, data *isolated* DNA dapat dikategorikan ke dalam *ordered data set,* yaitu urutan nukleotida yang memiliki informasi tertentu.



Gambar 2.3. Jenis-jenis data mentah sebelum di olah

2.7 *Similarity Based Distance*

Untuk mencari anggota kelompok dalam metode pengelompokan maupun pencarian kemiripan dalam data semantik dibutuhkan metode pencarian jarak terdekat, metode tersebut diantaranya adalah:

A. *Manhattan Distance (L1)*

Adalah metode untuk mengukur jarak dua buah objek dengan Persamaan (2.1) sebagai berikut:

$$d_{L1}(X_1, X_2) = \sum_{i=0}^{n} |X_1 i - X_2 i| \qquad (2.1)$$

Keterangan : *n*= jumlah total *record data*
i= iterasi mulai dari 0
$X_1 i$= nilai objek pertama
$X_2 i$=nilai objek ke-2

B. *Euclidean Distance* (d(x))

Dalam metode pengelompokan, hal yang perlu diperhatikan adalah pemilihan metrik yang digunakan untuk mengukur ketidakmiripan data yang dikelompokkan. Penggunaan metrik yang berbeda dapat memberikan hasil yang berbeda tergantung kasus yang diselesaikan. Metrik yang paling banyak digunakan adalah *Euclidean distance (*jarak Euclidean). Jarak Euclidean dapat dianggap sebagai jarak yang paling pendek antara dua titik dan pada dasarnya sama halnya dengan persamaan Pythagoras ketika digunakan di dalam dua dimensi. Secara matematis, jarak Euclidean dapat dituliskan di dalam Persamaan (2.2).

$$d(x) = ||x|| = \sqrt{\sum_{i=0}^{n} (X_{1_i} - X_{2_i})^2} \quad ; i = 1,2,3, \ldots, \qquad (2.2)$$

Keterangan : n = jumlah total *record data*

I = iterasi mulai dari 0

X_{1_i} = nilai objek pertama

X_{2_i} = nilai objek ke-2

Ketika menggunakan fungsi Euclidean untuk membandingkan jarak, tidak perlu menghitung akar dua sebab jarak selalu merupakan angka-angka positif. Untuk dua jarak d1 dan d2, jika :

$$\sqrt{d_1} > \sqrt{d_2} \quad \Leftrightarrow \quad d_1 > d_2 \tag{2.3}$$

C. *Chebyshev Distance (Chessboard Distance, L_Infinity Norm)*

$$d(X_1, X_2) = Max\ \{X_1 i - X_2 i\} \tag{2.4}$$

Keterangan : I = iterasi mulai dari 0

$X_1 i$= nilai objek pertama

$X_2 i$= nilai objek ke-2

D. *Mahalanobis Distance*

Metode untuk mengukur jarak dua buah objek dengan memikirkan korelasi antar objek dalam bentuk vektor variabel dari objek dan matrik *covariance* dari kedua objek tersebut dengan Persamaan (2.5).

$$d\ (X_1, X_2) = \sqrt{((X_1 - X_2)^2 * \ cov^{(-1)} * (X_1 - X_2))} \tag{2.5}$$

Keterangan : X_1= nilai objek pertama

X_2 =nilai objek ke-2

Apabila matrik *covariance* adalah matrik *identity*, maka *Mahalanobis distance* adalah *Euclidean Distance*, dan apabila matrik *covariance* adalah matrik diagonal, maka *Mahalanobis* adalah *Normalised Euclidean Distance* yaitu korelasi antar objek dianggap tidak ada. Dalam hal ini *Mahalanobis distance* dihitung sesuai dengan Persamaan (2.6).

$$d(X_1, X_2) = \sqrt{(\sum_{i=0}^{n} \frac{(X_1 i - X_2 i)^2}{(jumlah\ i)^2}} \quad (2.6)$$

Keterangan : n = jumlah total *record data*

i = iterasi mulai dari 0

$X_1 i$ = nilai objek pertama

$X_2 i$ = nilai objek ke-2

E. *Hamming Distance*

Metode *Hamming distance* digunakan untuk mengukur jarak antara dua *string* yang ukurannya sama dengan membandingkan simbol-simbol yang terdapat pada kedua *string* pada posisi yang sama. *Hamming distance* dari dua *string* adalah jumlah simbol dari kedua *string* yang berbeda. Pada Persamaan (2.7), jarak Hamming dapat dituliskan dengan $h(x, y)$, dengan *a* adalah karakter *a* dan *b* adalah karakter *b*, serta *i* dan *j* adalah dua *string* yang dibandingkan. Jarak terbaik yang dihasilkan Hamming adalah ketika dua *string* yang dibandingkan adalah 0 (nol). Hal itu terjadi jika karakter yang dibandingkan pada *string a* dan *string b* sama persis sehingga menghasilkan jarak nol. Namun jika terdapat karakter yang berbeda akan menghasilkan jarak 1. Sebagai contoh *Hamming distance* antara string ‘elektro’ dan ‘electra’ adalah 2. *Hamming Distance* juga digunakan untuk

mengukur jarak antar dua s*tring binary* misalnya jarak antara 11011001 dengan 10101010 adalah 6.

$$h(x,y) = \sum_{a=0}^{n} \begin{Bmatrix} min_{(x,y)} = 0 \rightarrow jika\ a_x \neq b_y \\ min_{(x,y)} = 1 \rightarrow jika\ a_x\ = b_y \end{Bmatrix} \tag{2.7}$$

F. *Levenshtein Distance*

Metode *Levenshtein Distance* digunakan untuk mengukur jarak antara dua *string* dengan menghitung jumlah pengoperasian yang perlu dilakukan untuk mengubah *string* yang satu menjadi *string* yang kedua yang diperbandingkan. Pengoperasian yang dilakukan termasuk operasi *insert*, *delete* dan substitusi. Sebagai contoh *Levenshtein distance* antara kata 'kitten' dan 'sitting' adalah 3 dengan pengoperasian substitusi k dengan s, substitusi e dengan i, dan *insert* g.

$$lev_{a,b}\ (i,j) = \begin{cases} \max(i,j) \rightarrow jika\ min_{(i,j)} = 0 \\ \min \begin{cases} lev_{a,b}\ (i-1,j) + 1 \\ lev_{a,b}\ (i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_i)} \end{cases} \end{cases} \tag{2.8}$$

Persamaan (2.7) adalah cara menghitung jarak kemiripan setiap urutan. Misalkan pola primer adalah dan urutan DNA b, maka variabel i adalah jumlah karakter a dan j adalah jumlah karakter b. Oleh karena itu, jarak antara pola utama dan urutan DNA menggunakan Levenshtein dapat ditulis dengan $lev_{a,b}\ (i,j)$. Penghitungan max (i, j) dapat terjadi jika jarak minimum *i* dan *j* nol atau dapat ditulis dengan $\min(i,j) = 0$. Jika jarak antara *i* dan *j* tidak 0, maka perhitungan levenshtein dapat menggunakan tiga kondisi minimum. Dimana, $1_{(a_i \neq b_i)}$ adalah fungsi indikasi yang ekivalen dengan 0 jika $a_i = b_j$, jika $a_i \neq b_i$ maka fungsi

indikasinya adalah 1. Pada persamaan minimum, $lev_{a,b}(i-1,j)+1$ adalah fungsi untuk *delete* (menghapus), $lev_{a,b}(i,j-1)+1$ adalah fungsi untuk *insert* (penambahan), dan $lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_i)}$ adalah fungsi untuk subtitusi.

G. *Hausdorff Distance*

Metode *Hausdorff Distance* digunakan untuk mengukur jarak berbasis nilai infimum (*Greatest lower bound)* dan supremum (*greatest upper bound)* dari kedua objek, yaitu semua variabel dari kedua objek tersebut mempunyai nilai *compact/closed*. *Hausdorff distance* dihitung seperti pada Persamaan (2.8).

$$d(X_1, X_2) = \max\{\sup(X_1 i \ in \ X_1) * \inf(X_1 i \ in \ X_2), \sup(X_2 i \ in \ X_2) * \inf(X_1 i \ in \ X_1)\} \quad (2.9)$$

Keterangan : $X_1 i \ in \ X_1$= nilai X_1i pada data X_1

$X_2 i \ in \ X_2$= nilai X_2i pada data X_2

BAB 3
DNA SEMANTIC SIMILARITY

Semantik adalah sebuah kata atau sekumpulan karakter yang ditinjau dari susunan karakternya berbeda namun memiliki makna yang sama. *Semantic Similarity* adalah metrik yang didefinisikan atas sekumpulan dokumen atau istilah yang gagasan jarak di antara mereka didasarkan pada kesamaan makna atau konten semantik yang bertentangan dengan kesamaan yang dapat diperkirakan mengenai representasi sintaksis mereka (misalnya format *string* mereka). Pada studi kasus ini, semantik data yang dimaksud adalah primer. Penelitian ini menggunakan 37 primer yang memiliki jumlah karakter dan susunan nukleotida yang berbeda-beda, namun memiliki makna yang sama. Seluruh primer tersebut memiliki makna yaitu kunci dari HCV. Primer digunakan sebagai parameter penentu bahwa suatu *isolated* DNA tersebut mengandung HCV walaupun susunan nukleotida yang diukur dapat saja berbeda-beda satu sama lain.

Pada tahap penelitian ini, ke-37 kata semantik (primer HCV) tersebut akan dibandingkan dengan data *isolated* DNA. Proses perbandingan karakter dilakukan dengan pendekatan pencocokan pola (metode *string matching*) dan *approximate string matching* (*Hamming distance* dan *Edit Levenshtein distance).* Data contoh yang kami gunakan adalah *isolated* DNA yang positif terinfeksi HCV. *Isolated* DNA tersebut kami ubah ke dalam bentuk FASTA terlebih dahulu. FASTA adalah urutan nukleotida suatu makhluk hidup dalam format *string data*, ditulis dengan karakter besar tanpa spasi hingga urutan terakhir. Sedangkan pada *isolated* DNA, urutan nukleotida ditulis setiap sepuluh karakter dan dipisahkan dengan spasi, karena dalam satu unting DNA terdapat 10 nukleotida. Setiap karakter mewakili satu dari empat nukleotida yang telah dijelaskan pada bab sebelumnya. Kemudian bentuk FASTA tersebut kami simpan dalam bentuk .txt file agar sistem lebih mudah membacanya. Pada waktu pemrosesan,

file kami panggil berdasarkan namanya dan kami baca satu perssatu sesuai dengan urutan iterasi.

3.1 Pencocokan Pola

Metode pencocokan pola digunakan sebagai metode awal dari *road map* penelitian. Metode ini berfungsi untuk menemukan adanya urutan yang sama dengan primer HCV di dalam *isolated* DNA. Dalam satu *isolated* DNA *homo sapiens* terdiri dari 10.000 hingga 15.000 urutan nukleotida, sedangkan data yang diujikan dapat mencapai ratusan hingga ribuan *isolated* DNA. Oleh karena itu dibutuhkan suatu metode pencocokan pola yang efisien dan efektif dalam menemukan pola tersebut pada seluruh *isolated* DNA yang diujikan. Pada sub bab ini akan dikaji tiga metode pencocokan pola untuk membandingkan performanya dalam kasus pencarian pola primer pada *isolated* DNA, ketiga metode yang diuji tersebut adalah Knuth Moris Pratt, Boyer Moore, dan Brute Force yang akan dijelaskan masing-masing cara kerjanya.

3.1.1 Algoritma Knuth Morris Pratt

Algoritma Knuth Morris Pratt menggunakan *prefix* dan *suffix* dari *pattern* untuk mengoptimasi pergeseran *pattern* dalam pencarian. Ketika memulai perbandingan, suatu *string* akan dibandingkan kesamaan *prefix* dan *suffix*-nya. Jika keduanya tidak sama, maka perbandingan akan melompat ke iterasi berikutnya. Sehingga karakter yang pertama kali diambil untuk perbandingan adalah karakter yang pertama dan yang terakhir. Ilustrasi perbandingan akan dijelaskan pada Gambar 3.2., melalui gambar tersebut dapat diimplementasikan ke dalam *pseudocode* pada Gambar 3.1.

```
Kompleksitas waktu total : O(n + m)
   Txt[i..i +n-1]
   then,
   incompatibility =  text[i+j]  & pattern [j], where
   0<j<n.
   So:
   Text[i..i+j-1] = pattern [0..j-1]
   And
    a=text[i+j]  ≠ b=pattern[j]
   Total time complexity: O(i + j)
```

Gambar 3.1. *Pseudocode* algoritma *Knuth-Morris-Pratt*

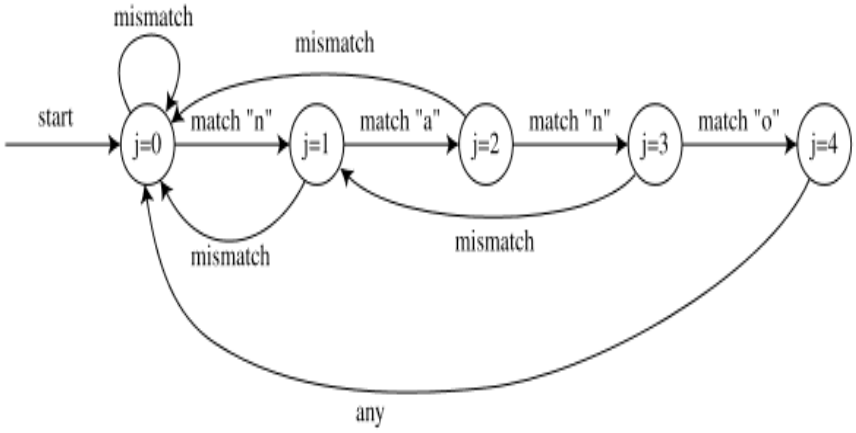
Misalkan:

Urutan DNA yang diujikan: TTAATTACCT

Pattern: AATT

Tabel 3.1. Tabel fungsi pembatas Knuth-Morris-Pratt

P	0	1	2
B(p)	0	1	2



Gambar 3.2. Alur algoritma Knuth Morris Pratt

f(MATCH) = MATCH – b(MATCH-1) untuk MATCH >= 1

f(MATCH) = 1 untuk MATCH = 0

T	T	A	A	T	T	A	C	C	T
X	**X**								
A	A	T	T						

Pergeseran = f(MATCH) = 1-b(0) = 0

T	T	A	A	T	T	A	C	C	T
	X	**X**	**X**						
	A	A	T	T					

Pergeseran = f(MATCH) = 1-b(0) = 1

T	T	A	A	T	T	A	C	C	T
		X	**X**	**X**	**X**				
		A	A	T	T				

Pergeseran = f(MATCH) = 1

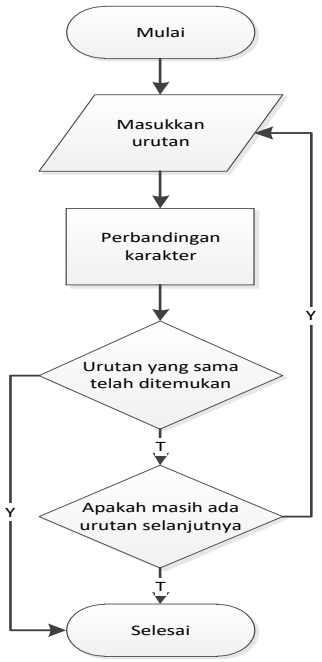
Jumlah perbandingan: 9

Dalam algoritma ini, hasil perbandingan sebelumnya akan tetap disimpan ke dalam histori untuk menghindari proses perbandingan yang sia-sia. KMP menggunakan prefix dan suffix dari pattern untuk mengoptimasi pergeseran *pattern* dalam pencarian.

3.1.2 Algoritma Brute Force

Metode *pattern matching* yang diujikan berikutnya adalah metode Brute Force. Metode Brute Force memiliki teknik pencarian sederhana dan memiliki aturan normal (*normal term),* sehingga untuk proses pencariannya lebih mudah. Metode pencarian dimulai dengan karakter paling awal kemudian ke karakter berikutnya hingga

menemukan karakter yang tidak cocok, maka perbandingan akan bergeser ke iterasi berikutnya. Jika *string data* cocok dengan *pattern* yang diberikan akan disimpan, kemudian melanjutkan pencarian. *Flowchart* metode *Brute Force* dapat dijelaskan dalam gambar 3.3.



Gambar 3.3. Diagram alur Algoritma Brute Force

Melalui diagram alur pada Gambar 3.3., dapat dijelaskan prosedur pencarian algoritma Brute Force seperti pada Gambar 3.4.

```
(input a1,a2,...an:string, x:string, output idx :
    integer)
K←1
While (k < n) and (ak ≠ x) do
K← k+1
Endwhile
{k= n or ak=x}
If ak = x then {x found}
Idx ← k
Else
Idx ← 0 {x not found}
Endif
```

Gambar 3.4. *Pseudocode* algoritma Brute Force

Dimana kompleksitas waktu terburuk pada algoritma Brute Force adalah O(MN) dan kompleksitas waktu terbaik adalah O(N).

Misalkan:

Urutan DNA yang diujikan: TTAATTACCT

Pattern: AATT

Maka pencarian *pattern* berdasarkan algoritma Brute Force adalah:

T	T	A	A	T	T	A	C	C	T
X									
A	A	T	T						

Pergeseran = 1

T	T	A	A	T	T	A	C	C	T
	X								
	A	A	T	T					

Pergeseran = 1

T	T	A	A	T	T	A	C	C	T
		X	**X**	**X**	**X**				
		A	A	T	T				

Pergeseran = 4

Jumlah perbandingan: 6

Brute Force akan mencocokkan *string* disetiap karakter untuk menentukan apakah *pattern* yang dimaksud terdapat di posisi tersebut. Dengan algoritma *Brute Force*, proses pembandingan *pattern* akan maju satu langkah ke kanan dan mulai mencocokkan lagi sampai bertemu dengan karakter yang tidak cocok, *pattern* yang dimaksud sudah ditemukan, atau pencarian sudah mencapai ujung teks.

3.1.3 Algoritma Boyer Moore

Algoritma pencocokan *string* ketiga yang diujikan adalah algoritma Boyer Moore. Algoritma ini membandingkan awalan dan akhiran dari suatu *string*, sehingga ketika tidak ditemui kecocokan pada pola, proses pencarian akan melanjutkan ke *string* berikutnya.

Misalkan:

Urutan DNA yang diujikan: TTAATTACCT

Pattern: AATT

Maka pencarian *pattern* berdasarkan algoritma Boyer Moore adalah:

T	T	A	A	T	T	A	C	C	T
X			**X**						
A	A	T	T						

Pergeseran = 1, karena akhiran dari *pattern* mengandung huruf "T" sehingga tidak cocok dengan urutan yang diujikan

T	T	A	A	T	T	A	C	C	T
	X			**X**					
	A	A	T	T					

Pergeseran = 1, karena akhiran dari *pattern* mengandung huruf "T" sehingga tidak cocok dengan urutan yang diujikan

T	T	A	A	T	T	A	C	C	T
		X	**X**	**X**	**X**				
		A	A	T	T				

Pergeseran = 4, karena akhiran dari *pattern* mengandung huruf "T" dan awalan "A" dan cocok dengan urutan yang diujikan, sehingga pencocokan dilanjutkan hingga karakter terakhir.

Jumlah perbandingan =6

```
Misalkan string yang ingin di uji adalah: text[i..i + n-1]
Proses perbandingan antara: text[i+j] dan pattern[j]
Di mana 0 < j < n,

Maka:
Text[i+j+1...i+n-1] = pattern[j+1..n-1]   dan
a=texttext[i+j] ≠ b=pattern[j]
Jika u = akhiran dari pattern sebelum b
     V = awalan dari pattern

Hingga,
text[i+j+1..i+n-1]=pattern[j+1..n-1]
```

Gambar 3.5. Pseudocode algoritma Boyer Moore

Melalui *pseudocode* pada Gambar 3.5. dapat dijelaskan algoritma pencocokan string Boyer Moore didasarkan pada 2 teknik, yaitu:

1. *Looking-glass technique*

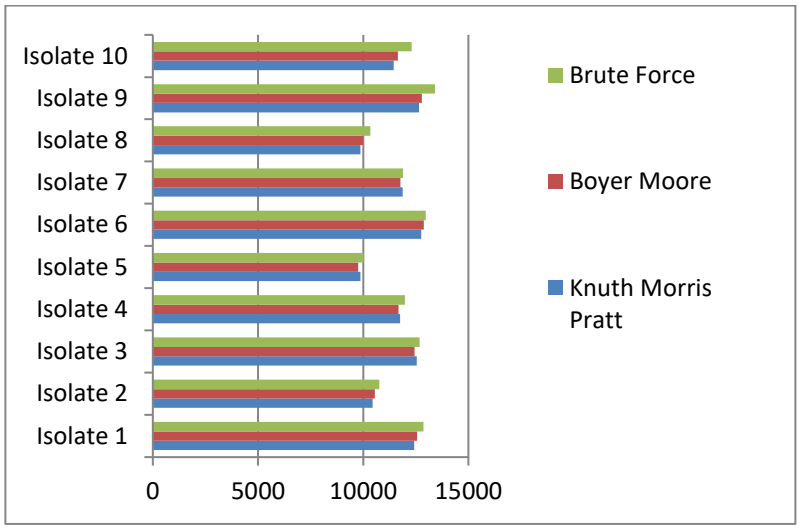
   Teknik ini merupakan cara untuk menemukan sebuah *pattern* dalam teks yang dibandingkan dengan memulai pencocokan dari akhir *string pattern.*

2. *Character-jump technique*

   Saat terjadi ketidakcocokan, pencarian akan dilanjutkan setelah menggeser pattern sebesar nilai tertentu untuk menghindari pencocokan yang sia-sia.

Tabel 3.2. Analisa perbandingan akurasi algoritma *pattern matching*

No	Algoritma	Jumlah nukleotida yang dibandingkan	Jumlah pergeseran dibandingkan dengan jumlah seluruh nukleotida	Akurasi
1	Knuth-morris Pratt	$10\ x10^5$	85%	100%
2	Booyer Moore	$10\ x10^5$	70%	100%
3	Brute Force	$10\ x10^5$	90%	100%



Gambar 3.6. Grafik analisa jumlah perbandingan ketiga metode pada sepuluh isolate yang diujikan

Melalui pengujian algoritma menggunakan data *isolated* DNA, dapat dianalisa bahwa ketiga metode memiliki nilai akurasi tinggi. Pada pengujian manual dengan jumlah urutan 10 karakter, metode Brute Force dan Boyer Moore menghasilkan jumlah pergeseran yang sama, namun pada pengujian data yang cukup besar metode Boyer Moore menghasilkan jumlah perbandingan yang paling kecil. Hal ini dikarenakan Boyer Moore melakukan pengecekan pada awal dan akhir dari *pattern* yang diujikan dan jika kedua karakter tersebut tidak cocok maka algoritma akan melanjutkan perbandingan. Sedangkan algoritma Brute Force akan terus melanjutkan pencocokan hingga karakter terakhir selama setiap karakter yang dibandingkan masih cocok, jika pada karakter terakhir ditemukan ketidak-cocokan maka hal ini sudah melakukan pencocokan yang percuma Tujuan dari uji coba ini adalah menemukan metode tercepat dalam menemukan *pattern* yang didasarkan dengan jumlah perbandingan yang kecil dan akurasi yang tinggi.

3.2 *Approximate String Matching*

Faktor mutasi merupakan salah satu faktor utama dari penelitian HCV, sedangkan permasalahan yang dihadapi pada proses *pattern matching* adalah metode tersebut tidak dapat mengenali adanya perubahan susunan DNA. Metode *pattern matching* akan mengembalikan nilai *null/false* ketika tidak ada *sequence* yang sama persis dengan primer yang dibandingkan, padahal ketika terjadi mutasi, satu atau lebih nukleotida dapat berubah.

Metode pencocokan *string* sering digunakan untuk mencari pola DNA. Namun, metode pencocokan *string* dasar tidak dapat mengenali kasus mutasi virus dan bakteri. Metode Hamming berbasis jarak dapat menerima ketidakcocokan karakter dalam pengaturan meskipun dapat memberikan hasil kinerja yang bervariasi tergantung pada jumlah pola yang dibandingkan. Oleh karena itu, diusulkan penggunaan metode *approximate string matching*, salah satu metodenya yaitu *Hamming distance.*

Hasil pengujian menggunakan Hamming menunjukkan bahwa metode tersebut mampu menoleransi adanya perbedaan karakter antara primer dan *sequence* pada *isolated* DNA. Namun, primer dengan jumlah karakter yang kecil akan cenderung menghasilkan jarak yang kecil pula, sebaliknya primer dengan jumlah karakter yang panjang akan menghasilkan jarak yang panjang pula. Sehingga primer dengan jumlah karakter yang pendek cenderung banyak menghasilkan *sequence isolated* DNA dengan nilai kemiripan tinggi, dengan arti lain cenderung tren. Padahal belum tentu primer dengan karakter pendek tersebut adalah primer yang saat ini sedang tren. Oleh karena itu diusulkan pengubahan metode Hamming dengan menambahkan fungsi normalisasi dengan hasil penghitungan jarak Hamming, dengan tujuan menyeimbangkan ketimpangan jumlah karakter antara primer yang satu dengan primer yang lainnya sehingga memberikan hasil perbandingan yang adil.

Metode pencocokan string, seperti Boyer Moore dan Knuth Morris Pratt, telah banyak digunakan untuk mengenali pola-pola dalam DNA dengan membandingkan kesamaan mereka (G. Kucherov; K. Salikhov ; D. Tsur, , 2014). Metode-metode

tersebut dapat menyatakan ketidaksamaan DNA hanya karena satu karakter yang berbeda (Guo, Hermelin, & Komusiewicz, 2014). Namun, metode pencocokan string konvensional (S. Cho; J.C. Na; K. Park; J.S. Sim, 2015), tidak dapat menghitung urutan DNA karena pola dengan karakter yang berbeda tidak selalu berarti mereka tidak cocok. Perbedaan karakter itu terjadi karena urutan DNA mengalami mutasi seiring waktu (Y. Chen; Y. Hu, 2006) (R. Beal, D. Adjeroh, 2015). Penular infeksi dan multi-resistansi agen seperti virus dan bakteri pembawa penyakit memiliki kemampuan beradaptasi terhadap lingkungan. Virus dan bakteri dapat memiliki perubahan pengaturan DNA dan pola RNA (T. Boby, A.M. Patch, S.J. Aves, 2005) (Pray, 2008). Fokus dari pencocokan *string* berdasarkan skor dalam penelitian ini adalah metode Hamming yang dapat menerima ketidakcocokan karakter dalam suatu pengaturan (C.S. Rao ; S.V. Raju, 2016) (S. Das; K.Kapoor, 2017) untuk mengatasi masalah mutasi pada DNA.

Beberapa karakter tidak dapat diubah, dihapus atau ditambahkan dalam kasus urutan mutasi DNA. Ada beberapa urutan utama dalam data sampel yang terisolasi yang positif terinfeksi HCV. Oleh karena itu metode Hamming cocok untuk menangani pencocokan string sekuens *isolated* DNA dengan berbagai pola primer. Seperti disebutkan sebelumnya, metode Hamming yang dimodifikasi dengan untuk melakukan analisis jarak pengaturan nukleotida dalam DNA yang memiliki infeksi HCV primer. Beberapa metode untuk modifikasi dipilih karena pekerjaan Penulis sebelumnya (B.A. Kindhi, T.A. Sardjono, 2015) dengan Brute Force (BF), Boyer-Moore (BM) dan Knuth-Morris-Pratt (KMP) tidak dapat mengenali adanya mutasi pada DNA.

Seperti disebutkan sebelumnya, modifikasi metode Hamming diusulkan untuk melakukan analisis jarak susunan nukleotida dalam DNA yang memiliki infeksi HCV. Beberapa metode *pattern matching* telah diuji coba untuk menemukan pola primer di dalam *isolated* DNA dengan Brute Force (BF), Boyer-Moore (BM) dan Knuth-Morris-Pratt (KMP) (B.A. Kindhi, T.A. Sardjono, 2015). BM menunjukkan hasil yang lebih baik karena memiliki jumlah perbandingan yang lebih sedikit dibandingkan dengan

metode lain. Namun, urutan primer tidak harus memiliki susunan urutan yang sama karena virus dapat memiliki siklus mutasi yang mengakibatkan beberapa nukleotida berubah. Metode-metode tersebut hanya mengenali urutan dengan kesamaan yang pasti dan tidak dapat menangani mutasi pada urutan primer. Dengan kata lain, metode *pattern matching* hanya akan menyatakan positif pada suatu *isolated* DNA jika pada saat pencocokan ditemukan urutan yang sama perseis dengan primer yang dibandingkan. Masalahnya adalah susunan nukleotida pada *isolated DNA* dapat saja berbeda, berarti bahwa tidak dapat menentukan urutan primer yang tepat. Oleh karena itu diusulkan perubahan metode dari *pattern matching* menjadi *approximate string matching* untuk menangani masalah pada tahap *semantic similarity*. Berikut adalah contoh penghitungan jarak Hamming:

$$h(p,s) = \sum_{a=0}^{20} \begin{Bmatrix} min_{(x,y)} = 0 \rightarrow jika\ a_p \neq b_s \\ min_{(x,y)} = 1 \rightarrow jika\ a_p\ = b_s \end{Bmatrix}$$

=0+1+1+1+1+1+1+0+0+1+0+0+1+1+1+1+1+1+0+1

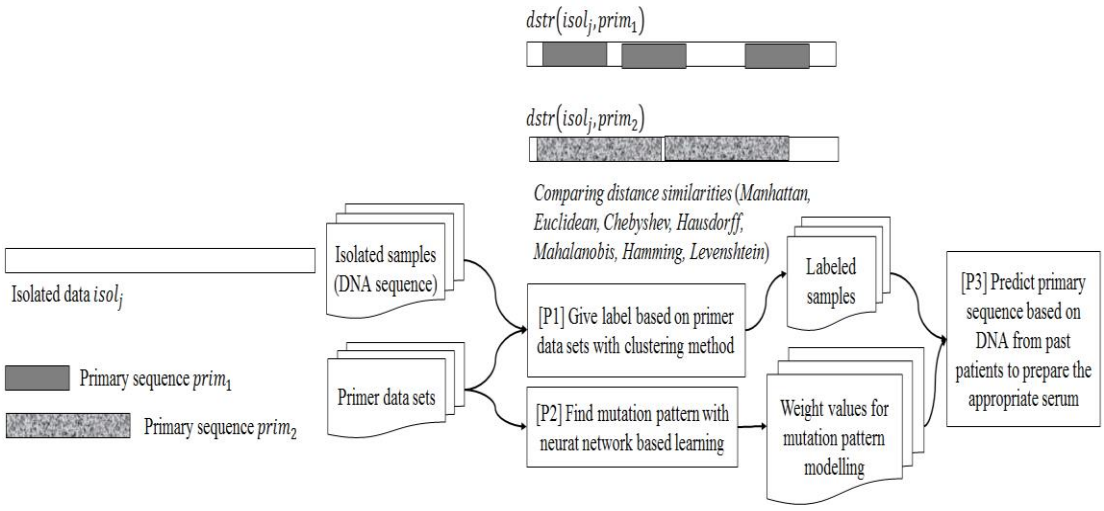
=14

Isolated DNA access code : KJ439774.1,

Tahun 2014,

Indeks ke: 3566

P	T	T	T	G	A	C	T	C	A	A	C	C	G	T	C	A	C	T	G	A
S	T	A	G	T	C	A	C	C	A	G	C	C	T	C	A	C	T	G	G	C
H	0	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1	1	1	0	1
Total jarak Hamming									:	14										



Gambar 3.7. Pendekatan metode yang diusulkan untuk membandingkan sampel terisolasi dengan urutan *primer*

Metode Hamming sering digunakan untuk masalah pencocokan *string* dengan menemukan persamaan dari jumlah karakter dan pengaturan maksimum jarak(C. Sammut; G.I. Webb, 2010). Metode Hamming juga diterapkan untuk menemukan pola dalam pengenalan sidik jari (S. Kundu and B. Ray, 2015), pengecekan bit paritas dan enkripsi data (J. Park, I. Kim, H. Y. Song, 2017) (Singh, 2016) (W. Shan, S. Zhang and Y. He, 2017) (C. Chen and R. Veldhuis, 2009) (S. Kim, H. Cho, 2017) (S. Pissis ,A. Retha, 2015).

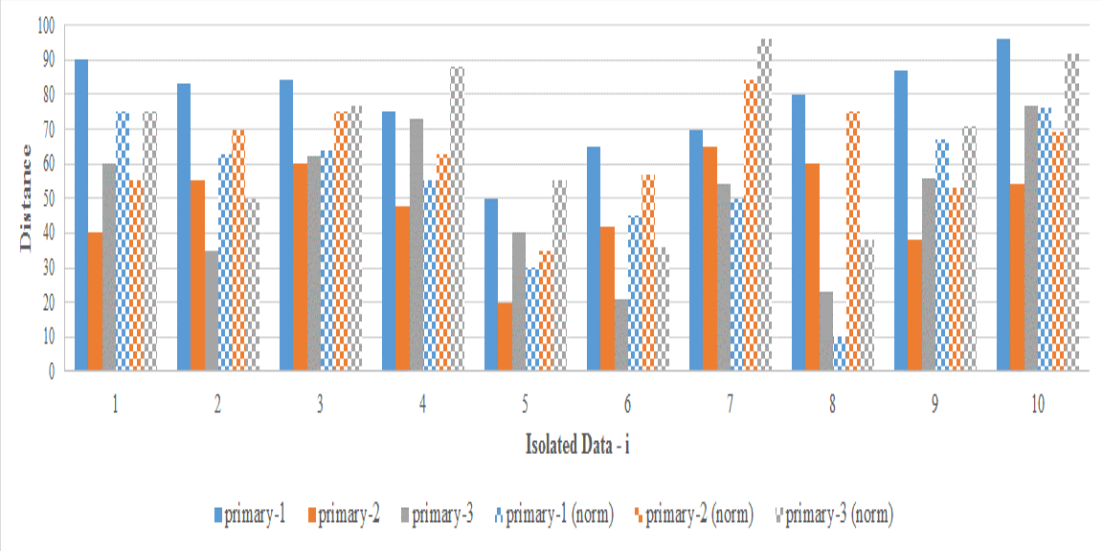
Metode hamming biasa digunakan untuk mendeteksi kesalahan atau kesalahan bit dalam mengirim data. Tetapi karena cara kerjanya sederhana, metode ini juga dapat diterapkan untuk menganalisis kemiripan string (Kondo, 2014). Ini adalah salah satu alasan Penulis memilih metode hamming, karena urutan nukleotida tidak dapat ditingkatkan atau dikurangi jumlahnya (A.S. Pinheiro, H.P. Pinheiro, P.K. Sen, 2012), namun dapat diamati pola dan jaraknya (H.P. Pinheiro,A.S. Pinheiro , P.K. Sen, 2005) (A. Apostolicoa, C. Guerraa , G.M. Landauc , C. Pizzie, 2016).

Namun, metode Hamming cenderung memberikan nilai skor kesamaan yang lebih tinggi ketika urutan primer memiliki karakter yang lebih sedikit. Dalam masalah

jarak untuk sekuens DNA, urutan pada *isolated DNA* sering diklasifikasikan ke dalam primer yang memiliki jumlah karakter terkecil. Salah satu solusinya adalah dengan menambahkan nilai normalisasi untuk menghindari masalah tersebut ke dalam perhitungan pencocokan pola dengan metode Hamming. Nilai normalisasi di sini digunakan untuk menyeimbangkan urutan primer dengan panjang yang berbeda.

Metode Hamming termodifikasi digunakan untuk menganalisa jarak kemiripan masing-masing primer terhadapa masing-masing urutan di dalam *isolated* DNA. Primer infeksi HCV diperoleh dari Institute of Tropical Disease, Universitas Airlangga, Indonesia sejumlah 10 primer, $P = \{prim_1 \dots prim_{10}\}$. Sampel terisolasi $isol_j$ dari World Gen Bank dibandingkan dengan data primer $prim_i$. Jadi pada proses perbandingan, terdapat kemungkinan dalam satu *isolated* DNA terdapat lebih dari satu urutan yang mirip dengan primer. Sebaliknya, dalam satu *isolated* DNA juga terdapat kemungkinan bahwa urutan-urutan tersebut memiliki nilai kemiripan pada lebih dari satu primer.

Karena urutan primer memiliki urutan panjang bervariasi antara 15 hingga 50 nukleotida, maka pada *pre-processing*, data akan dipotong sepanjang jumlah karakter masing-masing primer yang akan dibandingkan. Untuk skor Hamming dalam Persamaan 3.2, bertujuan untuk menghitung jumlah karakter yang serupa dalam urutan antara primer dan terisolasi untuk setiap pemotongan DNA. Nilai modifikasi (Persamaan 3.3) diperoleh dengan menambahkan langkah normalisasi $\boldsymbol{dstr}(\boldsymbol{isol_j}, \boldsymbol{prim_i})$ membandingkan data sampel terisolasi tertentu dengan semua sekuens HCV primer yang ada. Fungsi $\boldsymbol{len}(\boldsymbol{prim_i})$ didefinisikan sebagai jumlah karakter dalam urutan $\boldsymbol{prim_i}$. Nilai normalisasi ditambahkan untuk menyeimbangkan masalah panjang urutan. Yaitu hasil analisa penghitungan jarak pada metode Hamming



Gambar 3.8. Perbandingan hasil eksperimen Penulis untuk mengukur jarak dengan metode Hamming dan metode Hamming dengan nilai normalisasi

ditambahkan dengan nilai normalisasi. Nilai normalisasi tersebut adalah dengan mengurangi jumlah karakter suatu primer dengan jumlah karakter primer terpendek, kemudian membaginya dengan jumlah karakter primer terpanjang dikurangi jumlah karakter primer terpendek.

Metode penghitungan jarak kemiripan tersebut dapat diamati pada Persamaan 3.1. $dist(isol_j, prim_i)$ dimana $isol_j$ adalah bentuk *string* dari data sampel yang terisolasi dan $prim_i$ adalah bentuk *string* dari urutan primer, yaitu urutan primer HCV. Fungsi dari $dist(isol_j, prim_i)$ adalah untuk mencari jumlah perbedaan karakter dari urutan HCV primer dalam suatu urutan data sampel yang terisolasi. Nilai dari $max(isol_j)$ (Persamaan 3.4) berkaitan dengan data sampel terisolasi $isol_j$ dengan urutan primer yang paling mirip. Nilai dari $min(isol_j)$ (Persamaan 3.5) adalah persamaan untuk menghitung data sampel terisolasi $isol_j$ dengan urutan primer yang tidak sama. Urutan primer didefinisikan dengan $P = \{prim_1 \dots prim_{10}\}$ dan di sini kita memiliki sepuluh urutan primer. Contohnya, hasil dari proses $infected(isol_j, prim_i)$ (persamaan 3.6) dengan hasil {0,0,0,0,0,0,1,0,0,0} berarti bahwa data *isolated* $isol_j$

ditemukan mutasi pada primer yang ke tujuh $prim_7$. Nilai *treshold* dibutuhkan untuk menentukan ambang batas urutan DNA yang positif (M.J. Atallah, T.W. Duket, 2011). Hasil pencocokan urutan dianggap positif jika nilainya melebihi nilai *thres*, *thres* adalah nilai ambang 70 persen dari seratus persen (Simmonds.et.al, 2005).

$$dist(isol_j, prim_i) = hamm(isol_j, prim_i) + norm(prim_i) \quad (3.1)$$

$$hamm(isol_j, prim_i) = \frac{p\_match}{\#\ primary\ characters}\ x\ 100\% \quad (3.2)$$

$$norm(prim_i) = \frac{len(prim_i) - min(P)}{max(P) - min(P)} \quad (3.3)$$

$$max(P) = \max_{prim_k \in P} len(prim_k) \quad (3.4)$$

$$min(P) = \min_{prim_k \in P} len(prim_k) \quad (3.5)$$

$$infected(isol_j, prim_i)_{prim_i \in P} = \begin{cases} 0 & dstr(isol_j, prim_i) < thres \\ 1 & dstr(isol_j, prim_i) \geq thres \end{cases} \quad (3.6)$$

Dengan melakukan normalisasi pada metode Hamming, hasil dari penghitungan jarak akan menjadi sejajar atau seimbang. Normalisasi metode Hamming bertujuan untuk mengatasi ketimpangan pada jumlah karakter primer yang berbeda-beda sehingga menghasilkan analisa jarak yang efektif pula.

Untuk memvalidasi metode yang diusulkan, maka dibandingkan 100 data sampel terisolasi yang positif terinfeksi HCV dengan jenis kelamin dan rentang usia yang

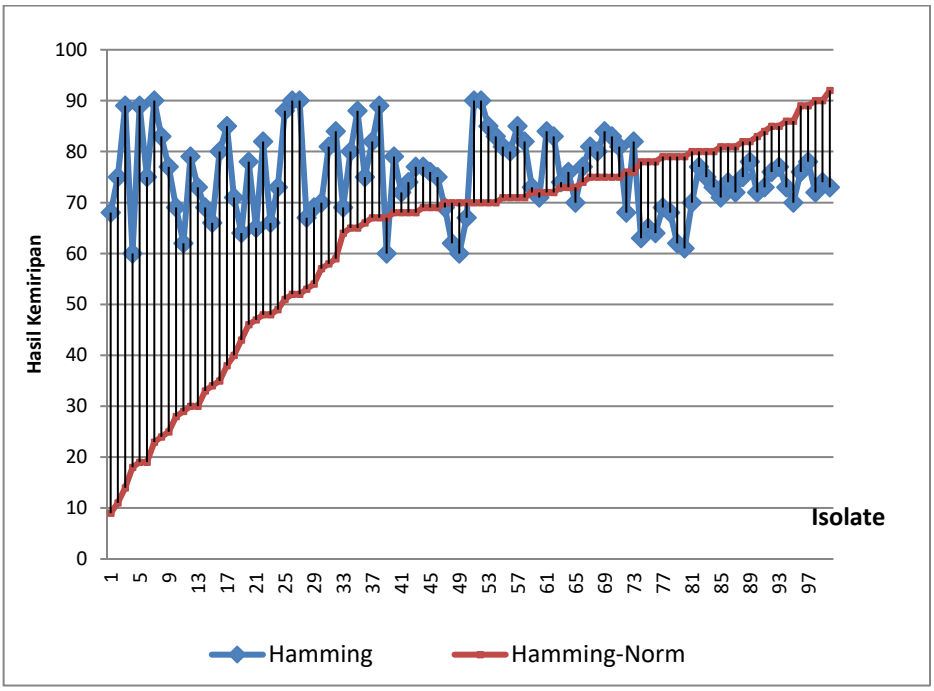
berbeda. Primary Sequence-1 (PS-1) dicatat dari tahun 1992, sementara Primary Sequence-2 (PS-2) adalah 1999 dan Primary Sequence-3 (PS-3) adalah 2014. Sampel terisolasi yang positif terinfeksi HCV dapat terdiri dari dari 9.000 hingga 15.000 sekuens DNA.

Tabel 3.3. menjelaskan bagaimana cara melakukan uji coba dengan membandingkan beberapa metode dalam penelitian sebelumnya. Dari salah satu hasil pengujian, seperti yang ditunjukkan pada Tabel 3.3., tiga metode BF, KMP, dan BM memberikan hasil yang "tidak ditemukan". Karena ketiga metode tersebut hanya mampu mendeteksi pola dengan kesamaan seratus persen. Selanjutnya diujicoba menggunakan jarak hamming dengan hasil yang kurang sesuai dengan harapan. Hasil pengujian hamming distance hampir selalu memberikan nilai tinggi pada primer yang memiliki sedikit karakter. Kemudian Penulis mencoba menambahkan jarak hamming dengan proses normalisasi, dengan hasil yang hampir mendekati kondisi nyata HCV saat ini.

Setelah menambahkan nilai normalisasi dalam skor Hamming, skor pencocokan pola tren dari data sampel yang terisolasi dari PS1 menurun sementara PS3 meningkat. Perbandingan dengan metode Hamming yang dimodifikasi membuat hasil yang bervariasi yang berarti bahwa PS1 tidak selalu memiliki nilai pencocokan pola tertinggi dengan data sampel yang terisolasi dibandingkan dengan metode Hamming biasa. Hal ini menunjukkan bahwa nilai normalisasi dapat menyeimbangkan urutan primer dengan panjang yang berbeda.

Tabel 3.3. Contoh perbandingan antara data sampel yang terisolasi dan Urutan Primer

Isolate-i	**Metode Perbandingan**	**Pattern Matching Skor**
1	Brute-Force	Tidak ditemukan
	Knuth-Morris-Pratt	Tidak ditemukan
	Boyer-Moore	Tidak ditemukan
	Hamming (A)	**90**
	Hamming-Norm (B)	**75**



Gambar 3.9. Gap hasil antara metode Hamming dan Metode Hamming dengan normalisasi

Data terisolasi dari {1, 3, 4, 5, 7, 9, 10} memiliki kecenderungan untuk mirip dengan PS3 karena data tersebut dicatat dalam tiga tahun terakhir (2014-2017) sementara PS3 diterbitkan pada tahun 2014. Dengan demikian data sampel terisolasi terbaru memiliki kecocokan yang sesuai dengan urutan primer termutasi saat ini. Hasil analisa tersebut ditunjukkan dalam kasus pencocokan pola dengan PS3.

Nilai *treshold* atau skor minimum urutan yang dinyatakan positif pada Persamaan 3.6. adalah $thres > 70\%$ (Simmonds. et. al, 2005). Ini berarti bahwa kemiripan nukleotida harus lebih dari 70% karena semua data sampel yang terisolasi positif terinfeksi HCV. Namun percobaan menunjukkan bahwa tidak semua data *isolated* DNA memberikan skor pencocokan pola yang lebih tinggi dari ambang ke urutan primer yang dibandingkan. Percobaan dengan metode Hamming untuk data terisolasi {1, 2, 3, 8, 9, 10} memiliki skor pencocokan pola yang lebih tinggi dari ambang ketika dibandingkan dengan PS1. Hasil itu adalah bukti PS1 yang disebutkan

di atas memiliki nilai pencocokan pola tertinggi dengan data sampel *isolated* DNA. Sedangkan bukti kemiringan yang cocok dari PS3 dengan data sampel terisolasi terbaru ditunjukkan dalam data terisolasi dari {4, 7, 10}.

Terdaapat kebutuhan untuk membandingkan dengan lebih banyak urutan primer karena eksperimen saat ini memiliki skor pencocokan pola kurang dari ambang batas meskipun data sampel yang terisolasi positif terinfeksi HCV. Pengujian kesamaan DNA menunjukkan tidak ada urutan pasti yang memiliki infeksi HCV primer karena bakteri dan virus bermutasi dan mengubah pengaturan pola urutan DNA. Skor kesamaan antara urutan primer dan urutan DNA terinfeksi positif tergantung pada jumlah perbedaan karakter hasil perbandingan.

Gambar 3.6. menggambarkan grafik hasil hasil kemiripan antara Hamming dan Hamming dengan normalisasi pada pengujian 3 primer (primer diterbitkan pada 2014). Hasil pengujian metode Hamming cenderung di atas tingkat 60% meskipun data uji yang dibandingkan adalah *isolatedd* DNA 18 tahun yang lalu. Seharusnya *isolated* DNA yang lebih dari 15 tahun jauh dari primer tidak memiliki nilai kesamaan yang tinggi, karena pada saat itu primer telah melakukan mutasi berkali-kali. Berbeda pada metode hamming-normalisasi memberikan hasil yang sama berbanding lurus dengan jarak tahun pertama.

3.3 Optimasi *Semantic Similarity*

Kecepatan dan ketepatan merupakan indikator dalam pemilihan metode yang tepat pada tahap *semantic similarity.* Semakin besar set data maka waktu yang dibutuhkan untuk melakukan pencarian pola juga semakin lama, oleh karena itu dibutuhkan metode yang mampu menemukan pola seccepatnya dengan jumlah perbandingan paling minimum.

Metode *Levenshtein Distance,* mampu menghitung jarak yang minimum antara satu pola dengan pola lainnya. Keunggulan metode ini dibandingkan dengan metode

Hamming yang sebelumnya digunakan adalah metode *Edit Levenshtein* ini memiliki tiga kemungkinan proses ketika menemui karakter yang berbeda, yaitu *insert, delete,* dan *subtitusi.*

Metode *Levensthein distance* ini selanjutnya akan selalu digunakan sebagai dasar semantic similarity pada langkah penelitian berikutnya. Beberapa contoh yang dipilih secara acak dari data urutan DNA yang telah melalui analisis kesamaan disajikan pada Tabel 3.4. Kolom pertama berisi informasi nomor aksesi DNA yang terisolasi dalam sistem bank gen dunia. Pola primer adalah primer HCV yang digunakan untuk referensi ketika menghitung kesamaan dan sebagai centroid dalam tahap pengelompokan. Berikut adalah contoh penghitungan *Edit Levenshtein Distance* pada data DNA:

$$lev_{a,b}\,(i,j) = \begin{cases} \max(i,j) \rightarrow jika\ min_{(i,j)} = 0 \\ \min \begin{cases} lev_{a,b}\,(i-1,j)+1 \\ lev_{a,b}\,(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_i)} \end{cases} \end{cases}$$

(3.7)

Isolated DNA access code: KC197231.1,

Tahun 2013,

Indeks Ke: 245

P	G	T	C	G	C	G	A	A	A	G	G	C	C	T	T	G	T	G	G	T	A	C	T	G	C	C	T	G	A	T
S	G	T	T	G	C	G	A	A	A	G	G	C	C	T	T	G	T	G	G	T	A	C	T	G	C	C	T	G	A	T
	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Total Jarak Edit Levenshtein Distance: subtitusi “T” dengan “C”, yaitu 1

Tabel 3.4. Sampel acak dari hasil kesamaan dari proses mengiris masing-masing isolat DNA berdasarkan pada masing-masing pola primer

Akses File dan tahun	**Primer Pattern**	**Pattern pada *isolated* DNA**	**Indeks posisi**	**Kemiripan (%)**	**Jarak tahun**	**Jarak**
EU255978.1-2014	TTTGACTCAACCGTCACTGA	TTTGACTCGTCTGTCCTCTG	4924	70	0	6
EU155241.1-2009	GGTGAGTGATTGGAGGGTTG	CCCTTGTGTCATCGGAGGGG	1991	55	10	9
HQ850280.1-2016	TCCCGGGGCACTCGCAAGCACCCTATCAGG	TGCCTCCGCCTCGGAAGAAGCGTACAGT	7400	57	24	13
KJ439773.1-2014	TGGGGATCCCGTATGATACCCGCTGCTTTGA	GGGRTTCTCATATGCACCCGCTGCTTTGA	8479	72	0	9
KJ439774.1-2014	TTTGACTCAACCGTCACTGA	TAGTCACCAGCCTCACTGGC	3566	60	0	8
KJ439774.1-2014	ACTGTCACTGAACAGGACAT	ACAGTTACTGAGAGTGACAT	8515	75	0	5
KF676351.1-2014	CTCAGGTTCCGCTCGTCCTC	CTCAGCCTACTCCTACTTTCC	9653	57	0	9

Akses File dan tahun	Primer Pattern	Pattern pada *isolated* DNA	Indeks posisi	Kemiripan (%)	Jarak tahun	Jarak
KF035125.1-2013	ACTGTCACTGAACAGGACAT	AGCCACTGAACCTGTGATAT	3358	65	1	7
KC197231.1-2013	GTCGCGAAAGGCCTTGTGGTACTGCCTGAT	GTTGCGAAAGGCCTTGTGGTACTGCCTGAT	245	97	21	1

Urutan yang dibandingkan dalam pola *isolated* DNA adalah urutan yang dipotong dari *isolated* DNA untuk dibandingkan dengan primer. Posisi indeks adalah posisi urutan yang dibandingkan dalam *isolated* DNA. Kesamaan *string* dihitung berdasarkan jumlah karakter identik yang ada. Jarak tahun adalah jarak antara tahun primer dan tahun DNA terisolasi, dan jarak pencocokan pola adalah hasil dari perhitungan jarak berdasarkan Persamaan (3.7), yaitu metode *Edit Levenshtein Distance.*

3.4 Sistem Pakar DNA Analisis

Saat ini, setiap rumah sakit di Indonesia memiliki basis penelitian sendiri. Data masing-masing rumah sakit berbeda dan tidak terintegrasi. Pengolahan data yang bagus akan lebih baik jika terintegrasi secara *online* dan dapat dimanfaatkan secara luas. Tenaga medis dapat secara langsung bertanggung jawab menggunakannya sebagai pendamping penelitian. Pemerintah menggunakan sistem sebagai alat untuk pengambilan keputusan kebijakan kesehatan. Oleh karena itu diperlukan sistem matang yang dapat memenuhi kebutuhan tersebut.

Pada bab ini dilakuakn analisa dan perbandingan kinerja dan hasil pengolahan sistem pakar DNA yang ada. Hasil penelitian ini akan menganalisis fitur dan platform apa yang terbaik untuk sistem pakar DNA untuk mengakomodasi kebutuhan analitis staf medis di masa depan. Prototipe yang akan dibangun dalam penelitian ini adalah desain infrastruktur yang telah dianalisis sesuai dengan kebutuhan Sistem Pakar DNA pemangku kepentingan di Indonesia.

Diusulkan penambangan DNA cloud yang cerdas, aman, dan terstruktur. *Cloud Expert System DNA Analysis* (CESDA) yang Penulis rancang memiliki spesifikasi database tiga tingkat. Untuk mempercepat fungsi analisis DNA semantik, Penulis juga mengusulkan perbedaan antara OLTP dan OLAP. Selain itu, penambangan data besar dan jumlah pengguna yang mengaksesnya pada saat yang sama membutuhkan paralel server berbasis cloud.

Survei dilakukan untuk mempelajari dan menganalisis beberapa aplikasi yang digunakan oleh beberapa rumah sakit di Indonesia saat ini. Cara kerja masing-masing aplikasi dan pemanfaatannya oleh tenaga medis. Dari hasil penelitian ada 4 aplikasi yang paling sering digunakan, empat aplikasi tersebut adalah:

A. *DNA BLAST* (World-Gene-Bank, accessed May, 2017)

DNA BLAST (Alat Pencarian Alignment Lokal Dasar) adalah aplikasi analisis DNA online yang dikelola oleh Bank Dunia. Setiap penemuan baru *isolated* DNA dari suatu penyakit, para peneliti harus melaporkan *isolated* DNA tersebut sebagai paten pada Gen Bank Dunia, ini adalah salah satu keuntungan dari aplikasi DNA Blast. Semakin banyak yang mendaftarkan paten di World Bank Genes, semakin lengkap data sampel pada DNA Blast, sehingga aplikasi ini sering digunakan oleh peneliti sebagai referensi urutan pencocokan suatu isolat dan primer. Namun, DNA Blast hanya dapat mencari kecocokan dari suatu pola, sementara apa yang peneliti perlukan bisa beragam seperti menemukan turunan dari *isolated* DNA tersebut.

B. *Clustal* (University College Dublin, Multiple Sequence Alignment, www.clustal.org)

Clustal adalah aplikasi pengurutan DNA yang merupakan penyelarasan urutan berganda. Urutan yang diimpor akan diproses menjadi pohon DNA. Fungsi pohon ini adalah untuk menganalisis keturunan atau kedekatan urutan dari primer. File yang diimpor dapat diperoleh dari aplikasi MEGA.

C. *MEGA* (Molecular-Evolutionary-Genetics-Analysis, accessed May 2017)

Merupakan perpanjangan dari *Molecular Evolutionary Genetics Analysis*, yaitu analisis genetika evolusi molekuler. Fitur MEGA termasuk Konstruksi Penjajaran Urutan, Penanganan Data, dan Bagian Tabel Kode Genetik.

D. *Genetyx* (www.genetyx.co.jp, accessed April 2017)

Genetyx Merupakan perangkat lunak analisis DNA berbayar yang memiliki fitur lengkap saat ini. MEGA tidak dapat membuat FASTA dan harus diekspor ke Clustalw. Clustalw mampu membuat pohon penyelarasan urutan berdasarkan sekelompok kedekatan. Sementara Genetyx mampu melakukan keduanya.

3.5 Usulan rancangan sistem pakar analisis DNA

Hasil eksperimen dari beberapa analisis sistem pakar DNA dapat disimpulkan bahwa beberapa perangkat lunak *open source* tidak dapat memproses *isolated* DNA menjadi FASTA, sehingga untuk membuat analisis genom harus diimpor ke perangkat lunak lain yang baru diproses. Sistem Pakar Analisis DNA berbasis Cloud SaaS sudah ada, tetapi fitur yang disajikan tidak seperti sistem pakar dengan aplikasi berbayar yang di install secara lokal. Oleh karena itu dibutuhkan SaaS Cloud-Based Expert System yang dapat mengakomodasi semua fitur yang dibutuhkan dalam pekerjaan analisis DNA. Ketika semua fitur dibuat secara online, spesifikasi perangkat keras yang diperlukan juga harus diselaraskan dengan persyaratan perangkat lunak. Oleh karena itu dibutuhkan Server Data yang dapat terus tumbuh seiring dengan meningkatnya

isolated DNA yang terdaftar ke dalam basis data, dan di mana ketika basis data sedang melakukan pemeliharaan, aplikasi akan tetap dapat menyediakan layanan selama permintaan pengguna alih-alih mengakses basis data. Untuk mengakses data DNA yang bisa mencapai jutaan, tentu saja membutuhkan waktu lama dalam proses analisis. Proses analisis dapat dilakukan dengan cepat jika ada beberapa PC yang memproses perangkat lunak secara bersamaan, sehingga hasil analisis dapat diketahui dengan cepat.

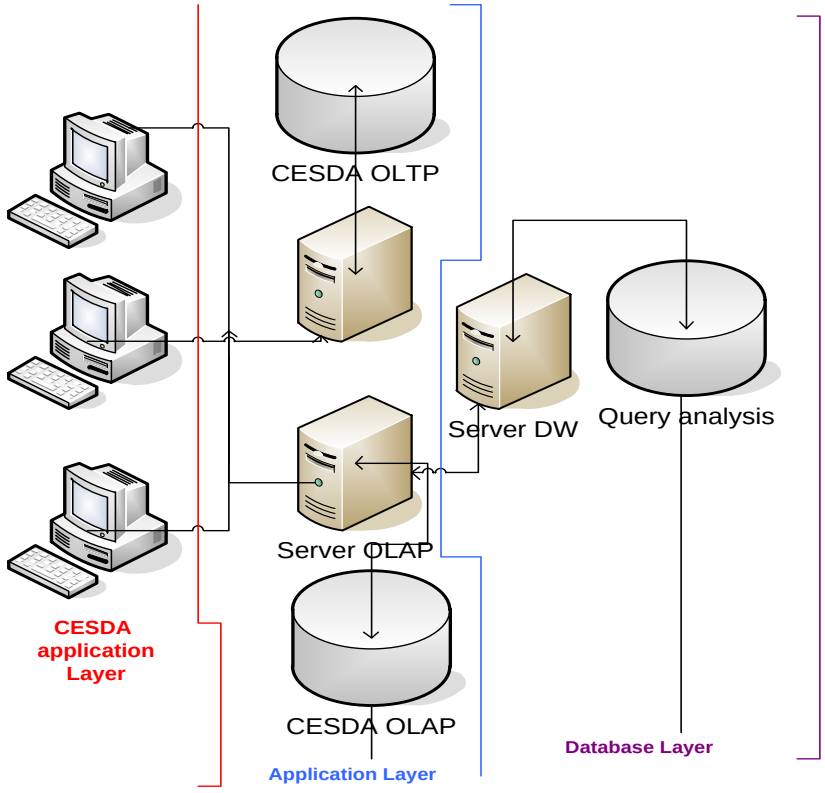
Sistem pakar yang telah di rancang mampu menampung analisis pergerakan perubahan nukleotida yang menyebabkan terjadinya mutasi genetik, mencari jam molekuler dan turunan virus. Analisis ini adalah kondisi biomolekuler yang terjadi saat ini atau di masa lalu. Beberapa studi di bidang rekayasa biomolekuler telah menerbitkan metode memprediksi urutan-urutan DNA (K. Kim , M. Kim ,Y.Wooc, 2008) (Brijesh K. Sriwastava, Subhadip Basu, and Ujjwal Maulik, 2015). Studi tersebut juga menyertakan analisa bahwa dibutuhkan dukungan pengolah data yang memadai untuk melakukan analisis data dalam jumlah besar. Melalui analisis kerentanan, studi mengenai beberapa aplikasi sistem pakar DNA yang saat ini digunakan, maka dapat dirancang suatu sistem yang mampu melengkapi dan menjembatani kebutuhan ahli dalam melakukan analisa data DNA. Sistem yang penulis rancacng memiliki kriteria sebagai berikut:

A. OLAP dan OLTP

Analisis DNA sistem pakar di masa depan, tidak hanya mampu melakukan OLAP (*Online Analytical Processing*), tetapi juga mampu melakukan OLTP (*Online Transaction Processing*), sehingga ketika ada pengguna yang memasukkan data baik data primer maupun *isolated* DNA, sistem dapat dengan cepat melakukan pembaruan analisis transaksi dan memperbarui data di OLAP. Transaksi dari OLTP tersebut antara lain pemasukan, perubahan, penghapusan, maupun pengaturan *trigger database* oleh admin, yang memungkinkan

dilakukan oleh banyak pengguna dalam waktu yang bersamaan, sehingga OLTP juga berperan mengatur *session* pada masing-masing pengguna yang sedang bertransaksi.

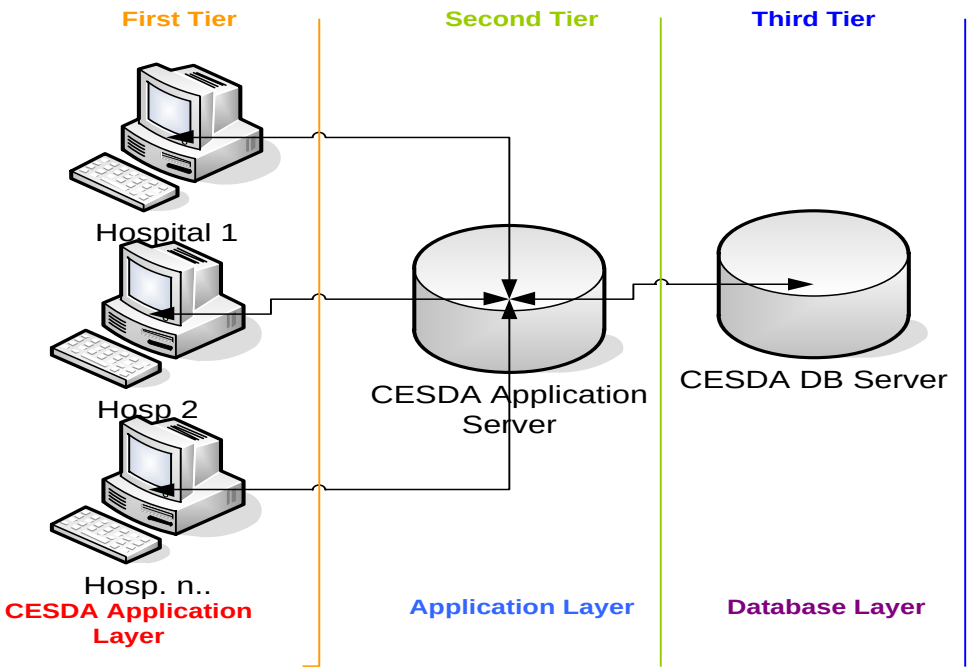
Pada system yang diusulkan OLAP dan OLTP di lakukan pada server yang berbeda, hal ini bertujuan untuk menghindari adanya *deadlock* ketika sistem sedang melayani transaksi *input-output* data, dan disisi lain user lain melakukan permintaan analisa data yang cukup besar.



Gambar 3.10. Rancangan system OLAP dan OLTP yang diusulkan

B. *Database Three-Tier*

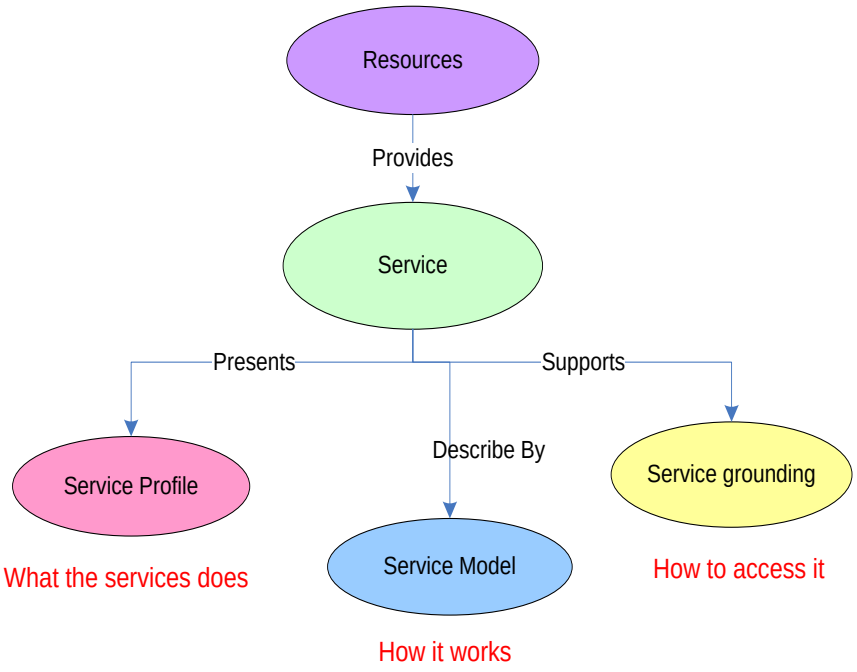
Model desain *database client-server* yang diusulkan adalah Basis Data Tiga Tingkat. Seperti yang ditunjukkan pada Gambar 3.8., *database* dan aplikasi dimasukkan ke dalam server terpisah. Sehingga ketika basis data berkembang pesat dan membutuhkan ruang baru, aplikasi tidak akan terganggu karena berada di tempat yang terpisah. Sebaliknya, ketika aplikasi membutuhkan perbaikan modul, basis data tidak akan terganggu dan proses *backup* data tetap dapat berjalan.



Gambar 3.11. Rancangan *Three Tier Client Server* yang diusulkan

C. *Cloud SaaS*

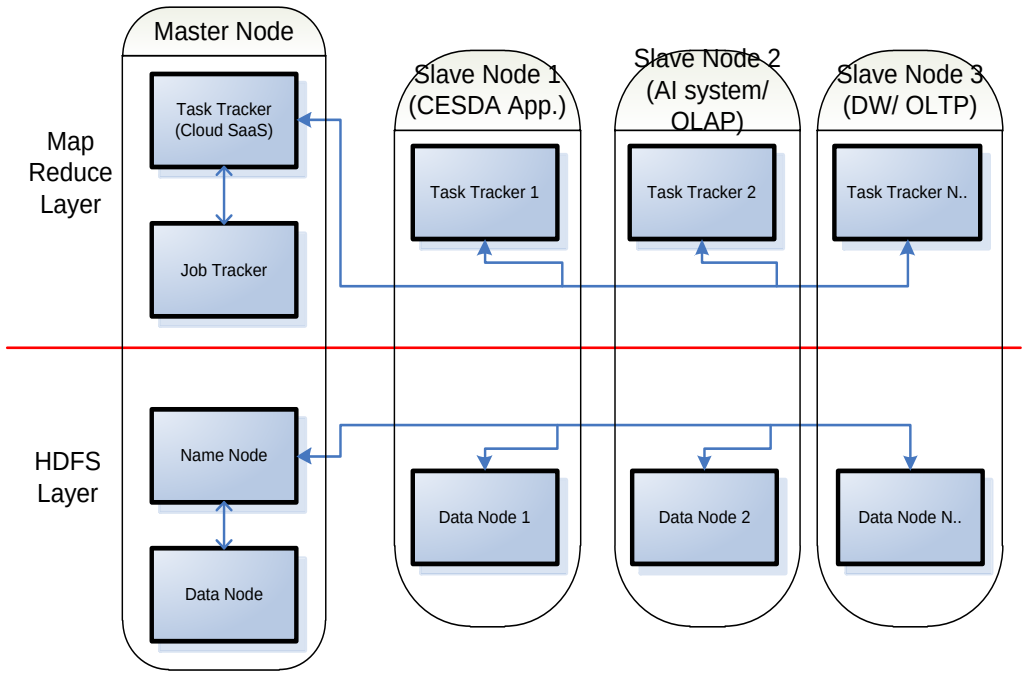
Sistem pakar yang dibangun kemudian harus dalam bentuk *Cloud* perangkat lunak sebagai layanan, yaitu *Cloud SaaS (Software as a Service)*. Pemilihan model *Cloud SaaS* karena aplikasi dapat terus dikembangkan tanpa harus dalam pemeliharaan (Stollberg Michael and Armin Haller, 2005) (J. Kopecký, T. Vitvar, C. Bournez and J. Farrell, 2007). *Cloud SaaS* mampu memberikan layanan kepada ratusan pengguna dengan berbagai layanan dan data. *Cloud SaaS* dianggap aman karena pengguna tidak secara langsung mengakses aplikasi tetapi melalui WSDL (*Web Service Definition Language*) yang kemudian diteruskan ke Server Aplikasi. Dalam *Cloud Computing*, layanan disediakan dalam bentuk profil layanan, model layanan, dan landasan layanan seperti yang dapat diamati pada Gambar 3.9.



Gambar 3.12. SaaS Web Service (D. Martin, M. Burstein, D. McDermott, S. Mc Ilraith, M. Paolucci, K.Sycara, D. L. Mc Guinness, E. Sirin and N. Srinivasan., 2007)

D. *Pararel Computing*

Selain *SaaS Cloud* dan *Three-Tier Database*, spesifikasi yang dibutuhkan berikutnya adalah komputasi paralel. Yaitu kondisi di mana server ada lebih dari satu PC dan bersama-sama melakukan tugas analisis dengan membagi sistem kerja. Model komputasi paralel mampu meredam kinerja pemrosesan data semantik dan membuat perangkat keras dari server lebih tahan lama karena tidak dipaksa untuk melakukan tugas yang berat. Model komputasi paralel yang digunakan dalam penelitian ini adalah Hadoop. Beberapa studi pendahulu menggunakan Hadoop untuk mengelola data besar karena sistem hirarkis setiap *node* dapat diatur sesuai kebutuhan. Seperti pada Gambar 3.10., PC utama yang bertugas menerima perintah dan mengirim hasil pengerjaan yang disebut Master Node dan PC lain yang membantu Master Node dalam melakukan tugas disebut *Slave Node*.



Gambar 3.13. Arsitektur level yang diusulkan menggunakan Hadoop dengan tiga bagian komputer server yang berfungsi sebagai *slave computer*

Pada Hadoop, karena data semakin besar, *Node Slave* dapat ditambahkan lagi tanpa harus mematikan atau membangun kembali sistem yang telah berjalan. Pada rancangan ini, *cloud SaaS* ditempatkan pada node master karena SaaS adalah lapisan pertama tempat pengguna mengakses, kemudian dari tindakan pengguna akan dikirim perintah ke setiap *slave node*.

3.6 Rancangan keseluruhan sistem pakar DNA

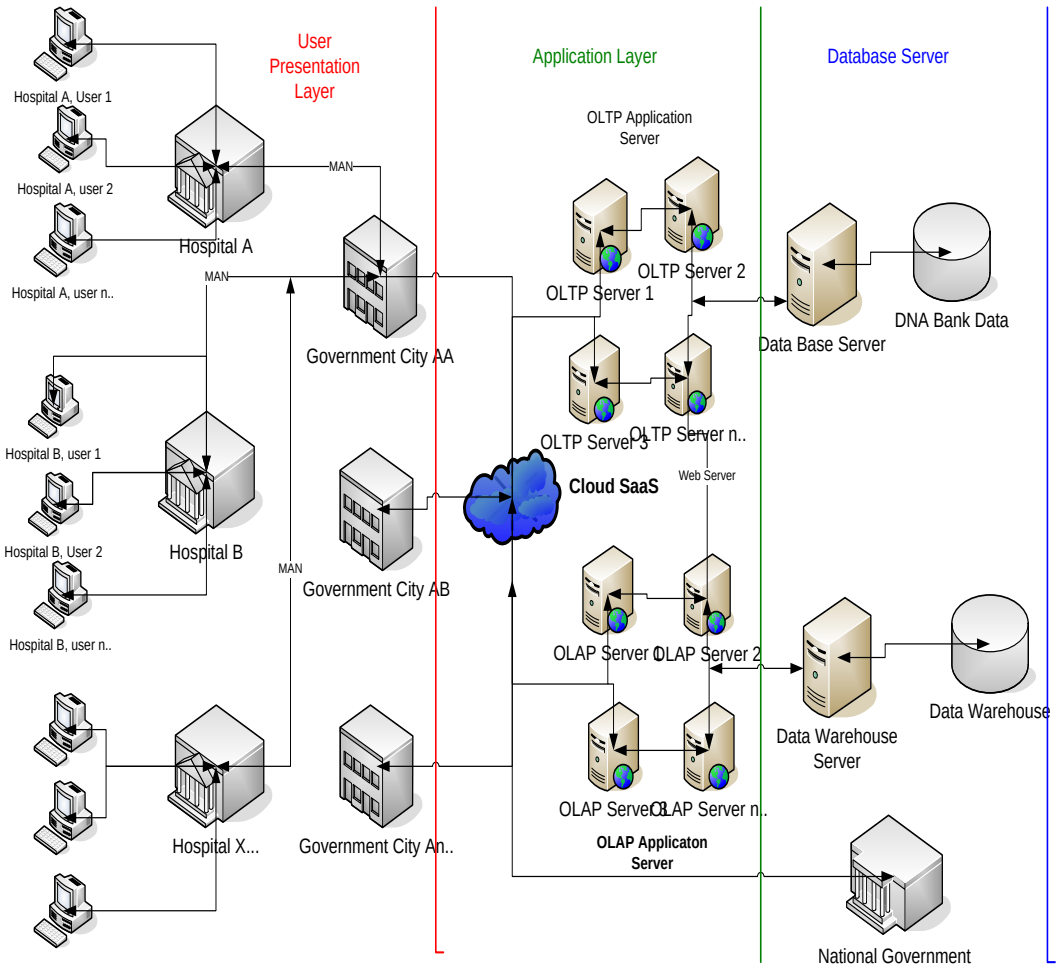
Setelah tahap analisis dapat dirumuskan perancangan sistem pakar analisis DNA yang berkelanjutan untuk masa depan. Sistem ini dirancang sesuai dengan kebutuhan dalam kendala yang dihadapi masalah dan mempromosikan konsep kota pintar. Melalui Gambar 3.5**.** dapat dijelaskan bahwa di setiap rumah sakit ada pengguna 1 untuk n-pekerjaan yang menganalisis berbagai penyakit berdasarkan DNA pasien. Data DNA apa pun yang dimasukkan oleh analis medis, riwayat pasien hanya disimpan di rumah sakit, sementara data urutan DNA akan masuk ke database pusat melalui jaringan MAN (*Metropolitan Area Network*). Analis medis dapat melakukan berbagai proses analisis dan fitur yang disediakan oleh sistem.

Data DNA yang dikirim ke server pusat pertama-tama akan memasuki proses OLTP (*Online Transaction Process*). OLTP adalah proses normalisasi data dan memasukkan data ke dalam bank data. Aktivitas yang terjadi pada server DNS dimasukkan, diperbarui, dan dihapus. Setiap hari, ribuan pengguna dari berbagai rumah sakit akan melakukan aktivitas OLTP dengan data yang besar dan dapat terjadi pada saat yang sama, oleh karena itu dibutuhkan komputasi paralel pada server OLTP.

Selain melalui proses OLTP, data DNA akan langsung melalui proses OLAP yaitu proses analisis online. Analisis yang dapat dihasilkan melalui OLAP mencakup jumlah pasien dengan penyakit X di suatu area, usia, tahun, dan bahkan tingkat kematian. Di masa depan, sistem tidak hanya mampu menganalisis peristiwa masa lalu, tetapi dapat melakukan sistem prediksi. Proses analisis ini dapat terus dikembangkan

sesuai dengan kebutuhan yang akan datang, oleh karena itu *server database* dirancang adalah bentuk database tiga-tier, sistem dapat melakukan pengembangan tanpa harus menghentikan sistem yang ada.

Untuk menghindari pencurian data atau penipuan, sistem dibangun di Cloud Computing. Dalam penelitian ini, jenis komputasi awan yang dibangun adalah SaaS (Perangkat Lunak sebagai layanan). SaaS jauh lebih aman daripada aplikasi berbasis web biasa, karena klien tidak mengakses sistem secara langsung tetapi harus terlebih dahulu membuat permintaan pada layanan web. Layanan web yang ada akan membaca



Gambar 3.14. Desain keseleruhan infrastruktur CESDA

permintaan dari klien dalam bentuk file XML. File XML akan diteruskan ke server aplikasi untuk diproses OLTP dan OLAP yang kemudian akan diteruskan ke server *database*.

Setiap server dibangun berdasarkan komputasi paralel, di mana setiap tugas yang ditetapkan akan diparalelkan oleh beberapa server. Jutaan catatan data DNA yang disimpan di bank data akan memakan waktu lama untuk diproses jika hanya dilakukan oleh satu server.

3.7 Metode pengujian yang diusulkan

Pengujian sistem dilakukan dengan membuat simulasi sistem mini menggunakan empat PC:

1. Satu PC bertindak sebagai Server Pemerintah yang dihubungkan oleh Cloud SaaS
2. Dua PC menginstal Hadoop. Satu PC bertindak sebagai Master Node dan satu PC lainnya bertindak sebagai Slave Node
3. Satu PC bertindak sebagai server Database atau Data Warehouse Server.

Setelah empat PC selesai proses pengaturan dan instalasi, maka pengujian sistem akan dilakukan. Fase pengujian sistem tersebut adalah:

1. Pengujian OLTP, sistem akan dikondisikan sebagai server OLTP. Sistem akan diberi tugas memasukkan lima data *isolated* DNA bersama, masing-masing *isolated* DNA terdiri dari 15.000 urutan. Sistem harus dapat melakukan normalisasi data secara bersamaan tanpa kesalahan apa pun. Hasil dari pemrosesan akan diteruskan ke server *database*.
2. Pengujian OLAP, sistem akan dikondisikan sebagai server OLAP. Data hasil normal pada proses OLTP akan diproses untuk analisis pada proses OLAP.

Setiap isolat akan dibandingkan dengan penyakit primer. Primer akan dibandingkan dengan setiap urutan isolat. Hasil OLAP adalah deteksi penyakit dengan toleransi kesalahan pencocokan lebih dari 70%. Toleransi> 70% diasumsikan sebagai mutasi genetik dari urutan. Hasil dari proses analitik akan disimpan ke dalam *server data warehouse.*

Pengukuran tingkat keberhasilan didasarkan pada hasil uji coba. Yaitu dengan melakukan simulasi sistem *mini-cloud*, dengan server paralel berbasis Hadoop dengan aplikasi sederhana yang dapat memproses DNA semantik.

Kecerdasan buatan untuk menganalisis DNA dengan berbagai tujuan telah dipelajari secara luas. Selain itu, aplikasi pendukung juga telah banyak diproduksi oleh perusahaan perangkat lunak di seluruh dunia, baik itu open source atau berbayar. Saat ini, di Indonesia, data DNA hanya dimiliki oleh masing-masing rumah sakit. Tidak dapat dimanfaatkan oleh lembaga lain dan pemerintah juga tidak dapat mengendalikannya. Oleh karena itu diusulkan infrastruktur yang terhubung dan terkoordinasi secara nasional. Tenaga medis dapat menggunakannya untuk kebutuhan penelitian, dan pemerintah dapat menggunakannya untuk kebijakan kesehatan. Tujuan OLAP adalah hasil dari analisis yang tepat dan dengan *Data Warehouse* (DW) data besar dapat dengan cepat diproses. *Cloud SaaS* memungkinkan aplikasi dikembangkan dengan mudah sesuai kebutuhannya di masa depan. CESDA adalah salah satu solusi untuk masalah analisis DNA di Indonesia saat ini.

3.8 Kesimpulan Penelitian

Pada tahap pertama penelitian, Penulis melakukan beberapa kali perubahan penetapan metode untuk tahap ini. Pada awalnya Penulis melakukan analisa string matching dengan membandingkan metode Knuth-Morris-Pratt, Boyer Moore, dan Brute Force untuk mencari metode terbaik. Hasilnya adalah metode Boyer Moore memiliki tingkat pergeseran perbandingan paling rendah, sehingga waktu yang

dibutuhkan untuk mencari pattern tersebut akan lebih pendek pula, dengan kata lain Boyer Moore dapat menemukan pattern lebih cepat dibandingkan dengan dua metode yang lain. Kemudian seiring berkembangnya penelitian Penulis, Penulis menambahkan factor mutasi di dalam *isolated* DNA yaitu memungkinkan adanya urutan *isolated* DNA yang hanya mirip saja bukan sama persis dengan urutan primer. Metode yang Penulis usulkan adalah metode Hamming. Pada proses pengujian Penulis menemukan kendala bahwa primer dengan jumlah karakter sedikit akan menghasilkan jarak yang lebih pendek dibandingkan dengan primer lainnya. Oleh karena itu Penulis menambahkan metode normalisasi pada tiap panjang karakter primer tersebut menggunakan metode minimax. Hasil dari normalisasi tersebut Penulis tambahkan ke dalam hasil penghitungan jarak Hamming, metode tersebut Penulis usulkan dengan nama "Hamming ternormalisasi".

Semakin besar data yang gunakan, semakin panjang pula waktu yang dibutuhkan pada proses *Semantic Similarity*. Metode Hamming ternormalisasi yang diusulkan sebelumnya membutuhkan dua proses penghitungan yaitu jarak Hamming dan normalisasi primer. Kemudian metode tersebut Penulis bandingkan dengan metode Edit Levenshtein Distance dengan akurasi yang sama. Namun Edit Levenshtein Distance mampu memberikan jarak minimum dibandingkan Hamming dan satu kali proses. Hal ini dikarenakan selain menghitung perbedaan, Edit Levenshtein Distance mampu melakukan operasi perubahan seperti menghapus dan menambah. Sehingga Penulis tetapkan bahwa untuk proses "Semantic Similarity" Penulis menggunakan metode Edit Levensthein Distance.

Setelah pemilihan metode *semantic similarity* yang tepat, tahapan penelitian selanjutnya adalah melakukan perancangan infrastruktur yang tepat untuk aplikasi sistem pakar DNA analisis di suatu negara, khususnya di Indonesia. Di masa sekarang ini, banyak analisa medis baik itu pengenalan penyakit maupun metode menyembuhannya di dasarkan pada pola suatu DNA. Dari hasil wawancara dengan ahli, selama ini, setiap rumah sakit memiliki system informasi pengolahan data DNA masing-masing dan tidak terintegrasi, Bank Genetik Dunia dipakai untuk rujukan

tersebut. Pada tahap ini diusulkan sebuah sistem pakar yang terintegrasi dengan seluruh rumah sakit di Indonesia dan Pemerintah. Sehingga setiap rumah sakit dapat bertukar data dan berkolaborasi dalam penelitian.

Pemerintah berfungsi sebagai penjembatan yang memfasilitasi proses tersebut. Sistem Pakar tersebut berada di SaaS (*Software as a System*) dengan bantuan pararel computing berbasis Hadoop yang mampu melakukan proses OLTP dan OLAP. Skema database yang diusulkan yaitu *database tree tier* yaitu skema dimana server di bagi menjadi tiga bagian sebagai penyimpan data, sebagai aplikasi, dan sebagai proses OLAP dan OLTP. Dengan catatan, bahwa penelitian ini masih dalam bentuk rancangan dan analisa belum ke tahap ujicoba dan implementasi. Pemerintah dalam hal ini adalah Kementrian Kesehatan yang mengatur regulasi antar rumah sakit di Indonesia serta regulasi tentang vaksin.

Keterangan:

Publikasi yang dihasilkan dari topik ini yaitu:

Berlian Al Kindhi, Tri Arief Sardjono, "Pattern Matching Performance Comparisons as Big Data Analysis Recommendations for Hepatitis C Virus (HCV) Sequence DNA", 3rd IEEE Artificial Intelligence and Modelling System, Malaysia, 2015

Berlian Al Kindhi, M. A. Hendrawan, D. Purwitasari, T. A. Sardjono, M. H. Purnomo, "Distance-based pattern matching of DNA sequences for evaluating primary mutation", IEEE 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Indonesia

Berlian Al Kindhi, Tri Arief Sardjono, Mauridhi Hery Purnomo, "Prototype Infrastructure Cloud Expert System DNA Analysis (CESDA) as the Basis of Sustainability DNA Software Improvement in Indonesia", IEEE 22nd European Modelling Symposium (EMS), Manchester, United of Kingdom, 2017

BAB 4
HYBRID CLUSTERING UNTUK ANALISA TREN DNA

Setiap untai DNA tunggal terdiri dari 10 urutan nukleotida. Urutan-urutan ini tidak dapat dipisahkan atau diatur secara acak karena setiap urutan DNA mengandung pengkodean genom tertentu. Ketika virus bermutasi, obat atau vaksin untuk virus yang telah diberikan kepada pasien akan menjadi tidak berguna. Oleh karena itu, ada kebutuhan untuk metode menganalisis kemungkinan arah mutasi DNA sehingga langkah-langkah pencegahan dapat diadaptasi lebih cepat. Virus tipe RNA mampu mengubah pola DNA yang terinfeksi, yang merupakan salah satu cara bagi virus tersebut untuk mempertahankan dirinya. Pada bab ini diusulkan metode pengelompokan hibrida baru yang menggabungkan kelebihan dari tiga metode K-means, fuzzy C-means, dan *hierarchical clustering* untuk memprediksi arah tren mutasi DNA. Gabungan dari ketiga metode ini diujikan pada dua set data dari 1000 data *isolated* DNA yang terdiri dari 500 *isolated* DNA terinfeksi dan tidak terinfeksi HCV dengan 37 primer HCV.

4.1 *Hybrid Clustering DNA*

Metode pengelompokan telah banyak diterapkan untuk analisis klasifikasi di berbagai bidang, salah satunya adalah analisis pola DNA. Proses pengelompokan yang sederhana dan cepat adalah salah satu keuntungan dari metode pengelompokan dibandingkan dengan metode pembelajaran mesin lainnya. Ketika membentuk kelompok sebagai bagian dari proses pengelompokan, centroid harus dihasilkan sebagai pusat gugus; centroid ini dapat dibuat secara acak atau ditentukan secara manual. Kemudian, untuk setiap titik data, jarak ke centroid akan dihitung, dan node dengan jarak terpendek ke centroid akan dianggap sebagai anggota gugus. Ketika menganalisis pola DNA, klasifikasi dan metode pengelompokan digunakan untuk

mengidentifikasi penyakit pada makhluk hidup, hubungan di antara spesies tertentu, dan perubahan dalam pola DNA.

K-means *clustering* adalah metode pemodelan tanpa pengawasan yang pertama kali digunakan oleh MacQueen. K-Means bekerja dengan membagi sejumlah objek ke dalam partisi berdasarkan kategori atau kondisi yang ada relatif terhadap titik tengah atau pusat terdekat. Metode ini meminimalkan variasi antara data dalam *cluster* yang sama dan memaksimalkan variasi antar data dalam berbagai *cluster*. Data DNA dapat dianalisis dengan mengevaluasi kesamaan antara *isolated* DNA dan primer. Analisis semacam itu bertujuan untuk menemukan pola dalam *isolated* DNA karena penelitian ini bertujuan untuk menganalisis apakah jenis primer yang diujikan tersebut terkait dengan *isolated* DNA dari berbagai negara. Untuk mengatasi masalah ini, model pengelompokan hierarkis yang sesuai harus diterapkan. Ada dua pendekatan untuk pengelompokan hierarkis: *agglomerative (bottom-up)* dan memecah belah (*top-down*). Dalam pengelompokan *agglomerative*, setelah menemukan jarak terpendek, tiga pendekatan dapat digunakan untuk menentukan bagaimana data harus ditugaskan ke *cluster*, yaitu pendekatan hubungan tunggal (jarak terpendek), pendekatan hubungan lengkap (jarak terjauh), dan keterkaitan rata-rata pendekatan.

Dalam DNA yang diisolasi dari satu organisme, ada sekitar 9.000 hingga 15.000 urutan nukleotida. Hal ini memungkinkan dalam satu *isolated* DNA untuk menunjukkan kesamaan dengan lebih dari satu primer. Dalam kasus seperti itu, pengelompokan Fuzzy C-means (FCM) adalah pendekatan yang sesuai karena pengelompokan FCM memungkinkan data untuk ditugaskan ke lebih dari satu klaster. Setiap *isolated* DNA memiliki beberapa derajat keanggotaan dengan masing-masing *centroid*. Dalam pengelompokan FCM, persyaratan minimum untuk keanggotaan dalam klaster ditentukan, dan *node* tersebut dapat menjadi milik lebih dari satu klaster jika memenuhi jarak minimum yang ditentukan.

Dalam prakteknya, proses pengelompokan tidak dapat dikatakan sesederhana yang diharapkan. Terkadang set data yang digunakan rumit karena data tidak

terstruktur dan klasifikasi bertingkat, dan target yang diharapkan bisa multidimensi. Setiap urutan adalah pola yang mengandung informasi tertentu. Namun, pola-pola ini kadang-kadang tidak terstruktur meskipun kesamaannya, dan perubahan dalam suatu pola disebut mutasi. Tiga metode pengelompokan yang dijelaskan pada paragraf sebelumnya masing-masing memiliki kelebihan yang sesuai dan cocok untuk kasus-kasus tertentu. K-means *clustering* cocok untuk pengelompokan eksklusif, pengelompokan hierarkis cocok untuk pengelompokan yang jelas, dan pengelompokan FCM cocok untuk pengelompokan yang tumpang tindih. Bagaimana jika data yang sedang dipelajari terkadang mengandung data tumpang tindih dengan kemungkinan mutasi, dan centroidnya berbeda, tetapi hasil yang diharapkan adalah hierarkis? Makalah ini mengusulkan metode pengelompokan hibrida yang mewarisi kelebihan dari ketiga metode, dengan hasil yang diharapkan adalah bahwa setiap kelompok sekuens DNA akan menunjukkan kecenderungan terhadap kesamaan dengan primer tertentu.

K-means *clustering* digunakan untuk mengidentifikasi kecenderungan suatu *isolated* DNA terhadap suatu primer, pengelompokan hierarkis digunakan untuk menganalisis penyebaran virus hepatitis C (HCV) di negara-negara tertentu berdasarkan hubungan antara primer dan asal DNA yang terisolasi, dan pengelompokan FCM digunakan untuk menganalisis tren mutasi HCV. Hasil dari pengujian dari metode yang diusulkan akan dibandingkan dengan delapan metode pengelompokan alternatif: pohon keputusan, mesin vektor pendukung, Apriori, pemaksimalan harapan, *k-nearest neighbors*, Klasifikasi dan Regresi Pohon, Naive Bayes, dan metode K-means umum .

Metode *Decision Tree* (pohon keputusan) adalah metode klasifikasi dan prediksi berdasarkan pada penentuan relevansi data berdasarkan hasil keputusan. Hasil tersebut dapat dihitung berdasarkan atribut reguler dan kriteria koeksistensi atau dengan menggunakan konsep privasi diferensial untuk menghitung metrik pangatur pada informasi gain dan indeks Gini. Mengingat sejumlah besar data, pohon keputusan dan *random forest* (hutan acak) dapat dimanfaatkan dalam pembelajaran mesin untuk

memodelkan interaksi kompleks dalam bioinformatika dan biologi. Model pohon keputusan berdasarkan *Pearson Corelation Coeficient - Tree* (PCC-Tree) sebagai ukuran baru kualitas fitur untuk mengkonfirmasi atribut pemisahan optimal dan titik pemisahan selama pertumbuhan pohon keputusan. Ketika setiap node pada tingkat yang sama dikaitkan dengan atribut yang sama, ini akan menyebabkan kesalahan selama pemilihan fitur; Pohon seperti itu juga disebut pohon keputusan yang terlupa. Oleh karena itu, metode menganalisis kesalahan klasifikasi dalam pohon keputusan dengan algoritma pembulatan acak diperlukan. Metode pohon keputusan dapat diterapkan untuk memprediksi kinerja layanan, dapat diimplementasikan dalam perangkat keras, dan dapat digunakan untuk analisis dalam bioinformatika. Berbeda dengan pendekatan pohon keputusan, K-means *clustering* adalah metode pengelompokan non-hirarkis di mana data yang ada dipartisi menjadi satu atau lebih cluster. Namun, centroid dalam pendekatan K-means bisa berubah-ubah; sebuah node yang memiliki jarak terpendek ke semua node lain dalam sebuah *cluster* akan menjadi *centroid*.

Dalam pembelajaran mesin, *Support Vector Machine* (SVM) adalah sistem pembelajaran yang menggunakan hipotesis spasial seperti fungsi linear dinyatakan dalam hal fitur dalam ruang dimensi tinggi. Pengklasifikasi SVM bekerja berdasarkan prinsip *Structural Risk Minimum* (SRM) untuk menemukan *hyperplane* terbaik yang memisahkan dua kelas dalam ruang input. Algoritma pembelajaran untuk SVM didasarkan pada teori optimasi untuk pelaksanaan bias pembelajaran dalam pembelajaran statistik. Beberapa peneliti telah menerapkan SVM untuk klasifikasi DNA untuk memprediksi lokasi protein. Untuk menyederhanakan pencarian, metode prediksi DNA dapat fokus pada wilayah kodon.

Untuk klasifikasi tingkat intensitas terjadinya pola tertentu, algoritma Apriori dapat digunakan. Algoritma ini mengklasifikasikan set data berdasarkan intensitas data relatif terhadap kriteria tertentu. Metode ini dapat diterapkan dalam sistem pendukung keputusan. Algoritma*Expectation Maximum* (EM) digunakan untuk menemukan estimasi kemungkinan maksimum (ML) dari nilai parameter model probabilistik yang

juga bergantung pada variabel yang tidak diketahui. Inti dari metode ini terdiri dari langkah E, di mana ekspresi untuk harapan kemungkinan dihitung dengan langkah M, di mana perkiraan ML dihitung dengan memaksimalkan nilai dari kemungkinan pada langkah E. Namun, kedua metode ini (Apriori dan EM) jarang diterapkan untuk klasifikasi dan prediksi DNA dalam literatur.

Algoritma k-NN mengklasifikasikan *instance query* berdasarkan node lain di lingkungannya. Untuk mencari tahu apakah sebuah simpul berada di lingkungan itu, metode jarak Euclidean dapat digunakan. Jarak ke semua node akan dihitung, dan node *k* dengan jarak terdekat dianggap sebagai tetangga dari *instance query*, di mana *k* adalah nilai yang ditentukan sebelum permulaan proses. Kemudian, menggunakan tetangga sebanyak *k* yang dipilih, hasil mayoritas di antara node ini akan dihitung untuk menghasilkan keputusan.

Metode lain yang dapat diterapkan untuk pembelajaran yang diawasi adalah metode Klasifikasi dan Regresi Pohon (CART). Metode ini menjelaskan hubungan antara variabel respon (variabel dependen) dan variabel prediktor (variabel independen) dengan cara nonparametrik statistik. Liu et al. menerapkan pendekatan regresi pohon karena variabel respon yang dibentuk oleh pengambilan sampel data terus menerus; Namun, jika data sampel yang ada sesuai dengan variabel respon kategoris, maka pendekatan pohon klasifikasi harus digunakan. Pohon klasifikasi terdiri dari tiga tahap, yang membutuhkan sampel pembelajaran (L). Tahap pertama adalah tahap pemisah, di mana setiap penyortir bergantung hanya pada satu variabel independen. Untuk variabel independen Xj yang berkesinambungan dengan ruang sampel n, ada penyorting n-1 yang berbeda. Untuk kategori nominal kelas L, pemisah diperoleh dengan tepat. Namun, jika Xj adalah kategori ordinal, maka sorting L1 dilakukan. Metode pengurutan untuk CART dapat didasarkan pada indeks Gini.

Metode terakhir yang dipertimbangkan untuk perbandingan adalah metode Naive Bayes. Metode Naive Bayes dipilih untuk perbandingan dengan metode yang diusulkan karena kemudahan penerapannya, berdasarkan perbandingan atribut ambigu

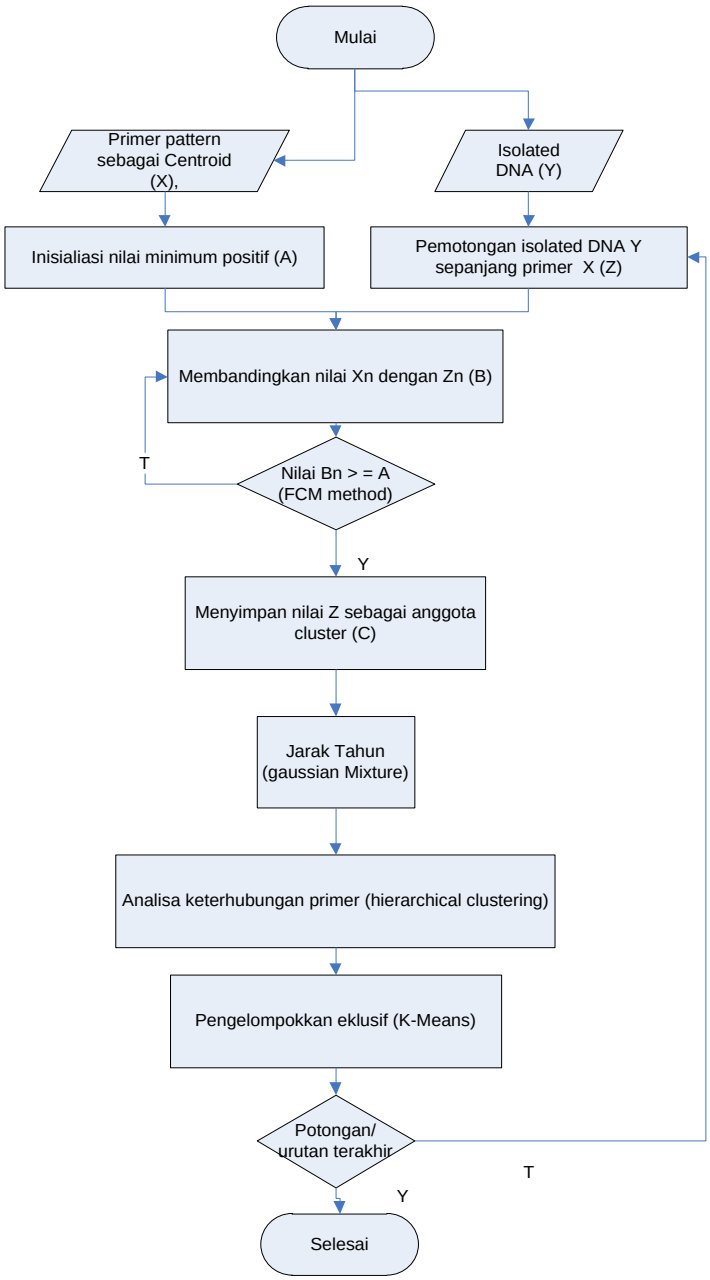
dan berbagai kondisi keputusan. Keputusan yang diperoleh oleh penggolongan Naive Bayes dapat berfungsi sebagai referensi untuk mengevaluasi hasil klasifikasi data. Atribut dalam metode Naive Bayes bebas dan tidak terkait satu sama lain; dalam proses analisis Bayesian, atribut disebut kondisi. Dalam kumpulan data tertentu, variabel *i* diasumsikan sesuai dengan vektor nilai atribut, dengan demikian, atribut yang menarik adalah *Xi (iϵ [1, n])*.

4.2 Proses Pengelompokan

4.2.1 Metode Pengelompokan *hybrid* yang diusulkan

Setiap urutan DNA yang terisolasi harus dibandingkan dengan masing-masing pola primer untuk menghitung nilai kedekatan masing-masing. Namun jumlah pola karakter tidak selalu sama tergantung pada jumlah karakter dalam primer HCV. Akibatnya, proses menemukan nilai kemiripan adalah proses yang tidak terstruktur. Oleh karena itu, selama proses pengelompokan, pemotongan urutan tidak didasarkan pada satu unting tunggal heliks DNA (satu putaran unting heliks terdiri dari sepuluh nukleotida) tetapi lebih pada panjang primer. Dengan demikian, prosedur perulangan yang digunakan dalam sistem juga didasarkan pada ukuran iterasi tak tentu pada pola tidak terstruktur.

Pada Gambar 4.1, proses pengelompokan hibrida dirinci, mulai dari input DNA terisolasi dari bank data dan pemotongan sepanjang pola primer dan berlanjut melalui proses penentuan anggota gugus. Data input dibagi menjadi dua jenis, yaitu, data primer HCV (x) dan data DNA (y) yang terisolasi. Setelah data primer telah dimasukkan, sistem akan menentukan nilai minimum yang urutannya dianggap positif terhadap setiap primer HCV (A). Kemudian, jarak kemiripan antara X dan Z (B) dihitung untuk setiap *n*. Jika nilai B lebih besar dari A, maka data terkait akan disimpan dalam basis data sebagai simpul positif sesuai dengan metode pengelompokan Fuzzy C-Means.



Gambar 4.1. Metode Hybrid *Clustering* yang di usulkan

Untuk memperjelas aliran metode yang diusulkan, *Pseudocode* disajikan di bawah ini. Gambar 4.2. adalah alur logika untuk proses pengelompokan, dari

memotong sekuens DNA terisolasi untuk menganalisis hasil dari proses pengelompokan.

Code Snippet 1 : Pseudocode dari metode yang diusulkan

```
Input: primer pattern = X[n]
         isolated DNA = Y[n]
Initialization:
         min positive tolerance = A
do , while last X[n]
do, while last Y[n]
 for(i=0 until i<= Y.length()
   Z[i] = slicing Y as long as X
   for(j=0 until j<= Z.size()
     similaritydistance()
     B=similarity distance between Z[i] and X[n]
     FCM methods
     if(B[j]>= A)
       count year distance
       saveToDB()
Primer linkage analysis (Hierarchical)
Exclusive Clustering Process (KMeans)
```

Gambar 4.2. *Pseudocode* dari metode yang diusulkan

Metode K-means terintegrasi dengan metode FCM seperti yang ditunjukkan dalam Persamaan (4.1)-(4.5) di mana x mewakili nilai *x* dari semua node $x = x_1, \dots, x_n$. *k* adalah jumlah cluster dan n adalah jumlah node di mana $(k \leq n)$. Dalam metode FCM, nilai masing-masing w_{ij} adalah konstan dan berada dalam jangkauan $\epsilon[0,1]$; dalam metode K-means, setiap μ_i, adalah variabel konstan. Dalam pengelompokan FCM, w_{ij} digunakan untuk menemukan derajat perbedaan antara setiap elemen dalam sebuah cluster, dan dalam K-means *clustering*, μ_i berfungsi untuk memaksimalkan kotak dari penyimpangan antara node dalam berbagai cluster. Variabel-variabel ini serupa dalam fungsi; dengan demikian, variabel w_{ij} dalam FCM dapat digabungkan dengan variabel μ_i dalam K-Means.

Jika

$$\arg\min C \sum_{i=1}^{n}\sum_{j=1}^{c} {w_{ij}}^{m} \tag{4.1}$$

dan

$$\arg\min S \sum_{i=1}^{k}\sum_{x\epsilon S_i} \|x-\mu_i\|^2 = \arg\min S \sum_{i=1}^{k} |S_i|\, Var\, S_i \tag{4.2}$$

Maka dapat diidentifikasi persamaan seperti berikut:

$$\sum_{x\epsilon S_i} \|x-\mu_i\|^2 = \sum_{x\neq y\in S_i} (x-\mu_i)(\mu_i - y) \tag{4.3}$$

μ_i adalah nilai konstan dalam $\epsilon[0,1]$ dan dapat dihitung dengan :

$$\mu_i = \frac{1}{\sum_{i=1}^{S} \left(\frac{\|x_i - S_j\|}{\|x_i - S_k\|}\right)^{\frac{2}{m-1}}} \tag{4.4}$$

Dimana

$$x = x_1, x_2, \dots, x_n \tag{4.5}$$

4.2.2 Prosedur validasi metode yang diusulkan

Langkah berikutnya adalah melakukan analisis ketepatan pengelompokan dengan membandingkan hasil metode yang diusulkan dengan delapan metode

penambangan data lainnya yang telah diuji dalam beberapa karya dalam literatur. Analisis dilakukan dengan menentukan ukuran kesenjangan antara satu metode dan metode lainnya dan nilai median di antara semua metode. Memprediksi mutasi virus dan bakteri tidaklah mudah; Namun, dalam beberapa literatur, proses prediksi telah dilakukan dengan menganalisis tren yang ada. Pengaturan urutan primer diperoleh berdasarkan data *isolated* DNA yang ada kemudian dianalisis menggunakan metode biologis yang disebut uji protein mengikat.

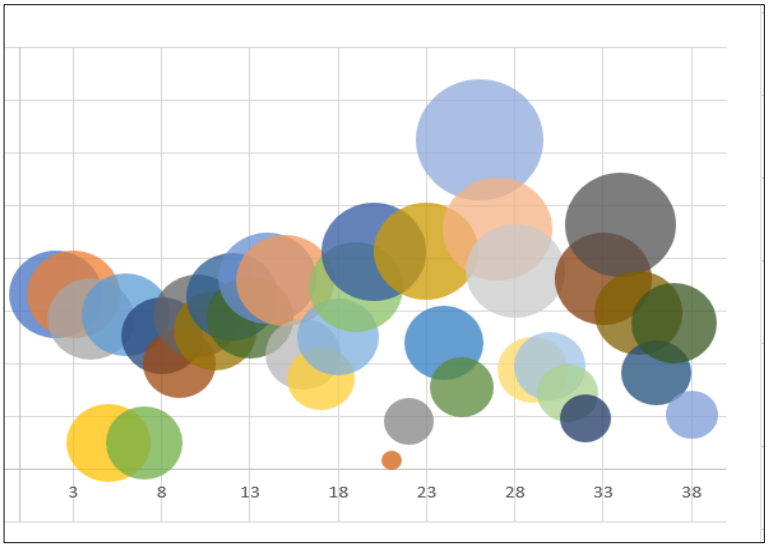
Untuk membandingkan kinerja metode yang diusulkan dengan metode lain, maka dievaluasi matriks internal. Matriks internal dihitung hanya menggunakan informasi dalam setiap *cluster* untuk mengevaluasi apakah kelompok telah dipisahkan dengan benar. Untuk matriks eksternal, dilakukan dengan menghitung jumlah rata-rata data yang terkait dengan kluster standar di antara semua metode sebagai referensi untuk matriks kebingungan. Matriks eksternal ini berfungsi sebagai dasar untuk uji statistik pada struktur data.

Matriks eksternal dibangun berdasarkan pembagian node. Jika, pada saat pengujian, lebih dari separuh metode yang diuji menunjukkan bahwa sebuah node termasuk dalam kelompok A, maka node tersebut akan digunakan dalam matriks kebingungan (*Confussion matrix)* sebagai *node True Positive* (TP), dan *node* yang tersisa adalah *True Negative* (TN). Kemudian, pada saat validasi, jika metode menunjukkan bahwa node adalah milik cluster A meskipun matriks eksternal mengatakan sebaliknya, maka node tersebut dapat dianggap sebagai *False Positive* (FP), dan *node False Negative* (FN) dapat diidentifikasi dengan cara yang sama.

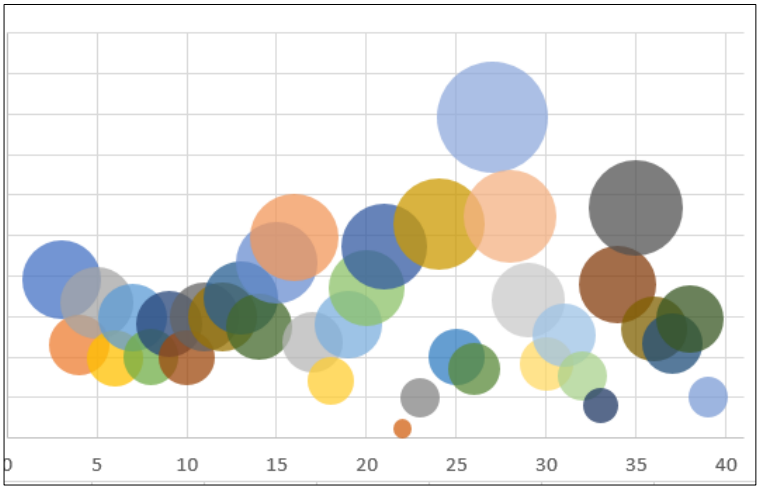
4.3 Hasil Pengujian

Hasil pengelompokan untuk dua set data, yaitu, kumpulan data yang terinfeksi HCV dan tidak terinfeksi, menunjukkan perbedaan yang signifikan. Dengan nilai kesamaan minimal 40%, setiap cluster yang terbentuk dari kumpulan data DNA terisolasi HCV-positif memiliki sejumlah besar anggota. Sebaliknya, menerapkan

proses pengelompokan ke set data DNA yang tidak terinfeksi dengan HCV menghasilkan rata-rata kurang dari 1000 anggota per cluster. Hasil ini menunjukkan bahwa penggunaan primer diperlukan untuk mendeteksi keberadaan virus atau bakteri dalam DNA yang terisolasi. Perbandingan jumlah anggota cluster antara kumpulan data yang terinfeksi HCV dan tidak terinfeksi dapat dilihat pada Gambar 4.3.



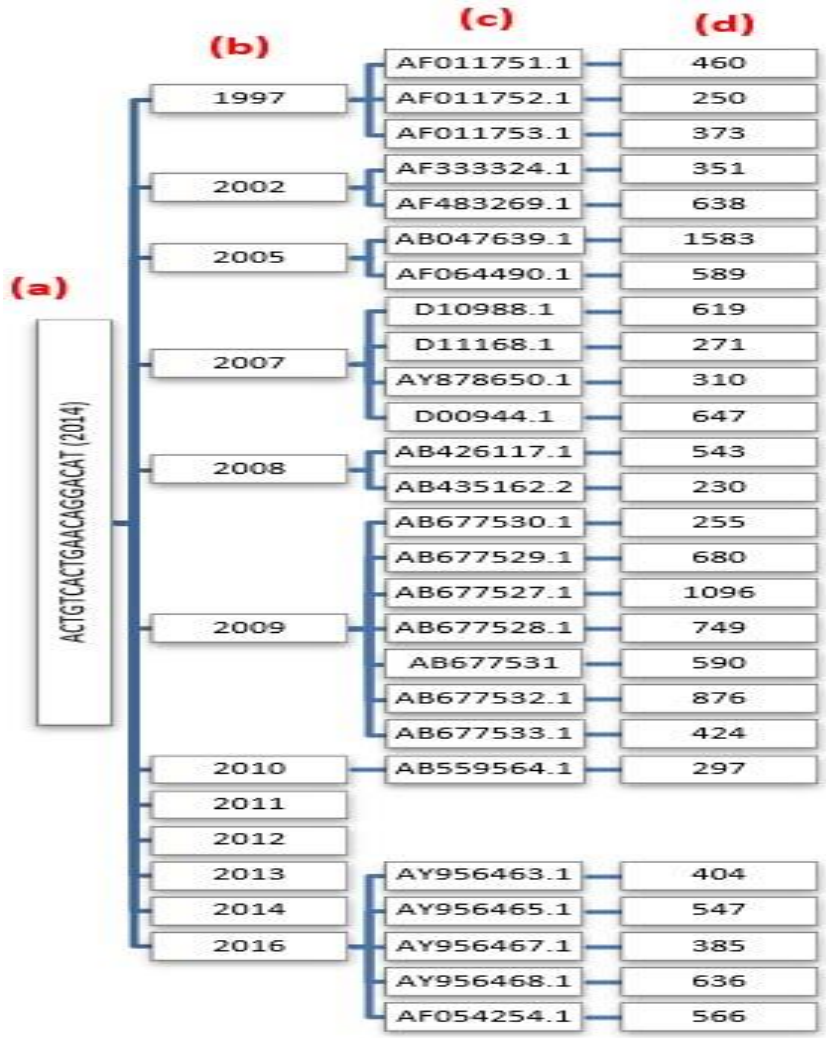
(a)



(b)

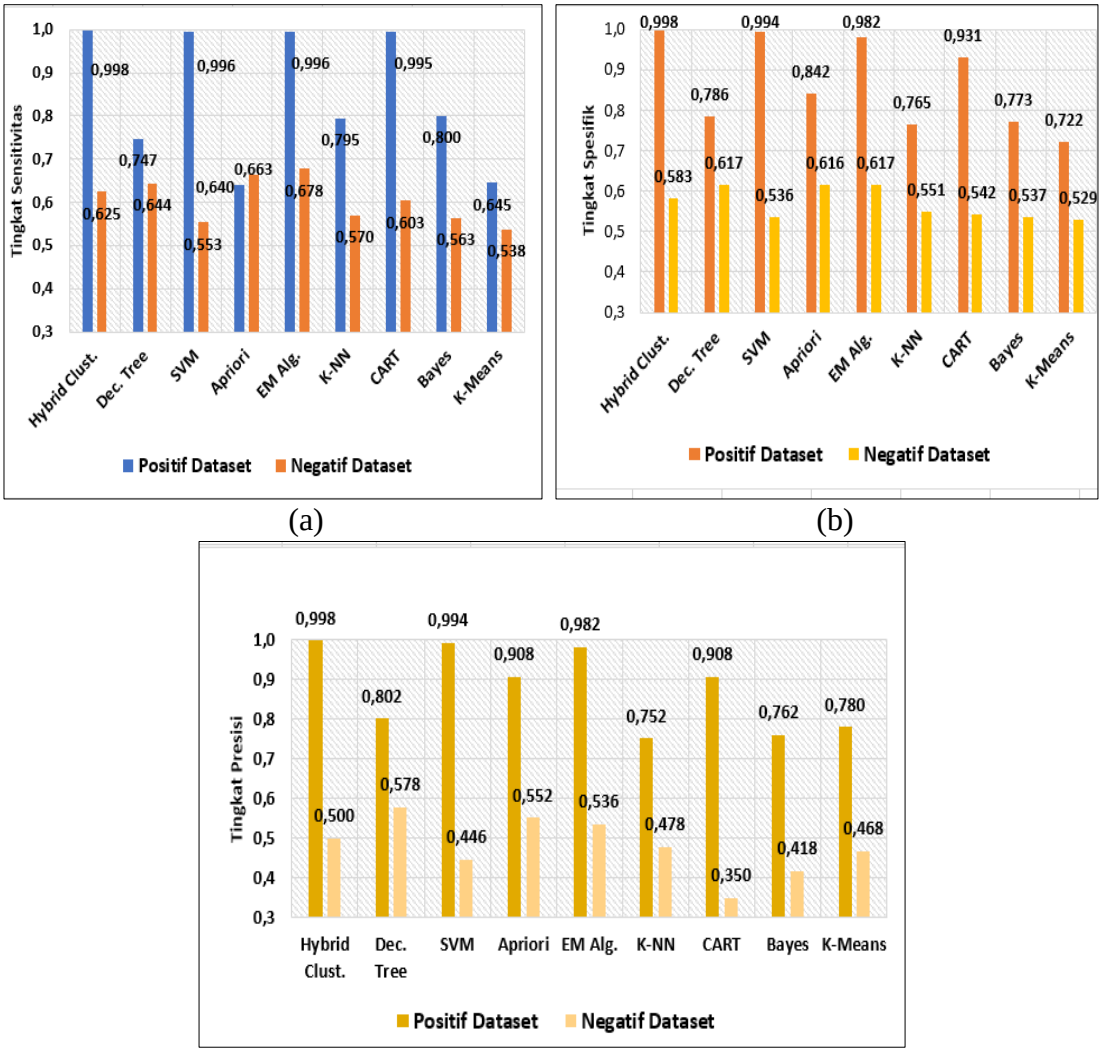
Gambar 4.3. Hasil pengelompokan hibrida pada 1000 isolat DNA yang (a) positif untuk infeksi HCV dan (b) negatif untuk infeksi HCV, di mana nomor primer HCV berada pada sumbu X

Pada Gambar. 4.3., dapat diamati bahwa ketika pengelompokan hibrida dilakukan pada DNA Homo sapiens yang tidak terinfeksi HCV dengan tingkat kesamaan minimal 40%, kelompok terbesar mengandung 1.583 *node*, dan kelompok terkecil memiliki 49 node anggota. Sebaliknya pengelompokan pada data positif HCV memiliki jumlah anggota sangat tinggi pada masing-masing kelompoknya hingga mencapai dua puluh ribu anggota dalam satu kelompok.



Gambar 4.4. *Hierarchical clustering* dari 50 isolat DNA HCV-positif terhadap salah satu primer yang diuji: (a) adalah pola primer, (b) daftar tahun-tahun publikasi dari isolat DNA, (c) daftar kode akses file dari isolat DNA, dan (D) daftar jumlah urutan di setiap isolat yang cocok positif untuk primer

Gambar 4.4. menunjukkan contoh hasil pengelompokan hirarkis dari 50 *isolated* DNA HCV positif terhadap satu primer, ACTGTCACTGAACAGGACAT (2014). Pengelompokan awal didasarkan pada masa tahun *isolated* DNA didaftarkan gen bank dunia; kemudian, tingkat pengelompokan berikutnya didasarkan pada kode akses dan jumlah simpul yang memenuhi ambang kesamaan minimum.



Gambar 4.5. Analisis komparatif dari metode yang diusulkan dan delapan metode alternatif pada parameter evaluasi kinerja: (a) sensitivitas, (b) spesifisitas, dan (c) presisi

Dalam pengujian tersebut, ambang kesamaan minimum adalah 50% dari jumlah karakter dalam primer. Selanjutnya pengelompokan hierarkis dapat dilakukan dengan menganalisis lokasi setiap urutan (posisi indeks yang menunjukkan di mana letak urutan nukelotida tersebut pada *isolated* DNA), dan kemudian, urutan itu sendiri dapat ditampilkan di bawah itu. Pendekatan ini dimaksudkan untuk memudahkan pencarian posisi dari sekuen yang diketahui dalam DNA yang terisolasi. Pada Gambar 4.4., ditampilkan empat lapis hirarki, hingga jumlah simpul positif, untuk mewakili hasil pengelompokan hierarkis.

Jumlah rata-rata dari simpul FP dapat diamati dari Gambar 4.5, yang menyajikan dua set grafik batang, satu untuk set data positif (yang terinfeksi HCV) dan satu untuk kumpulan data negatif, di mana sumbu y mewakili sensitivitas, spesifisitas, dan ketepatan dari masing-masing metode yang diuji. Gambar. 4.5. menampilkan analisis komparatif dari metode pengelompokan yang berbeda. Berkenaan dengan hasil tingkat FP, spesifisitas dan sensitivitas cenderung lebih tinggi untuk metode SVM daripada metode lain yang dipertimbangkan untuk perbandingan, dan metode pengelompokan hibrida yang diusulkan cenderung untuk mencapai nilai yang lebih tinggi.

4.4 Analisa Hasil Pengelompokan *Hybrid*

Pengaturan urutan primer diperoleh berdasarkan data *isolated* DNA yang ada yang dianalisis menggunakan metode biologis yang disebut uji protein mengikat; dengan demikian, dapat disimpulkan bahwa primer seperti itu, diperoleh melalui analisis data yang ada, bukan urutan baru. Apakah mungkin membuat prediksi dengan menghasilkan urutan primer baru? Mungkin saja, tetapi proses verifikasi akan cukup rumit, dan hasilnya akan sulit untuk divalidasi. Dengan demikian, dalam penelitian selanjutnya, akan berfokus untuk membuat prediksi berdasarkan pola dan urutan primer yang baru diisolasi.

Hasil dari proses pengelompokan hibrida yang diusulkan pada data terinfeksi HCV positif menunjukkan bahwa primer dengan cluster terkecil memiliki 627 node anggota, sedangkan cluster terbesar berpusat pada primer dengan 24.992 node anggota. Hasilnya menunjukkan bahwa satu isolat mungkin menunjukkan kesamaan yang tinggi dengan lebih dari satu primer, dan oleh karena itu, daerah gugus dapat ditumpuk seperti ditunjukkan pada Gambar 4.2.

Selain itu, juga diuji data Homo sapiens yang terisolasi yang negatif untuk infeksi HCV karena kesamaan urutannya dengan primer HCV. Hasilnya menunjukkan bahwa ada banyak urutan yang memiliki nilai kesamaan tinggi sehubungan dengan primer HCV. Temuan ini menegaskan bahwa gen tertentu mengandung mutasi dalam urutan nukleotida tertentu yang dapat diprediksi yang merupakan indikasi seseorang yang rentan terhadap HCV, gen tersebut disebut juga gen IL28. Namun demikian, jumlah sekuens nukleotida positif dalam *isolated* DNA Homo sapiens yang tidak terinfeksi tidak setinggi pada *isolated* yang terinfeksi HCV, dan ada banyak *isolated* yang tidak terinfeksi yang sama sekali tidak mirip dengan primer mana pun. Ketika dilakukan uji coba lagi setelah menurunkan ambang kesamaan minimum hingga 40% dari total karakter primer, lebih banyak urutan nukleotida dalam kumpulan data negatif (non-HCV-terinfeksi) dapat tergabung ke kluster primer.

Primer yang diuji, yang berasal dari beberapa studi dalam literatur, merupakan urutan nukleotida yang diharapkan cocok untuk perbandingan terhadap semua *isolated* DNA positif HCV di bank gen dunia karena tujuannya mengisolasi dan menduplikasi DNA sampel. Seperti ditunjukkan pada Gambar 4.2., hasil tes pada 500 *isolated* DNA yang terinfeksi HCV menunjukkan bahwa primer yang diuji selalu dapat dikaitkan dengan salah satu urutan dalam *isolated* DNA. Namun, pada DNA Homo sapiens normal yang tidak terinfeksi, jumlah urutan yang mirip dengan primer sangat kecil bahkan ketika batas bawah pada nilai kesamaan diturunkan menjadi 40%. Semakin banyak data sampel dimasukkan ke dalam *database*, semakin besar sensitivitas sistem, memungkinkan perubahan tren dianalisis secara lebih rinci.

Metode K-means menentukan *isolated* DNA mana yang paling terkait erat dengan suatu primer. Hal ini dicapai dengan mempertimbangkan baik jumlah urutan positif sehubungan dengan primer dan jarak tahun antara primer dan *isolated* DNA. *Isolated* DNA dengan sejumlah besar sekuens positif dan jarak tahun yang dekat dengan primer akan ditugaskan ke kelompok yang sama dengan primer yang terkait, seperti yang diamati pada Tabel 4.1. Setiap *isolated* DNA dapat ditugaskan hanya untuk satu kelompok; ini dilakukan dengan menentukan primer yang jumlah urutan tertinggi adalah pencocokan positif.

Pengelompokan FCM memungkinkan setiap *isolated* DNA dikaitkan dengan lebih dari satu primer. Setiap urutan dalam *isolated* DNA akan ditugaskan ke klaster primer yang merupakan pencocokan positif. Oleh karena itu, dalam satu *isolated* DNA, bisa ada ratusan urutan milik *cluster* primer yang berbeda. *Cluster* primer yang paling banyak urutannya dapat disimpulkan untuk mewakili tren primer utama; Oleh karena itu, vaksin HCV dapat dirancang berdasarkan pola primer ini. Hasil pengelompokan FCM tumpang tindih ini dapat dilihat pada Gambar 4.2.

Pendekatan hierarchical *clustering* dapat mengidentifikasi tren primer yang terkait dengan tahun publikasi dan negara asal. Pada Gambar. 4.4. kita dapat mengamati bahwa primer dengan pola ACTGTCACTGAACAGGACAT (a) dikaitkan dengan jumlah urutan positif terbesar pada tahun 2009 (b), untuk kode file *isolated* DNA yang diawali dengan huruf AB, seperti AB677527.1,(c), yang berarti bahwa isolate DNA berasal dari Jepang.

Tahap berikutnya dari analisis ini adalah untuk membandingkan hasil pengelompokan dari metode yang diusulkan dengan delapan metode pengelompokan yang paling sering digunakan. Hasilnya menunjukkan bahwa untuk kasus pengelompokan data microarray DNA, metode SVM umum lebih cocok daripada metode lainnya. Namun, ketika metode SVM dibandingkan dengan metode yang diusulkan, jumlah rata-rata hasil FP jauh lebih rendah untuk metode pengelompokan hibrida. Tingkat FP adalah jumlah urutan nukleotida yang salah diidentifikasi sebagai

memiliki kesamaan yang tinggi dengan primer. Selain itu, jumlah anggota di setiap kelompok K-means sangat kecil sehingga keakuratan menurun, seperti yang dapat kita amati pada Gambar 4.4. Hal ini karena metode K-means memungkinkan setiap node hanya milik satu klaster.

4.5 Pembahasan Hasil Pengelompokan *Hybrid*

Pada bab ini diusulkan metode pengelompokan yang sederhana dan efektif yang cocok untuk analisis DNA. Dengan beberapa metode, pengelompokan hanya dapat dilakukan pada data numerik; oleh karena itu diusulkan metode pengelompokan untuk data string DNA yang terdiri dari urutan nukleotida di mana preprocessing similaritas semantik digunakan untuk mengkonversi data string menjadi data numerik. Metode pengelompokan hibrida yang diusulkan adalah kombinasi dari tiga metode pengelompokan yang mencapai peningkatan kinerja dengan memanfaatkan keuntungan dari metode ini untuk menekankan tiga aspek data yang berbeda: pengelompokan hierarkis untuk pengelompokan jelas, K-means pengelompokan untuk pengelompokan eksklusif, dan pengelompokan FCM untuk tumpang tindih pengelompokan. Dengan pendekatan K-means, dapat ditemukan kecenderungan masing-masing *isolated* DNA dikaitkan dengan primer tertentu, menunjukkan bahwa *isolated* DNA terinfeksi dengan subtipe HCV yang terkait dengan primer tersebut. Melalui pendekatan FCM, kita dapat menemukan tren primer dengan mengidentifikasi primer dengan jumlah anggota kelompok urutan yang memiliki kecocokan positif tertinggi, memungkinkan arah mutasi HCV untuk dianalisis. Dengan pendekatan hierarkis, dapat ditentukan tahun dan negara di mana primer paling sering muncul berdasarkan *isolated* DNA mana yang cocok positif dengan primer itu.

Di antara delapan metode yang dipertimbangkan untuk perbandingan dengan metode yang diusulkan, metode SVM memiliki kinerja tertinggi, tetapi sensitivitas dan spesifisitas metode pengelompokan hibrida lebih tinggi daripada metode SVM masing-masing sebesar 0,002 dan 0,004. Untuk pendekatan pengelompokan hibrida, nilai-nilai

dari metrik kinerja ini mencapai 0,998 pada satu set data yang terdiri dari *isolated* DNA yang positif untuk infeksi HCV. Karena metode yang diusulkan melebihi metode yang ada untuk pengelompokan data DNA *microarray*, metode pengelompokan hibrida yang diusulkan disimpulkan menjadi metode yang cocok untuk menganalisis mutasi virus dan bakteri. Dalam penelitian ini, ditemukan bahwa 90% dari data DNA terisolasi yang diuji (dengan publikasi tahun 2014-2017 dipublikasikan) menunjukkan kecenderungan yang kuat terhadap salah satu primer, dan hasil yang terkait dapat dianalisis untuk memprediksi arah mutasi HCV. Pekerjaan lain dalam metode hibrida yang diusulkan adalah mengklasifikasikan jumlah urutan positif berdasarkan tahun, dan mengekstraksi tren mutasi HCV dari tahun ke tahun berdasarkan jumlah anggota kelompok dengan primer sebagai pusatnya.

4.6 Kesimpulan Penelitian

Tahap ketiga dari penelitian ini adalah melakukan pengelompokan urutan DNA hasil dari prose "semantic similarity". Pada penelitian ini diusulkan penggabungan tiga metode pengelompokan untuk memberikan hasil yang berbeda-beda sesuai dengan kebutuhan. Tiga metode tersebut adalah K-Means, Fuzzy C-Means, dan Hierarchical *Clustering*. Ciri K-Means adalah melakukan pengelompokan secara ekslusif, yaitu satu node hanya dapat tergabung dalam satu kelompok. Tujuan dari pengelompokan dengan pendekatan K-Means adalah untuk mencari kecenderungan suatu *isolated* DNA terhadap primer tertentu. Berbeda dengan K-Means, Fuzzy C-Means mengijinkan anggotanya untuk tergabung ke dalam lebih dari satu kelompok, tujuan dari pendekatan ini adalah untuk mencari primer yang tren. Primer tren tersebut diidentifikasi dengan banyaknya jumlah *node* yang tergabung pada kelompok tersebut. Semakin banyak jumlah *node*, berarti primer tersebut sangat mudah menemukan HCV di dalam *isolated* DNA. Metode Pengelompokan terakhir yang digabungkan adalah metode Hirarkikal, pendekatan ini bertujuan untuk menganalisa asal infeksi HCV dari suatu *isolated* DNA. Dalam satu *isolated* DNA dilakukan suatu urutan akar dan menganalisa primer mana saja yang dapat ditemukan dalam *isolated* DNA tersebut. Metode Hybrid *Clustering*

yang diusulkan mampu memberikan beberapa solusi untuk kebutuhan analisa di dunia medis.

Keterangan:

Publikasi dari topik penelitian ini adalah:

Berlian Al Kindhi, T. A. Sardjono, M. H. Purnomo, G. J. Verkerke “Hybrid K-Means, Fuzzy C-Means, and Hierarchical Clustering for DNA Hepatitis C Virus Trend Mutation Analysis”, Expert System with Application Journal, 2018

Halaman ini sengaja dikosongkan

BAB 5
PREDIKSI HCV DALAM *ISOLATED* DNA

Virus dan bakteri mempertahankan diri dengan cara bermutasi. Dari hari ke hari, jumlah *pattern* mutasi tersebut semakin bertambah dan beragam. Sehingga tidak semua primer mampu mengenali adanya HCV di dalam *isolated* DNA tertentu. Suatu primer kemungkinan tidak cocok dengan *isolated* DNA tertentu yang berasal dari suatu negara, sebaliknya, suatu primer dapat saja sangat cocok dengan *isolated* DNA tertentu.

Pengolahan *isolated* DNA untuk mencari pola pada *sequence* nukleotida telah banyak dilakukan pada penelitian sebelumnya (Jun Hu; Yang Li; Ming Zhang; Xibei Yang; Hong-Bin Shen; Dong-Jun Yu, 2017) (Bin Liu; Shanyi Wang; Qiwen Dong; Shumin Li; Xuan Liu, 2016) (Jianmin Ma; Minh N. Nguyen; Jagath C. Rajapakse, 2009). Tujuan dari pencarian pola ini beragam, ada yang dipakai sebagai forensik, mutasi genetik, atau menganlisa adanya suatu penyakit di dalam suatu DNA (Luis Alberto Hernandez Montiel, 2016). DNA forensik biasanya dilakukan untuk mengenali suatu individu berdasarkan hubungan keluarga. Pengenalan pola DNA juga dapat digunakan untuk mendeteksi adanya penyakit di dalam suatu *isolated* DNA, seperti misalnya pengenalan penyakit kanker. Selain itu, pengenalan pola DNA dilakukan untuk memprediksi adanya mutasi genetik, yaitu perubah suatu susunan DNA akibat terinfeksi oleh virus atau bakteri.

Untuk mencari pola di dalam suatu *isolated* DNA, dibutuhkan sebuah metode pengenalan baik itu metode *string matching*, pattern distance, maupun semantic similarity. Metode *string matching* yang sering digunakan pada pengenalan pola DNA adalah Booyer Moore, Knuth Morris Pratt, dan Brute Force, dan dari ketiga metode tersebut, metode Boyer Moore memiliki performansi yang paling tinggi (Berlian Al Kindhi; Tri Arief Sardjono, 2015). Untuk metode *pattern distance*, dapat menggunakan

metode hamming, hausdorff dan edit levenshtein distance (Berlian Al Kindhi; Muhammad Afif Hendrawan; Diana Purwitasari; Tri Arief Sardjono; Mauridhi Hery Purnomo, 2017). Dari ketiga metode tersebut, metode Edit Levenshtein *distance* memiliki performansi paling baik. Sedangkan untuk *semantic similarity* dapat menggunakan metode berbasis skor terhadap *pattern* yang dibandingkan. Pada penelitian ini digunakan metode Edit Levenshtein *distance* sebagai parameter *pre-processing*, karena sudah melalui proses ujicoba pada penelitian yang berbeda.

Hasil dari pengenalan pola dapat digabungkan dengan metode prediksi seperti jaringan syaraf tiruan, fuzzy, dan SVM. Tujuan dari penggabungan metode ini adalah untuk mendeteksi adanya mutasi atau penyakit di dalam susunan DNA. Pada penelitian ini, digabungkan pengenalan pola dengan berbagai jenis metode SVM. Tujuannya adalah memprediksi adanya infeksi HCV di dalam DNA manusia. Ujicoab SVM kernel dilakukan dengan enam pendekatan kernel yang berbeda dan dibandingkan performansinya masing-masing.

Salah satu tujuan pencarian suatu sequence di dalam *isolated* DNA adalah untuk menganalisa adanya suatu mutasi virus atau bakteri di dalam *isolated* DNA. Dalam satu *isolated* DNA dapat terdiri dari puluhan ribu sequence nukleotida, dan dalam bank gen terdapat jutaan *isolated* DNA. Hal ini akan membutuhkan waktu yang lama jika pencarian sequence tersebut dilakukan secara manual. Selain itu mutasi dari virus atau bakteri tersebut dapat berbeda-beda sesuai dengan bagaimana RNA pada *isolated* tersebut mempertahankan diri. Penelitian ini mempelajari cara menentukan apakah sebuah sequence tersebut adalah suatu mutasi dari HCV dan cara mengenalinya pada *isolated* DNA.

Pada algoritma *Support Vector Machine (SVM)*, waktu yang diperlukan untuk mengklasifikasikan titik data yang tidak diketahui adalah proporsional dengan jumlah vektor dukungan (Sebastián Maldonado; Julio López, 2018) (David de la Mata-Moya; María Pilar Jarabo-Amores; Jaime Martín de Nicolás; Manuel Rosa-Zurera, 2017). Tergantung pada kompleksitas struktur kelas, kadang-kadang jumlah vektor dukungan

dari model SVM meningkat dengan jumlah titik data pelatihan (Deepak KumarJain; Surendra Bilouhan Dubey; Rishin Kumar Choubey; Amit Sinhal; Siddharth Kumar Arjari; Amar Jain; Haoxiang Wang, 2018). Salah satu solusinya adalah dengan mengurangi jumlah vektor dukungan, namun tetap mempertahankan tingkat akurasi yang sama seperti SVM normal yang tidak menggunakan pengurangan vektor dukungan (Rupan Panja; Nikhil R. Pal, 2018). Sebuah SVM menemukan *hyperplane* yang memisahkan memaksimalkan margin pemisahan dan karenanya, lokasi *hyperplane* terutama tergantung pada satu set "titik batas" (Saurabh Paul; Malik Magdon-Ismail; Petros Drineas, 2016). Algoritma SVM dapat diterapkan pada data pelatihan yang dikurangi untuk menghasilkan model klasifikasi. Model klasifikasi ini dilakukan dengan menilai kinerja dengan melonggarkan definisi titik batas (M. A. Ebrahimi; M. H. Khoshtaghaza; S. Minaei; B. Jamshidi, 2017). Selain itu, optimasi SVM dapat juga dilakukan dengan memperluas algoritma ke ruang fitur menggunakan transformasi kernel. Dalam hal ini, sebuah *vector* dukungan dihasilkan dalam ruang fitur menggunakan matriks kernel terkait (Samia Djemai; Belkacem Brahmi; Mohand Ouamer Bibi, 2016).

Terdapat dua cara yang standar dalam mesin pembelajaran, yaitu *supervised* dan *unsupervised*, namun beberapa literatur telah menemukan algoritma yang mampu menyelesaikan masalah data tak terstruktur yaitu dengan pengubahan algoritma menjadi *semi-supervised* (semi-terbimbing) (Yong Liu; Shizhong Liao, 2017) (Sidheswar Routray; Arun Kumar Ray; Chandrabhanu Mishra; G. Palai, 2018). Pembelajaran semi-terbimbing adalah salah satu paradigma pembelajaran yang paling menjanjikan dalam banyak aplikasi praktis di mana beberapa sampel berupa data tak terstruktur. Di antara model pembelajaran semacam itu, SVM adalah yang paling umum dan menonjol (Jing Zhou; Ying Yang; Steven X. Ding; Yanyang Zi; Muheng Wei, 2018). Namun, SVM *semi-supervised* yang khas tidak dapat memperkirakan distribusi sampel positif dan negatif dengan baik. Salah satu solusi dari masalah tersebut adalah dengan menyajikan kombinasi dari dua strategi multi-klasifikasi untuk mengurangi waktu berjalan dan meningkatkan akurasi klasifikasi secara bersamaan.

Metode tersebut disebut juga dengan *ensemble S3 SVM* (Dan Zhang; Licheng Jiao; Xue Bai; Shuang Wang; Biao Hou, 2018). Metode SVM dapat menangani masalah klasifikasi semi-terbimbing bahkan dengan distribusi yang tidak diketahui atau data yang tidak seimbang.

Dalam bioinformatika, SVM sering digunakan untuk klasifikasi suatu citra maupun data kesehatan. Selain data tidak terstruktur, data multi dimensi juga merupakan salah satu faktor dibutuhkannya inovasi pada algoritma yang sudah ada. Liu dkk. mengusulkan sebuah aplikasi yang mampu mereduksi data multi dimensi, yaitu *Minimax Concave Ridge Support Vector Machine (MCR SVM)* yang secara bersamaan melakukan klasifikasi dan pengurangan dimensi (Jian-wei Liu; Li-peng Cui; Xiong-lin Luo, 2016). Pengklasifikasi SVM yang diusulkan oleh Liu dkk. ini menggabungkan keuntungan dari ketidaksempurnaan estimator SVM dan kemampuan seleksi grup fitur SVM untuk mengatasi kerugian. Selain itu, juga memberikan pembenaran teoritis untuk fitur sparsial yang dipilih.

Dalam aplikasi pembelajaran terbimbing, SVM juga mampu menentukan efektivitas pengujian obat di rumah sakit. Tugas yang paling membutuhkan waktu dari obat-obatan adalah mendiagnosis dan memilih pengobatan. Secara tradisional, dokter telah memecahkan masalah ini, hanya mengandalkan pada intuisi dan pengalaman mereka sendiri. SVM mampu memberikan analisa diagnosa dan saran obat dengan melakukan pengelompokan berdasarkan gejalanya. Parameter yang digunakan untuk pengujian SVM ini adalah deskripsi matematis dan perumusan masalah. Tujuan pemanfaatan SVM pada sistem pakar adalah membuat prasyarat untuk diagnosis pencegahan penyakit pasien. Dengan menggunakan SVM dan sistem pemantauan, spesialis dapat mendiagnosis dan mengembangkan perawatan yang optimal dengan benar (C. Venkatesan; P. Karthigaikumar; Anand Paul; S. Satheeskumaran; R. Kumar, 2018)

Metode SVM dapat digabungkan dengan metode yang lain sebagai aktivitas sebelum klasifikasi. SVM dapat digabungkan dengan metode *Convolutional Neural*

Network (CNN) untuk mendeteksi adanya leukemia di dalam tubuh manusia (Luis H.S.Vogado; Rodrigo M.S. Veras; Flavio. H.D. Araujo; Romuere R.V. Silva; Kelson R.T. Aires, 2018). CNN berfungsi sebagai penentu prediksinya, kemudian untuk membagi antara kelompok yang terjangkit leukemia dan bukan dapat digunakan metode SVM. Pada penelitian ini, diterapkan metode *Edit Levenshtein distance* dengan optimasi SVM. *Edit Levenshtein Distance* berfungsi untuk menganalisa adanya pola di dalam *isolated* DNA, kemudian hasil dari pengenalan pola tersebut akan menjadi matriks set data sebagai masukan dari SVM.

5.1 SVM Kernel

Pada sub bab ini akan dijelaskan ke enam jenis metode kernel SVM yang diujikan, yaitu Linear SVM, Quadratic SVM, Cubic SVM, Fine Quadratic Gaussian SVM, Medium Gaussian SVM, dan Coarse Gaussian SVM. Perbedaan dari ke-enam jenis metode SVM ini adalah cara menentukan *hyperplane*-nya, *hyperplane* adalah kunci dalam SVM untuk menentukan sebuah node tergabung dalam kelompok yang mana.

Data set pada penelitian ini adalah 1000 *isolated* DNA yang terdiri dari 500 *isolated* HCV dan 500 *isolated* homo sapiens non infeksi HCV. Data tersebut diunduh dari gen bank dunia. Kemudian sebagai parameter pembanding adalah primer HCV yang diperoleh dari Institut Penyakit Tropis, Universitas Airlangga. Setiap data *isolated* akan dinormalisasi ke dalam bentuk FASTA terlebih dahulu. Kemudian data FASTA tersebut akan dihitung nilai kemiripannya menggunakan metode *Edit Levenshtein Distance*. Hasil dari pengolahan metode *Edit Levenshtein Distance* tersebut akan dimasukkan ke dalam data matrix sebagai variabel *x* dalam SVM. Sehingga besarnya matrix data tersebut adalah sebesar jumlah *pattern* dikalikan dengan jumlah *isolated* DNA yaitu 37 x 1000. Sedangkan target atau variabel *y* adalah nilai positif atau negatifnya *isolated* tersebut terhadap HCV dan besarnya matrix target adalah 1 x 1000.

Data set pada penelitian ini dapat diterapkan pada SVM sesuai dengan Persamaan (5.1), dengan y_i adalah nilai antara -1 dan 1 yang mengindikasikan kelas dari x_i.

$$(x_i, y_1), .., (x_n, y_n)$$

$$\text{x} = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix};\ y = \begin{bmatrix} y_1 \\ .. \\ y_n \end{bmatrix} \tag{5.1}$$

A. Linear SVM

Linear SVM adalah algoritma pembelajaran mesin dari penambangan data untuk memecahkan masalah klasifikasi multikelas dari kumpulan data ultra besar yang mengimplementasikan versi asli dari algoritma pemisahan data untuk merancang SVM yang linear. Linear SVM adalah rutinitas linear skalabel yang berarti bahwa metode tersebut menciptakan model SVM dalam waktu CPU yang berskala secara linear dengan ukuran kumpulan data pelatihan. Metode ini membagi data menjadi dua kelompok yang terpisah secara jelas berdasarkan satu *hyperplane* linear.

Pada linear SVM, nilai x_i adalah p-dimensional dari vector, sehingga untuk mendapatkan jarak terdekat *hyperplane* ke dua grup yaitu antara x_i dengan nilai $y_i = 1$ dan $y_i = -1$ adalah dengan memaksimalkan margin dari *hyperplane.* Persamaan (5.2) adalah bentuk sederhana pemisahan dataset x_i menjadi dua kelompok.

$$\overrightarrow{\boldsymbol{w}} \cdot \vec{\boldsymbol{x}} - \boldsymbol{b} = \mathbf{0} \tag{5.2}$$

Ketika dataset yang ada tidak tentu, sedangkan hasil yang diinginkan adalah dapat terbagi secara linear maka dapat digunakan persamaan linear yang adaptif seperti pada Persamaan (5.3), dengan y_i adalah target yang diinginkan, dalam hal

ini adalah 1 untuk positif dan -1 untuk negatif. Sedangkan $(\vec{w} \cdot \vec{x_i} - b)$ adalah keluaran dari penghitungan SVM.

$$max(0,1 - y_i(\vec{w} \cdot \vec{x_i} - b)) \quad (5.3)$$

B. *Quadratic SVM*

Quadratic SVM adalah salah satu pendekatan pada SVM yang menyederhanakan fungsi kernel dengan melakukan fungsi kuadrat. Fungsi kuadrat ini dapat diterapkan pada data set yang saling terkait maupun yang data set jenis *time series* (Dagher, 2008). Penerapan fungsi kuadratik SVM dapat diamati pada Persamaan (5.4), dengan *T* adalah fungsi kuadrat urutan data atau *time series.*

$$\overrightarrow{w^T} \cdot x + b > 0, untuk\ y = 1$$

atau

$$\overrightarrow{w^T} \cdot x + b < 0, untuk\ y = -1$$

Maka *hyperplane:*

$$\overrightarrow{w^T} \cdot x + b = 0, (\vec{w}, b) \quad (5.4)$$

C. *Cubic SVM* dan *Fine Quadratic SVM*

Pendekatan SVM dapat diperluas ke permukaan non-linear dengan menggunakan trik kernel. Fungsi non-linear dapat memindahkan ruang asli ke-

ruang dimensi yang lebih tinggi. Trik kernel tersebut dapat diterapkan pada dua jenis metode SVM yaitu *Cubic* dan *Quadratic Gaussian* dengan d=3 untuk *Cubic* dan d=2 untuk *Quadratic Gaussian.* Trik kernel SVM untuk kedua metode dapat diamati pada Persamaan (5.5).

$$\boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x_j}) = (\boldsymbol{x_i}, \boldsymbol{x_j})^{\boldsymbol{d}} \tag{5.5}$$

D. *Medium* dan *Coarse Gaussian SVM*

Fungsi *Gaussian* pada SVM dikenal juga dengan fungsi radikal basis. *Gaussian SVM* termasuk salah satu jenis non-linier SVM yaitu setiap titiknya ditentukan oleh non-linier *kernel function.* Penghitungan trik kernel berdasarkan pada fungsi gauss yang diterapkan pada *hyperplane.* Yang membedakan antara *medium* dan *coarse gaussian* terletak pada cara penghitungan *hyperplane.* Hal ini memungkinkan didapatkannya *hyperplane* yang cocok untuk semua anggota, hasil penempatan *hyperplane* akan berbentuk kurva gauss. Penghitungan *gaussian* atau radikal basis dapat diamati pada Persamaan (5.6).

$$\mathbf{k}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{exp}\ (-\boldsymbol{\gamma}\left\|\boldsymbol{x_i}, \boldsymbol{x_j}\right\|^{\mathbf{2}}) \tag{5.6}$$

Pada Persamaan (5.6) dapat diterapkan untuk $\gamma = 0 \mid \gamma = 0{,}7$, jika parameternya berbeda maka $\gamma = 1/\ (2\sigma^2)$.

E. Optimasi SVM

Dari enam metode yang diujicoba, ditambahkan fungsi optimasi agar hasil pengujian dapat sesuai harapan. Optimasi tersebut adalah dengan menghitung nilai *hyperplane* seminimum mungkin. Secara umum, jarak antara dua *hyperplane* adalah $2/\|\vec{w}\|$, sehingga untuk memaksimalkan jarak antara dua *hyperplane* adalah dengan meminimalkan nilai $\|\vec{w}\|$. Optimasi SVM dapat diamati pada Persamaan (5.7), dengan γ adalah parameter yang menentukan antara meningkatkan ukuran margin dan memastikan bahwa $\overrightarrow{x_i}$ berada pada sisi yang benar dari margin.

$$y_i(\vec{w} \cdot \overrightarrow{x_i} - b) \geq 1, untuk\ i = 1, .., n$$

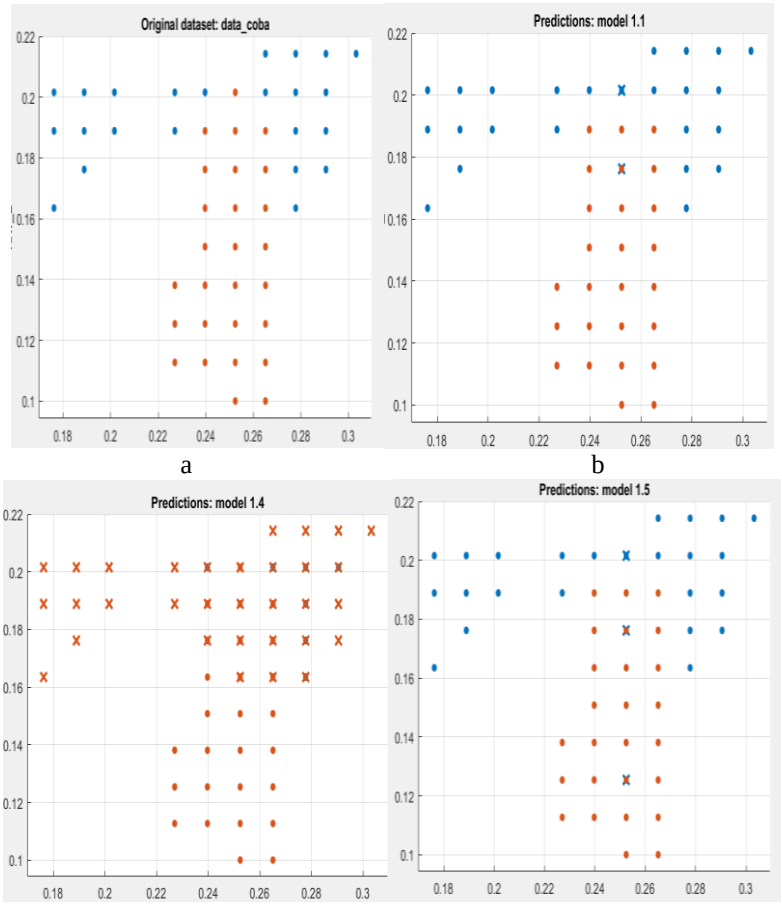
Sehingga,

$$\left[\frac{1}{n}\sum_{i=1}^{n} max\ (0,1 - y_i(\vec{w} \cdot \overrightarrow{x_i} - b))\right] + \gamma\|\vec{w}\|^2 \quad (5.7)$$

5.2 Hasil pengujian SVM Kernel

Seluruh metode SVM yang telah dijelaskan, diujikan menggunakan data set yang sebelumnya telah dinormalisasi ke dalam matriks berukuran 38x1000.

Data set tersebut berasal dari proses pengenalan pola pada *isolated* menggunakan metode *Edit Levenshtein Distance*, kemudian pada masing-masing *isolated*, dari seluruh *sequence* yang diujikan terhadap primer, akan diambil jarak terpendek terhadap masing-masing *isolated*. Jarak terpendek tersebut yang akan menjadi nilai variable *x* pada SVM. Sedangkan variable *y* adalah nilai prediksi apakah *isolated* tersebut positif atau negative yaitu 1 dan -1.



Gambar 5.1. Pemetaan plot data pada SVM a) data asli, b) plot data Linear, Quadratic, Cubic, Coarse Gaussian, c)plot data Fine Gaussian d)plot data Medium Gaussian

Hasil pemetaan data dapat diamati pada Gambar 5.1., plot biru adalah data negatif (grup 0) dan plot merah adalah positif (grup 1). Gambar 5.1.a menunjukkan plot data asli sebelum mengalami pemisahan kelompok dengan SVM, dimana plot merah dan plot biru terpisah dengan jelas. Gbr 1.b adalah pemetaan plot data dari ke empat metode yaitu *Linear, Quadratic, Cubic, dan Coarse Gaussian SVM.* Karena hasil pemisahan datanya yang sama, maka pemetaan plot datanya pun juga sama. Gambar 5.1.b memiliki dua plot error, yaitu data yang seharusnya bernilai positif namun mendapat hasil prediksi SVM bernilai negative. Gambar 5.1.c adalah pemetaan plot data metode Fine Gaussian SVM, pada gambar ini tampak bahwa banyak sekali

data yang mengalami kesalahan prediksi. Gambar 5.1.d. adalah pemetaan plot data Medium Gaussian SVM, tampak pada gambar tersebut terdapat 3 plot eror yaitu kesalah prediksi yang seharusnya bernilai positif diprediksi menjadi negatif.

Analisa dari pengujian prediksi masing-masing kernel SVM akan diukur performansinya dengan tingkat sensitivitas, spesifisitas, dan presisinya. Sensitivitas tes (juga disebut tingkat positif sejati – *True Positif*) didefinisikan sebagai proporsi orang dengan penyakit yang akan memiliki hasil positif. Dengan kata lain, hasil pengujian yang sangat sensitif adalah pengujian yang mengidentifikasi pasien dengan penyakit dengan benar. Pengujian yang 100% sensitif akan mengidentifikasi semua pasien yang memiliki penyakit. Pengujian dengan sensitivitas 90% akan mengidentifikasi 90% pasien yang memiliki penyakit, tetapi akan kehilangan 10% pasien yang memiliki penyakit. Hasil pengujian yang sangat sensitif dapat berguna untuk menyingkirkan penyakit jika seseorang memiliki hasil negatif. Sebagai contoh, hasil negatif pada HCV mungkin berarti orang tersebut tidak menderita kanker hati (sirosis). Akronim yang banyak digunakan adalah SnNout (*high Sensitivity, Negative result = rule out*).

Spesifisitas tes (juga disebut *True Negative Rate*) adalah proporsi orang tanpa penyakit yang akan memiliki hasil negatif. Dengan kata lain, pengujian tingkat spesifisitas merujuk pada seberapa baik metode mengidentifikasi pasien yang tidak memiliki penyakit. Pengujian yang memiliki spesifisitas 100% akan mengidentifikasi 100% pasien yang tidak memiliki penyakit. Sebaliknya, pengujian yang spesifik 90% akan mengidentifikasi 90% pasien yang tidak memiliki penyakit. Pengujian dengan spesifisitas tinggi (tingkat negatif benar tinggi) berguna ketika hasilnya positif. Pengujian yang sangat spesifik dapat berguna untuk menentukan pasien yang memiliki penyakit tertentu. Akronimnya adalah Spin *(high Specificity, rule in).*

Presisi adalah seberapa tingkat tingkat kebenaran pengujian metode dalam memprediksi bahwa pasien yang sakit dengan hasil positif dan bukan pasien dengan hasil negatif. Presisi mengukur seberapa imbangnya tingkat sensitivitas dan spesifitas suatu metode dengan menganalisa jumlah data yang diprediksi benar dan

mempertimbangkan bukan pasien yang diprediksi positif terhadap penyakit (*False Positive*).

Untuk dapat lebih memperjelas kinerja masing-masing metode SVM dapat diamati pada Tabel 5.1. Analisa yang dijelaskan pada Tabel 5.1. merupakan hasil pengolahan algoritma SVM menggunakan data set matriks hasil *semantic similarity* masing-masing *isolated* DNA terhadap masing-masing primer.

Tabel 5.1. Menunjukkan kecepatan masing-masing metode dalam mengolah data set. Dapat diamati bahwa metode linear SVM memiliki kecepatan prediksi paling lambat yaitu sekitar 6.000 obs/detik, dan waktu yang dibutuhkan untuk melatih data jauh lebih lama dibandingkan dengan metode yang lain. Waktu yang dibutuhkan metode Linear untuk melakukan pelatihan data set adalah sebesar 4,4208 detik, atau setara dengan 3-7 kali lebih lambat dibandingkan dengan metode lainnya. Metode Quadratic memiliki kecepatan prediksi paling tinggi, sedangkan waktu pelatihan paling rendah adalah metode Fine Gaussian SVM.

Tabel 5.1. Analisa Kecepatan prediksi masing-masing metode SVM

No.	Metode	Parameter Penentu Kernel	Kecepatan Prediksi (obs/dtk)	Waktu Pelatihan (detik)
1	Linear SVM	-	~ 6.000	4,4208
2	Quadratic SVM	d = 2	~ 19.000	0,84011
3	Cubic SVM	d = 3	~19,000	0,74427
4	Fine Gaussian SVM	-	~ 15.000	1,2663
5	Medium Gaussian SVM	-	~17.000	0,71109
6	Coarse Gaussian SVM	$\gamma = 0.7$	~ 18.000	0,69129

5.3 Pembahasan hasil prediksi SVM

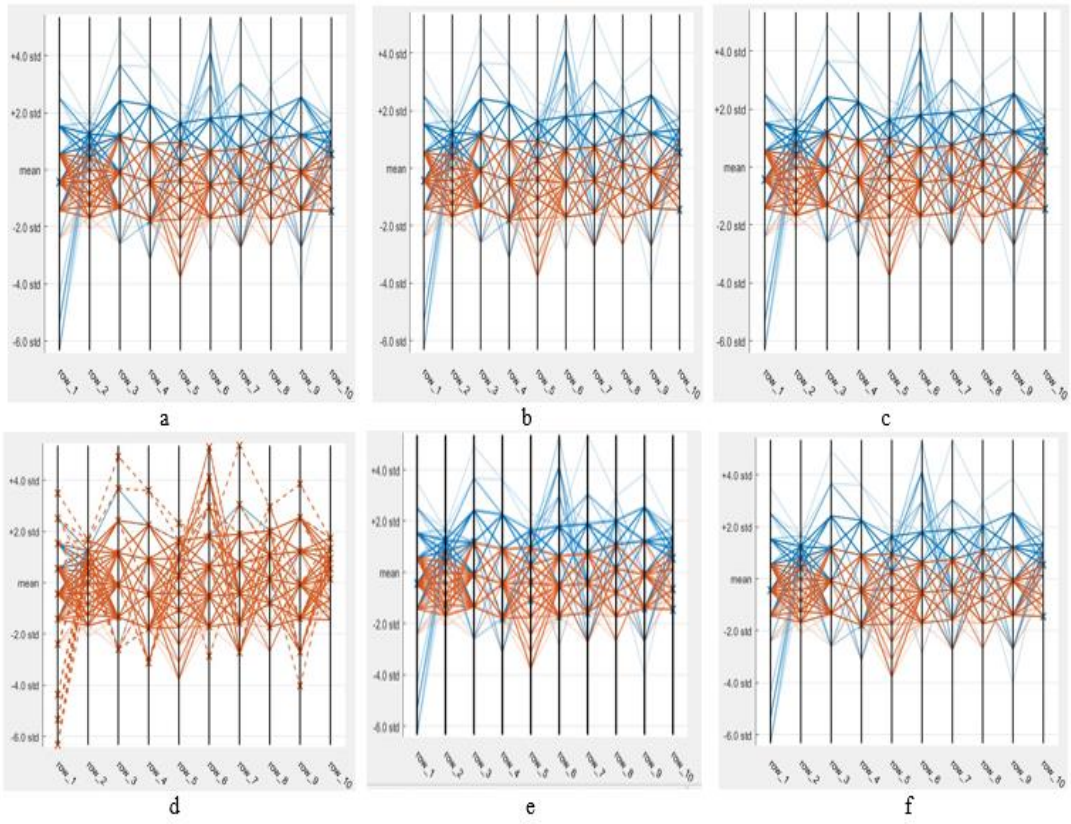
Dari ke-enam metode yang diujikan, keseluruhannya mampu mengolah data set. Hasil pengujian dapat diamati pada Gambar 5.2, dimana setiap plot merah dan biru dapat terpisah secara kontinyu. Sedangkan untuk metode Fine Gaussian SVM, tampak bahwa plot merah dan plot biru saling bertumpukkan (tidak terpisah) sehingga yang tampak adalah koordinat plot merah saja.

Pengukuran akurasi dari hasil percobaan pada penelitian ini menggunakan metode *10 K-Fold Cross Validation*, yaitu 1000 dataset akan dibagi menjadi 10 kelompok secara acak. Tiap satu kelompok terdiri dari 100 data sebagai target dan 900 data sisanya adalah sebagai data pelatihan. Proses ini akan berulang sebanyak jumlah kelompok yang ditentukan hingga seluruh kelompok telah menjadi data target pada prose SVM. Tujuan dari pengujian ini adalah untuk mengukur tingkat akurasi dari masing-masing metode SVM yang telah dioptimasi. Hasil pengolahan metode SVM dapat diamati pada Gambar 5.3., dimana 4 dari ke-enam metode memiliki nilai akurasi sebesar 99,8%, satu diantaranya 99,7%, dan satu lainnya sebesar 77,4% yaitu metode Fine Gaussian SVM.

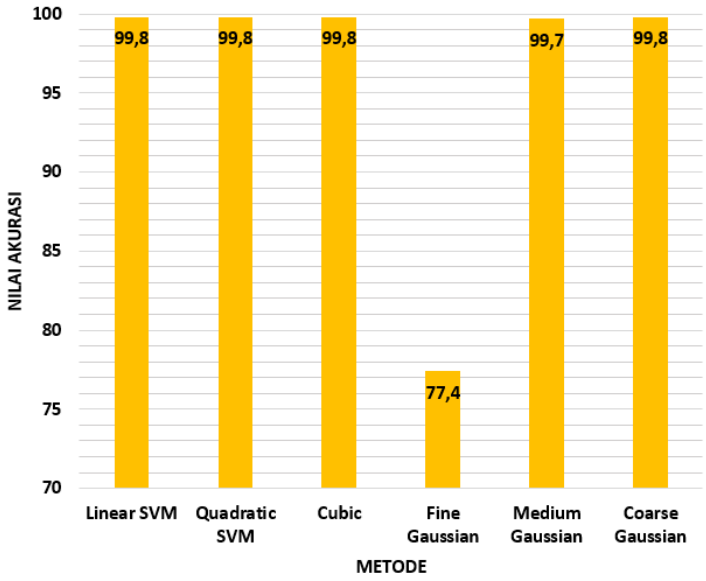
Tabel 5.2. Pengukuran tingkat akurasi masing-masing kernel SVM

No.	Metode	TP	TN	FN	FP
1	Linear SVM	498	500	2	0
2	Quadratic SVM	498	500	2	0
3	Cubic SVM	498	500	2	0
4	Fine Gaussian SVM	274	500	226	0
5	Medium Gaussian SVM	497	500	3	0
6	Coarse Gaussian SVM	498	500	2	0

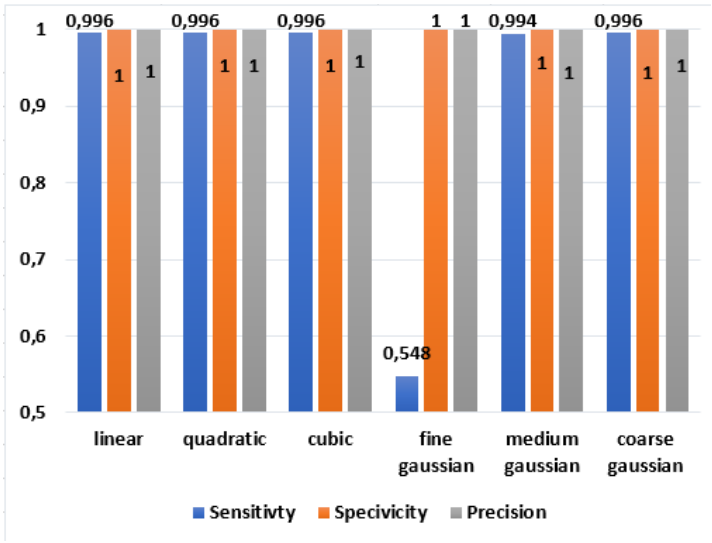
Analisa jumlah dataset dengan prediksi benar dan salah dapat diamati pada Tabel 5.2. TP adalah *True Positive* yaitu data yang dianggap positif oleh SVM dan memang data tersebut positif. TN adalah *True Negative*, yaitu data yang dianggap negatif oleh hasil pengolah SVM dan memang data tersebut negatif. FP adalah *False Positive* yaitu data yang dianggap positif oleh SVM padahal data tersebut sebenarnya negatif. FN adalah *false negative,* yaitu data yang dianggap negatif oleh SVM padahal data tersebut sebenarnya positif.



Gambar 5.2. Grafik pararel koordinat plot masing-masing metode SVM, a) Linear, b) Quadratic, c) Cubic , d) Fine Gaussian, e) Medium Gaussian, f)Coarse Gaussian SVM



Gambar 5.3. Tingkat akurasi masing-masing metode SVM



Gambar 5.4. Grafik sensitivitas (warna biru), spesifik (warna merah), dan presisi (warna abu-abu) masing-masing metode

Pengujian masing-masing SVM kernel menunjukkan hasil yang sangat baik, dari seribu data yang diujikan rata-rata hanya melakukan *error* prediksi pada dua-tiga data (kecuali kernel *Fine Gaussian* SVM). Melalui hasil TP, TN, FP, dan FN maka dapat dihitung nilai sensitivitas, spesifik, dan presisi. Hasil penghitungan sensitivitas tersebut dapat diamati pada grafik, pada Gambar 5.4.

Metode SVM banyak diterapkan untuk memisahkan set data yang berdasarkan *hyperplane*, yaitu garis yang membagi dua buah kelompok. Untuk memisahkan set data, SVM memiliki beberapa pendekatan berdasarkan bagaimana *hyperplane* membagi set data menjadi dua buah kelompok. Pada penelitian ini, digunakan beberapa pendekatan SVM antara lain *linear, quadratic, cubic, fine gaussian, medium gaussian,* dan *coarse gaussian*. Ke-enam metode kernel SVM tersebut diuji untuk menentukan suatu *isolated* DNA terinfeksi oleh HCV atau tidak.

Hasil pengujian menunjukkan, semua metode menghasilkan nilai performansi yang tinggi yaitu 99.8%, dengan kata lain, dari seribu data hanya terdapat dua data yang error prediksi. Metode Fine Gaussian adalah metode SVM yang memiliki performansi paling rendah pada studi kasus ini, yaitu sebesar 77,4% dengan 226 data error prediksi. Uji coba ini membuktikan bahwa sebagian besar pendekatan SVM mampu mengenali adanya mutasi HCV di dalam *isolated* DNA. Rencana penelitian lanjutan adalah menguji coba berbagai metode prediksi dan pengelompokan yang hasilnya akan dibandingkan dengan SVM untuk membuktikan metode yang terbaik untuk studi kasus penelitian ini.

5.4 Kesimpulan Penelitian

Tahapan akhir dari penelitian ini adalah melakukan prediksi adanya HCV pada *isolated* DNA. Semakin banyak jenis HCV ditandai dengan semakin bertambahnya subtype HCV, akan semakin sulit menemukan adanya mutasi di dalam *isolated* DNA. Maka diperlukan metode yang mampu memprediksi HCV dengan baik, salah satunya adalah dengan menganalisa 6 kernel metode SVM untuk melakukan pembelajaran

terbimbing dan menemukan adanya/tidak adanya HCV di 1000 *isolated* DNA yang diujikan. Hasilnya hampir semua kernel SVM mampu mengenali mana *isolated* DNA yang positif dan negatif terinfeksi oleh HCV. Namun metode SVM yang menggunakan Gaussian sebagai kernel, kurang bagus dalam mendeteksi sehingga menghasilkan lebih dari 15% eror prediksi.

Kesuksesan SVM dalam memprediksi adanya HCV di suatu *isolated* DNA tak lepas dari input data set yang bagus. Data set tersebut dibuat dari matriks berukuran 37x1000. Dimana 37 adalah jumlah primer yang dibandingkan dan 1000 adalah jumlah *isolated* DNA yang di uji. Dalam setiap kolom pada satu row, disimpan nilai terkcecil dari hasil penghitungan *Edit Levenshtein Distance* pada proses *Semantic Similarity*. Hasil pengujian pada metode SVM dengan berbagai kernel (selain kernel Gaussian) menunjukkan rata-rata hanya terdapat dua error prediksi dari 1000 data *isolated* DNA yang diujikan. Dari seluruh proses penelitian disertasi ini, disimpulkan bahwa akurasi yang tinggi pada proses pengelompokan dan prediksi tak lepas dari proses *semantic similarity* yang baik. Data set masukan yang baik akan mempermudah sistem dalam menjalankan algoritma dari metode *clustering* dan prediksi.

Keterangan:

Publikasi pada topik penelitian ini adalah:

Berlian Al Kindhi, T. A. Sardjono, M. H. Purnomo, " Optimasi Support Vector Machine untuk memprediksi adanya mutasi pada DNA Hepatitis C Virus", Jurnal Nasion Terakreditasi, JNTETI, Edisi Agustus 2018

Halaman ini sengaja dikosongkan

BAB 6
KESIMPULAN

6.1 Kesimpulan

Disertasi ini merupakan bagian dari inovasi di bidang teknik biomedika pada HCV. Pola mutasi HCV yang cepat mengakibatkan terus bertambahnya sub-tipe HCV dan hingga saat ini belum ada vaksin untuk HCV. Oleh karena itu urgensi untuk melakukan penelitian ini cukup tinggi. Beberapa pendekatan untuk mencapai penemuan anti-HCV (vaksin) dari segi teknik biomedika adalah dengan melakukan analisa tren mutasi primer tersebut. Disisi lain, dengan semakin beragamnya pola mutasi, mengakibatkan tidak semua primer mampu mendeteksi adanya HCV dalam *isolated* DNA, sehingga dibutuhkan metode prediksi yang mampu mempelajari pola tersebut dan dapat mengenali adanya HCV pada *isolated* DNA yang diuji-cobakan yang berasal dari berbagai negara di seluruh dunia.

Disertasi ini dibagi menjadi empat bagian yaitu proses *semantic similarity,* proses perancangan sistem pakar analisis DNA untuk pemerintahan, proses pengelompokan DNA untuk analisa tren, serta prediksi adanya HCV di dalam *isolated* DNA. Proses *semantic similarity* merupakan bagian penelitian yang paling membutuhkan waktu diantara sub penelitian lain yang dilakukan. Pada bagian penelitian ini dilakukan 3 kali ujicoba dengan 5 metode yang berbeda untuk mendapatkan metode yang terbaik. *Semantic Similarity* memilah-milah jutaan nukleotida dan mengukur tingkat kedekatannya dengan masing-masing primer. Hasil penelitian menunjukkan bahwa *Edit Levenshtein Distance* merupakan metode yang paling sesuai untuk studi kasus mutasi DNA. Setelah didapatkan nilai kedekatan masing-masing urutan terhadap primer, sistem akan melakukan proses pengelompokan berbasis metode *hybrid clustering* yang diusulkan, yaitu bagian penelitian ke tiga dari disertasi ini. Urutan-urutan yang memenuhi batas nilai positif akan di kelompokkan ke dalam klaster

dimana primer bertindak sebagai centroid. Pengelompokan ini menghasilkan tiga analisa sekaligus dalam satu kali proses, yaitu pengelompokan dengan pendekatan K-Means untuk mencari kecenderungan *isolated* DNA, pendekatan Fuzzy C-Means untuk mencari tren primer HCV, dan Hirarkikal untuk menganalisa runutan keterkaitan primer dengan *isolated* DNA. Selain itu, dengan menggunakan hasil pengolahan *semantic similarity*, data positif dan negatif akan dimasukkan ke dalam matriks sebagai *input features* metode SVM. Bagian terakhir dari penelitian ini memprediksi pola mutasi HCV dan menentukan apakah *isolated* DNA tersebut terinfeksi HCV menggunakan metode SVM. Hasil pengujian menunjukkan bahwa 5 dari kernel SVM yang diujikan menghasilkan tingkat akurasi yang tinggi. Sehingga dapat disimpulkan bahwa analisa pengelompokan dan prediksi akan memudahkan peneliti medis menuju ke arah yang lebih dekat dengan pembuatan vaksin.

6.2 Rencana Penelitian Lanjutan

Rencana penelitian lanjutan penulis adalah:

1. Mencoba metode Pengelompokan yang lain agar dapat memberikan analisa dan manfaat lebih luas untuk ahli. Selain itu, penulis akan
2. Mencoba mengembangkan metode prediksi yang lain baik dari segi optimasi maupun inovasi untuk mendapatkan hasil prediksi yang terbaik.
3. Untuk mengatasi masalah perbedaan primer dari satu *isolated* DNA dengan lainnya juga dibutuhkan suatu metode yang mampu memprediksi dengan akurat sehingga ke depannya ahli cukup menggunakan mesin pembelajaran tanpa primer untuk memprediksi adanya HCV di dalam *isolated* DNA.

4. Uji coba menggunakan data set yang lain juga merupakan salah satu fokus dalam penelitian lanjutan penulis. Penulis berharap bahwa penelitian ini tidak hanya untuk data set HCV saja namun juga dapat di aplikasikan untuk pengujian seluruh virus dan bakteri.

DAFTAR PUSTAKA

A. Apostolicoa, C. Guerraa , G.M. Landauc , C. Pizzie. (2016, July). Sequence similarity measures based on bounded hamming distance. *Theoretical Computer Science, 638*, 76-90.

A. Jansen, C. Frank, J. Koch, K. Stark. (2008). Surveillance of vector-borne diseases in Germany: trends and challenges in the view of disease emergence and climate change. *Parasitology Research*, S11-S17.

A.S. Pinheiro, H.P. Pinheiro, P.K. Sen. (2012). The Use of Hamming Distance in Bioinformatics. *Handbook of Statistics, 28*, 129-162.

Airlangga, U. (2012). Ilmu Ilmiah Lanjut Penyakit Tropis.

Airlangga, U. (2012). *Penyakit Tropis Ilmu Ilmiah Dasar*. Surabaya: Universitas Airlangga.

B.A. Kindhi, T.A. Sardjono. (2015). Pattern Matching Performance Comparison as Big Data Analysis Recomendations for Hepatitis C Virus (HCV) Sequence DNA. *The 3rd International Conference on Artificial Intelligence, Modelling and Simulation.* Sabah, Malaysia.

Berlian Al Kindhi; Muhammad Afif Hendrawan; Diana Purwitasari; Tri Arief Sardjono; Mauridhi Hery Purnomo. (2017). Distance-based pattern matching of DNA sequences for evaluating primary mutation. *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE).* Yogyakarta.

Berlian Al Kindhi; Tri Arief Sardjono. (2015). Pattern Matching Performance Comparisons as Big Data Analysis Recommendation for Hepatitis C Virus (HCV) Sequence DNA. *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS).* Kinabalu.

Bin Liu; Shanyi Wang; Qiwen Dong; Shumin Li; Xuan Liu. (2016). Identification of DNA-Binding Proteins by Combining Auto-Cross Covariance Transformation

and Ensemble Learning. *IEEE Transactions on NanoBioscience, 15*(4), 328-334.

Brijesh K. Sriwastava, Subhadip Basu, and Ujjwal Maulik. (2015). Predicting Protein-Protein Interaction Sites with a Novel Membership Based Fuzzy SVM Classifier. *IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 12*(6), 1394-1404.

Brisbois BW, Ali SH. (2010). *Climate Change, Vector-Borne Disease and Interdisciplinary Research: Social Science Perspectives on an Environment and Health Controversy. Ecohealth.* Heidelberg: Springer.

C. Chen and R. Veldhuis. (2009). Extracting biometric binary strings with minimal area under the FRR curve for the hamming distance classifier. *2009 17th European Signal Processing Conference.* Glasgow.

C. Sammut; G.I. Webb. (2010). *Encyclopedia of Machine Learning.* Springer US.

C. Venkatesan; P. Karthigaikumar; Anand Paul; S. Satheeskumaran; R. Kumar. (2018). ECG Signal Preprocessing and SVM Classifier-Based Abnormality Detection in Remote Healthcare Applications. *IEEE Access, 6*, 9767-9773.

C.D. Ramesh , P. Sharmila , G. P. S. Dhillon , P. D. Aditya . (2010). Climate change and threat of vector-borne diseases in India: Are we prepared? New York/Heidelberg: Springer-Verlag;.

C.S. Rao ; S.V. Raju. (2016). Next generation sequencing (NGS) database for tandem repeats with multiple pattern 2°-shaft multicore string matching. *Genomics Data, 7*, 307–317.

Country progress report on HIV/AIDS response . (April 2012.). (Minstry of Health and Federal HIV/AIDS Prevention and Control Office).

Cronkite, D. (2002). *Cell and Heredity.* New-Jersey: Prentice-Hall,Inc.

D. Martin, M. Burstein, D. McDermott, S. Mc Ilraith, M. Paolucci, K.Sycara, D. L. Mc Guinness, E. Sirin and N. Srinivasan. (2007). Bringing Semantics to Web Services with OWL-S. *world wide web 2007*, hal. 243-277.

Dagher, I. (2008). Quadratic kernel-free non-linear support vector machine. *Journal of Global Optimization, 41*(1), 15–30.

Dan Zhang; Licheng Jiao; Xue Bai; Shuang Wang; Biao Hou. (2018). A robust semi-supervised SVM via ensemble learning. *Applied Soft Computing, 65*, 632-643.

Database, D. (accessed 15 Juni 2015). *DDBJ Database RDP1,*. (http://www.ddbj.nig.ac.jp/cgi-bin/wgetz).

David de la Mata-Moya; María Pilar Jarabo-Amores; Jaime Martín de Nicolás; Manuel Rosa-Zurera. (2017). Approximating the Neyman–Pearson detector with 2C-SVMs. Application to radar detection. *Signal Processing, 131*, 364-375.

Deepak KumarJain; Surendra Bilouhan Dubey; Rishin Kumar Choubey; Amit Sinhal; Siddharth Kumar Arjari; Amar Jain; Haoxiang Wang. (2018). An approach for hyperspectral image classification by optimizing SVM using self organizing map. *Journal of Computational Science, 25*, 252-259.

Departemen Kesehatan Republik Indonesia. (2007). Pencegahan dan Pemberantasan Demam Berdarah Degue di Indonesia.

G. Kucherov; K. Salikhov ; D. Tsur, . (2014). Approximate String Matching Using a Bidirectional Index. *Symposium on Combinatorial Pattern Matching.* Moscow, Russia.

G. R. Alfonso, E. Isturiz. (2010). Update on the Global Spread of Dengue. *International Journal of Antimicrobial Agents*, 40–42.

Gabriela Czibula, Istvan Gergely Czibula, Adela-Maria Sîrbu, Ioan-Gabriel Mircea. (2015). A novel approach to adaptive relational association rule mining. *Applied Soft Computing , 36*, 519–533.

Guo, J., Hermelin, D., & Komusiewicz, C. (2014). Local search for string problems: Brute-force is essentially optimal. *Theoretical Computer Science, 525*, 30-41.

H.P. Pinheiro,A.S. Pinheiro , P.K. Sen. (2005, march). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference, 130*(1-2), 325-339.

Henning Fernau, Markus L.Schmid. ((2015)). Pattern matching with variables: A multivariate complexity analysis. *Information and Computation, 242*, 287–305.

J. Kopecký, T. Vitvar, C. Bournez and J. Farrell. (2007). SAWSDL: Semantic Annotations for WSDL and XML Schema. *IEEE Internet Computing, 11*(6), 60-67.

J. Park, I. Kim, H. Y. Song. (2017). Construction of parity-check-concatenated polar codes based on minimum Hamming weight codewords. *Electronics Letters, 53*(14), 924-926.

Jianmin Ma; Minh N. Nguyen; Jagath C. Rajapakse. (2009). Gene Classification Using COdon USage and Support Vector Machine. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6*(1), 134-143.

Jian-wei Liu; Li-peng Cui; Xiong-lin Luo. (2016). MCR SVM classifier with group sparsity. *Optik - International Journal for Light and Electron Optics, 127*(17), 6915-6926.

Jing Zhou; Ying Yang; Steven X. Ding; Yanyang Zi; Muheng Wei. (2018). A Fault Detection and Health Monitoring Scheme for SHip Propulsion Systems Using SVM Technique. *IEEE Access, 6*, 16207-16215.

Julie Fredonnet, Julie Foncy, Sophie Lamarre, Jean-Christophe Cau, Emmanuelle Trévisiol. (2013). Dynamic PDMS inking for DNA patterning by soft lithography. *Microelectronic Engineering, 111* , 379–383.

Jun Hu; Yang Li; Ming Zhang; Xibei Yang; Hong-Bin Shen; Dong-Jun Yu. (2017). Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-Based Features and Boosting Multiple SVMs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14*(6), 1389-1398.

Juniastuti et al. (2014). High Rate of Seronegative HCV infection in HIV-Positive Patients. *Biomedical Reports, 2*, 79-84.

K. Kim , M. Kim ,Y.Wooc. (2008). A DNA sequence alignment algorithm using quality information and a fuzzy inference method. *Progress in Natural Science, 18*, 595–602.

Kondo, T. (2014, October). Gradient orientation pattern matching with the Hamming distance. *Pattern Recognition, 47*(10), 3387-3404.

Luis Alberto Hernandez Montiel. (2016). Hybrid Algorithm Applied on Gene Selection and Classification from DIfferent Siseases. *IEEE Latin America Transactions, 14*(2), 930-935.

Luis H.S.Vogado; Rodrigo M.S. Veras; Flavio. H.D. Araujo; Romuere R.V. Silva; Kelson R.T. Aires. (2018). Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. *Engineering Applications of Artificial Intelligence, 72*, 415-422.

M. A. Ebrahimi; M. H. Khoshtaghaza; S. Minaei; B. Jamshidi. (2017). Vision-based pest detection based on SVM classification method. *Computers and Electronics in Agriculture, 137*, 52-58.

M.J. Atallah, T.W. Duket. (2011, july). Pattern matching in the Hamming distance with thresholds. *Information Processing Letters, 111*(14), 674-677.

Marco Capó, Aritz Pérez, Jose A. Lozano. (2017). An efficient approximation to the K-means clustering for massive data. *Knowledge-Based Systems, 117*, 56-69.

Mika Göös,Tuomo Lempiäinen,Eugen Czeizler,Pekka Orponen. (2014). Search methods for tile sets in patterned DNA self-assembly. *Journal of Computer and System Sciences, 80*, 297–319.

Mohammed J. Zaki, Wagner Meira. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms.* London: Cambridge University Press.

Molecular-Evolutionary-Genetics-Analysis. (accessed May 2017). *www.megasoftware.net.*

Neelam Goel, Shailendra Singh, Trilok Chand Aseri. (2015). An improved method for splice site prediction in DNA sequences using support vector machines. *Procedia Computer Science , 57*, 358 – 367.

O.A. Akalu, A. Endale, N. Tesfaye, D. Woldemichael. (2010). Federal HIV/AIDS Prevention and Control Office. Monthly HIV care and ART update. (Ethiopian Ministry of Health).

Pray, L. A. (2008). *Discovery of DNA Nature and Function: Watson and Crick.*

R. Beal, D. Adjeroh. (2015). Efficient pattern matching for RNA secondary structures. *Theoretical Computer Science, 592*, 59-71.

Rupan Panja; Nikhil R. Pal. (2018). MS-SVM: Minimally Spanned Support Vector Machine. *Applied Soft Computing, 64*, 356-365.

S. Cho; J.C. Na; K. Park; J.S. Sim. (2015). A fast algorithm for order-preserving pattern matching. *Information Processing Letters, 115*(2), 397-402.

S. Das; K.Kapoor. (2017). Weighted approximate parameterized string matching. *AKCE International Journal of Graphs and Combinatorics, 14*(1), 1-12.

S. Kim, H. Cho. (2017). Position-restricted approximate string matching with metric Hamming distance. *2017 IEEE International Conference on Big Data and Smart Computing (BigComp).* Jeju.

S. Kundu and B. Ray. (2015). New hamming score based correlation method for fingerprint identification. *2015 6th International Conference on Computers and Devices for Communication (CODEC).* Kolkata.

S. Pissis ,A. Retha. (2015). Generalised Implementation for Fixed-Length Approximate String Matching under Hamming Distance and Applications. *2015 IEEE International Parallel and Distributed Processing Symposium Workshop.* Hyderabad.

S. Tapan and D. Wang. (2016, january). A Further Study on Mining DNA Motifs Using Fuzzy Self-Organizing Maps. *IEEE Transactions on Neural Networks and Learning Systems, 27*(1), 113-124.

Samia Djemai; Belkacem Brahmi; Mohand Ouamer Bibi. (2016). A primal–dual method for SVM training. *Neurocomputing, 211*, 34-40.

Sandra Iurescia, Daniela Fioretti, Vito Michele Fazio,Monica Rinaldi. (2012). Epitope-driven DNA vaccine design employing immunoinformatics against B-cell lymphoma: A biotech's challenge. *Biotechnology Advances , 30* , 372–383.

Sasitorn Plakunmonthon, Nattanan Panjaworayan T-Thienprasert, Kritsada Khongnomnana, Yong Poovorawanc, Sunchai Payungporna. (2014). Computational prediction of hybridization patterns between hepatitisC viral genome and human microRNAs. *Journal of Computational Science, 5*, 327–331.

Saurabh Paul; Malik Magdon-Ismail; Petros Drineas. (2016). Feature selection for linear SVM with provable guarantees. *Pattern Recognition, 60*, 205-214.

Sebastián Maldonado; Julio López. (2018). Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification. *Applied Soft Computing, 67*, 94-105.

Sidheswar Routray; Arun Kumar Ray; Chandrabhanu Mishra; G. Palai. (2018). Efficient hybrid image denoising scheme based on SVM classification. *Optik, 157*, 503-511.

Simmonds.et.al. (2005). Consensus Proposals for a Unified System of. *Hepatology, 42*(4), 962-973.

Singh, A. K. (2016). Error detection and correction by hamming code. *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC).* Jalgaon.

Srinivasareddy Putluri, Md Zia Ur Rahman, Shaik Yasmeen Fathima. (2018). Cloud-based adaptive exon prediction for DNA analysis. *IET Journals & Magazines, 5*(1), 25-30.

Stollberg Michael and Armin Haller. (2005). Semantic Web Service Tutorial. *3rd International Conference on Web Services (ICWS 2005).* Orlando.

T. Boby, A.M. Patch, S.J. Aves. (2005). TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics, 21*(6), 811-816.

T. Feng, T. S. Li , P. Kuo. (2015). Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming. *Applied Mathematical Modelling , 39* , 7401–7419.

Tarek Hamrouni, Sarra Slimani, Faouzi Ben Charrada. (2015). A data mining correlated patterns-based periodic decentralized replication strategy for data grids. *The Journal of Systems and Software, 110*, 10–27.

Toshiaki Takayanagi. (2013). Modeling chronic hepatitis B or C virus infection during antiviral therapy usingananalogy to enzyme kinetics: Long-term viral dynamics without rebound and oscillation. *Computers in Biology and Medicine, 43*, 2021–2027.

(t.thn.). *University College Dublin, Multiple Sequence Alignment, www.clustal.org.*

V.G., R. (2010). Dengue conundrums. *International Journal of Antimicrobial Agents;*, 36-39.

W. Shan, S. Zhang and Y. He. (2017). Machine learning based side-channel-attack countermeasure with hamming-distance redistribution and its application on advanced encryption standard. *Electronics Letters, 53*(14), 926-928.

WHO. (2017). *Global Hepatitis Report 2017.* Geneva: Licence: CC BY-NC-SA 3.0 IGO.

World-Gene-Bank. (accessed May, 2017). *https://blast.ncbi.nlm.nih.gov/Blast.cgi.*

World-Health-Organization. (2017). *Global Hepatitis Report 2017.* Executive Summary, World Health Organization, Geneva.

(accessed April 2017). *www.genetyx.co.jp.*

Y. Chen; Y. Hu. (2006). Constraint-based sequential pattern mining: The consideration of recency and compactness. *Decision Support Systems, 42*(2), 1203-1215.

Yong Liu; Shizhong Liao. (2017). Granularity selection for cross-validation of SVM. *Information Sciences, 378*, 475-483.

LAMPIRAN 1

1. Contoh DNA record , sumber: Genomic World Bank:

Hepatitis C virus (isolate JFH-1) genomic RNA, complete genome

GenBank: AB047639.1

FASTA Graphics

Go to:

```
LOCUS       AB047639               9678 bp    RNA     linear   VRL
12-NOV-2005
DEFINITION  Hepatitis C virus (isolate JFH-1) genomic RNA, complete
genome.
ACCESSION   AB047639
VERSION     AB047639.1  GI:13122261
KEYWORDS    .
SOURCE      Hepatitis C virus JFH-1
  ORGANISM  Hepatitis C virus JFH-1
            Viruses; ssRNA viruses; ssRNA positive-strand viruses,
no DNA
            stage; Flaviviridae; Hepacivirus.
REFERENCE   1
  AUTHORS   Kato,T., Furusaka,A., Miyamoto,M., Date,T., Yasui,K.,
Hiramoto,J.,
            Nagayama,K., Tanaka,T. and Wakita,T.
  TITLE     Sequence analysis of hepatitis C virus isolated from a
fulminant
            hepatitis patient
  JOURNAL   J. Med. Virol. 64 (3), 334-339 (2001)
   PUBMED   11424123
REFERENCE   2  (bases 1 to 9678)
  AUTHORS   Kato,T., Wakita,T. and Furusaka,A.
  TITLE     Direct Submission
  JOURNAL   Submitted (23-AUG-2000) Takanobu Kato, The Tokyo
Metropolitan
            Institute for Neuroscience, Department of Microbiology;
Musashidai
            2-6, Fuchu, Tokyo 183-8526, Japan (E-
mail:takato@tmin.ac.jp,
```

Tel:81-423-25-3881(ex.4605), Fax:81-423-21-8678)

FEATURES Location/Qualifiers

source 1..9678

/organism="Hepatitis C virus JFH-1"

/mol_type="genomic RNA"

/db_xref="taxon:356411"

/clone="JFH-1"

/note="isolated from fulminant hepatitis patient/genotype 2a"

CDS 341..9442

/codon_start=1

/product="polyprotein"

/protein_id="BAB32872.1"

/db_xref="GI:13122262"

/translation="MSTNPKPQRKTKRNTNRRPEDVKFPGGGQIVGGVYLLPRRGPRL

GVRTTRKTSERSQPRGRRQPIPKDRRSTGKAWGKPGRPWPLYGNEGLGWAGWLLSPRG

SRPSWGPTDPRHRSRNVGKVIDTLTCGFADLMGYIPVVGAPLSGAARAVAHGVRVLED

GVNYATGNLPGFPFSIFLLALLSCITVPVSAAQVKNTSSSYMVTNDCSNDSITWQLEA

AVLHVPGCVPCERVGNTSRCWVPVSPNMAVRQPGALTQGLRTHIDMVVMSATFCSALY

VGDLCGGVMLAAQVFIVSPQYHWFVQECNCSIYPGTITGHRMAWDMMMNWSPTATMIL

AYVMRVPEVIIDIVSGAHWGVMFGLAYFSMQGAWAKVIVILLLAAGVDAGTTTVGGAV

ARSTNVIAGVFSHGPQQNIQLINTNGSWHINRTALNCNDSLNTGFLAALFYTNRFNSS

GCPGRLSACRNIEAFRIGWGTLQYEDNVTNPEDMRPYCWHYPPKPCGVVPARSVCGPV

YCFTPSPVVVGTTDRRGVPTYTWGENETDVFLLNSTRPPQGSWFGCTWMNSTGFTKTC

GAPPCRTRADFNASTDLLCPTDCFRKHPDATYIKCGSGPWLTPKCLVHYPYRLWHYPC

TVNFTIFKIRMYVGGVEHRLTAACNFTRGDRCDLEDRDRSQLSPLLHSTTEWAILPCT

YSDLPALSTGLLHLHQNIVDVQYMYGLSPAITKYVVRWEWVVLLFLLLADARVCACLW

MLILLGQAEAALEKLVVLHAASAANCHGLLYFAIFFVAAWHIRGRVVPLTTYCLTGLW

PFCLLLMALPRQAYAYDAPVHGQIGVGLLILITLFTLTPGYKTLLGQCLWWLCYLLTL

GEAMIQEWVPPMQVRGGRDGIAWAVTIFCPGVVFDITKWLLALLGPAYLLRAALTHVP

YFVRAHALIRVCALVKQLAGGRYVQVALLALGRWTGTYIYDHLTPMSDWAASGLRDLA

VAVEPIIFSPMEKKVIVWGAETAACGDILHGLPVSARLGQEILLGPADGYTSKGWKLL

APITAYAQQTRGLLGAIVVSMTGRDRTEQAGEVQILSTVSQSFLGTTISGVLWTVYHG

AGNKTLAGLRGPVTQMYSSAEGDLVGWPSPPGTKSLEPCKCGAVDLYLVTRNADVIPA

RRRGDKRGALLSPRPISTLKGSSGGPVLCPRGHVVGLFRAAVCSRGVAKSIDFIPVET

LDVVTRSPTFSDNSTPPAVPQTYQVGYLHAPTGSGKSTKVPVAYAAQGYKVLVLNPSV

AATLGFGAYLSKAHGINPNIRTGVRTVMTGEAITYSTYGKFLADGGCASGAYDIIICD

ECHAVDATSILGIGTVLDQAETAGVRLTVLATATPPGSVTTPHPDIEEVGLGREGEIP

FYGRAIPLSCIKGGRHLIFCHSKKKCDELAAALRGMGLNAVAYYRGLDVSIIPAQGDV

VVVATDALMTGYTGDFDSVIDCNVAVTQAVDFSLDPTFTITTQTVPQDAVSRSQRRGR

TGRGRQGTYRYVSTGERASGMFDSVVLCECYDAGAAWYDLTPAETTVRLRAYFNTPGL

PVCQDHLEFWEAVFTGLTHIDAHFLSQTKQAGENFAYLVAYQATVCARAKAPPPSWDA

MWKCLARLKPTLAGPTPLLYRLGPITNEVTLTHPGTKYIATCMQADLEVMTSTWVLAG

GVLAAVAAYCLATGCVSIIGRLHVNQRVVVAPDKEVLYEAFDEMEECASRAALIEEGQ

RIAEMLKSKIQGLLQQASKQAQDIQPAMQASWPKVEQFWARHMWNFISGIQYLAGLST

LPGNPAVASMMAFSAALTSPLSTSTTILLNIMGGWLASQIAPPAGATGFVVSGLVGAA

VGSIGLGKVLVDILAGYGAGISGALVAFKIMSGEKPSMEDVINLLPGILSPGALVVGV

ICAAILRRHVGPGEGAVQWMNRLIAFASRGNHVAPTHYVTESDASQRVTQLLGSLTIT

SLLRRLHNWITEDCPIPCSGSWLRDVWDWVCTILTDFKNWLTSKLFPKLPGLPFISCQ

KGYKGVWAGTGIMTTRCPCGANISGNVRLGSMRITGPKTCMNTWQGTFPINCYTEGQC

APKPPTNYKTAIWRVAASEYAEVTQHGSYSYVTGLTTDNLKIPCQLPSPEFFSWVDGV

QIHRFAPTPKPFFRDEVSFCVGLNSYAVGSQLPCEPEPDADVLRSMLTDPPHITAETA

ARRLARGSPPSEASSSVSQLSAPSLRATCTTHSNTYDVDMVDANLLMEGGVAQTEPES

RVPVLDFLEPMAEEESDLEPSIPSECMLPRSGFPRALPAWARPDYNPPLVESWRRPDY

QPPTVAGCALPPPKKAPTPPPRRRRTVGLSESTISEALQQLAIKTFGQPPSSGDAGSS

TGAGAAESGGPTSPGEPAPSETGSASSMPPLEGEPGDPDLESDQVELQPPPQGGGVAP

GSGSGSWSTCSEEDDTTVCCSMSYSWTGALITPCSPEEEKLPINPLSNSLLRYHNKVY

CTTSKSASQRAKKVTFDRTQVLDAHYDSVLKDIKLAASKVSARLLTLEEACQLTPPHS

ARSKYGFGAKEVRSLSGRAVNHIKSVWKDLLEDPQTPIPTTIMAKNEVFCVDPAKGGK

KPARLIVYPDLGVRVCEKMALYDITQKLPQAVMGASYGFQYSPAQRVEYLLKAWAEKK

DPMGFSYDTRCFDSTVTERDIRTEESIYQACSLPEEARTAIHSLTERLYVGGPMFNSK

GQTCGYRRCRASGVLTTSMGNTITCYVKALAACKAAGIVAPTMLVCGDDLVVISESQG

TEEDERNLRAFTEAMTRYSAPPGDPPRPEYDLELITSCSSNVSVALGPRGRRRYYLTR

DPTTPLARAAWETVRHSPINSWLGNIIQYAPTIWVRMVLMTHFFSILMVQDTLDQNLN

FEMYGSVYSVNPLDLPAIIERLHGLDAFSMHTYSHHELTRVASALRKLGAPPLRVWKS

RARAVRASLISRGGKAAVCGRYLFNWAVKTKLKLTPLPEARLLDLSSWFTVGAGGGDI
FHSVSRARPRSLLFGLLLLFVGVGLFLLPAR"

ORIGIN

```
        1 acctgcccct aataggggcg acactccgcc atgaatcact cccctgtgag
gaactactgt
       61 cttcacgcag aaagcgccta gccatggcgt tagtatgagt gtcgtacagc
ctccaggccc
      121 ccccctcccg ggagagccat agtggtctgc ggaaccggtg agtacaccgg
aattgccggg
      181 aagactgggt cctttcttgg ataaacccac tctatgcccg gccatttggg
cgtgcccccg
      241 caagactgct agccgagtag cgttgggttg cgaaaggcct tgtggtactg
cctgataggg
      301 cgcttgcgag tgccccggga ggtctcgtag accgtgcacc atgagcacaa
atcctaaacc
      361 tcaaagaaaa accaaaagaa acaccaaccg tcgcccagaa gacgttaagt
tcccgggcgg
      421 cggccagatc gttggcggag tatacttgtt gccgcgcagg ggccccaggt
tgggtgtgcg
      481 cacgacaagg aaaacttcgg agcggtccca gccacgtggg agacgccagc
ccatccccaa
      541 agatcggcgc tccactggca aggcctgggg aaaaccaggt cgcccctggc
ccctatatgg
      601 gaatgaggga ctcggctggg caggatggct cctgtccccc cgaggctctc
gcccctcctg
      661 gggccccact gacccccggc ataggtcgcg caacgtgggt aaagtcatcg
acaccctaac
      721 gtgtggcttt gccgacctca tggggtacat ccccgtcgta ggcgccccgc
ttagtggcgc
      781 cgccagagct gtcgcgcacg gcgtgagagt cctggaggac ggggttaatt
atgcaacagg
      841 gaacctaccc ggtttccccc tttctatctt cttgctggcc ctgttgtcct
gcatcaccgt
      901 tccggtctct gctgcccagg tgaagaatac cagtagcagc tacatggtga
ccaatgactg
```

961 ctccaatgac agcatcactt ggcagctcga ggctgcggtt ctccacgtcc ccgggtgcgt

1021 cccgtgcgag agagtgggga atacgtcacg gtgttgggtg ccagtctcgc caaacatggc

1081 tgtgcggcag cccggtgccc tcacgcaggg tctgcggacg cacatcgata tggttgtgat

1141 gtccgccacc ttctgctctg ctctctacgt gggggacctc tgtggcgggg tgatgctcgc

1201 ggcccaggtg ttcatcgtct cgccgcagta ccactggttt gtgcaagaat gcaattgctc

1261 catctaccct ggcaccatca ctggacaccg catggcatgg gacatgatga tgaactggtc

1321 gcccacggcc accatgatcc tggcgtacgt gatgcgcgtc cccgaggtca tcatagacat

1381 cgttagcggg gctcactggg gcgtcatgtt cggcttggcc tacttctcta tgcagggagc

1441 gtgggcgaag gtcattgtca tccttctgct ggccgctggg gtggacgcgg gcaccaccac

1501 cgttggaggc gctgttgcac gttccaccaa cgtgattgcc ggcgtgttca gccatggccc

1561 tcagcagaac attcagctca ttaacaccaa cggcagttgg cacatcaacc gtactgcctt

1621 gaattgcaat gactccttga acaccggctt tctcgcggcc ttgttctaca ccaaccgctt

1681 taactcgtca gggtgtccag ggcgcctgtc cgcctgccgc aacatcgagg ctttccggat

1741 agggtggggc accctacagt acgaggataa tgtcaccaat ccagaggata tgaggccgta

1801 ctgctggcac taccccccaa agccgtgtgg cgtagtcccc gcgaggtctg tgtgtggccc

1861 agtgtactgt ttcaccccca gcccggtagt agtgggcacg accgacagac gtggagtgcc

1921 cacctacaca tggggagaga atgagacaga tgtcttccta ctgaacagca cccgaccgcc

1981 gcagggctca tggttcggct gcacgtggat gaactccact ggtttcacca agacttgtgg

2041 cgcgccacct tgccgcacca gagctgactt caacgccagc acggacttgt tgtgccctac

2101 ggattgtttt aggaagcatc ctgatgccac ttatattaag tgtggttctg ggccctggct

2161 cacaccaaag tgcctggtcc actaccctta cagactctgg cattacccct gcacagtcaa

2221 ttttaccatc ttcaagataa gaatgtatgt agggggggtt gagcacaggc tcacggccgc

2281 atgcaacttc actcgtgggg atcgctgcga cttggaggac agggacagga gtcagctgtc

2341 tcctctgttg cactctacca cggaatgggc catcctgccc tgcacctact cagacttacc

2401 cgctttgtca actggtcttc tccaccttca ccagaacatc gtggacgtac aatacatgta

2461 tggcctctca cctgctatca caaaatacgt cgttcgatgg gagtgggtgg tactcttatt

2521 cctgctctta gcggacgcca gagtctgcgc ctgcttgtgg atgctcatct tgttgggcca

2581 ggccgaagca gcattggaga agttggtcgt cttgcacgct gcgagtgcgg ctaactgcca

2641 tggcctccta tattttgcca tcttcttcgt ggcagcttgg cacatcaggg gtcgggtggt

2701 ccccttgacc acctattgcc tcactggcct atggcccttc tgcctactgc tcatggcact

2761 gccccggcag gcttatgcct atgacgcacc tgtgcacgga cagataggcg tgggtttgtt

2821 gatattgatc accctcttca cactcacccc ggggtataag accctcctcg gccagtgtct

2881 gtggtggttg tgctatctcc tgaccctggg ggaagccatg attcaggagt gggtaccacc

2941 catgcaggtg cgcggcggcc gcgatggcat cgcgtgggcc gtcactatat tctgcccggg

3001 tgtggtgttt gacattacca aatggctttt ggcgttgctt gggcctgctt acctcttaag

3061 ggccgctttg acacatgtgc cgtacttcgt cagagctcac gctctgataa gggtatgcgc

3121 tttggtgaag cagctcgcgg ggggtaggta tgttcaggtg gcgctattgg cccttggcag

3181 gtggactggc acctacatct atgaccacct cacacctatg tcggactggg ccgctagcgg

3241 cctgcgcgac ttagcggtcg ccgtggaacc catcatcttc agtccgatgg agaagaaggt

3301 catcgtctgg ggagcggaga cggctgcatg tggggacatt ctacatggac ttcccgtgtc

3361 cgcccgactc ggccaggaga tcctcctcgg cccagctgat ggctacacct ccaaggggtg

3421 gaagctcctt gctcccatca ctgcttatgc ccagcaaaca cgaggcctcc tgggcgccat

3481 agtggtgagt atgacggggc gtgacaggac agaacaggcc ggggaagtcc aaatcctgtc

3541 cacagtctct cagtccttcc tcggaacaac catctcgggg gttttgtgga ctgtttacca

3601 cggagctggc aacaagactc tagccggctt acggggtccg gtcacgcaga tgtactcgag

3661 tgctgagggg gacttggtag gctggcccag ccccccctggg accaagtctt tggagccgtg

3721 caagtgtgga gccgtcgacc tatatctggt cacgcggaac gctgatgtca tcccggctcg

3781 gagacgcggg gacaagcggg gagcattgct ctccccgaga cccatttcga ccttgaaggg

3841 gtcctcgggg gggccggtgc tctgccctag gggccacgtc gttgggctct tccgagcagc

3901 tgtgtgctct cggggcgtgg ccaaatccat cgatttcatc cccgttgaga cactcgacgt

3961 tgttacaagg tctcccactt tcagtgacaa cagcacgcca ccggctgtgc cccagaccta

4021 tcaggtcggg tacttgcatg ctccaactgg cagtggaaag agcaccaagg tccctgtcgc

4081 gtatgccgcc caggggtaca aagtactagt gcttaacccc tcggtagctg ccaccctggg

4141 gtttggggcg tacctatcca aggcacatgg catcaatccc aacattagga ctggagtcag

4201 gaccgtgatg accggggagg ccatcacgta ctccacatat ggcaaatttc tcgccgatgg

4261 gggctgcgct agcggcgcct atgacatcat catatgcgat gaatgccacg ctgtggatgc

4321 tacctccatt ctcggcatcg gaacggtcct tgatcaagca gagacagccg gggtcagact

4381 aactgtgctg gctacggcca cacccccggg gtcagtgaca acccccccatc ccgatataga

4441 agaggtaggc ctcgggcggg agggtgagat ccccttctat gggagggcga ttcccctatc

4501 ctgcatcaag ggagggagac acctgatttt ctgccactca aagaaaaagt gtgacgagct

4561 cgcggcggcc cttcggggca tgggcttgaa tgccgtggca tactatagag ggttggacgt

4621 ctccataata ccagctcagg gagatgtggt ggtcgtcgcc accgacgccc tcatgacggg

4681 gtacactgga gactttgact ccgtgatcga ctgcaatgta gcggtcaccc aagctgtcga

4741 cttcagcctg gaccccacct tcactataac cacacagact gtcccacaag acgctgtctc

4801 acgcagtcag cgccgcgggc gcacaggtag aggaagacag ggcacttata ggtatgtttc

4861 cactggtgaa cgagcctcag gaatgtttga cagtgtagtg ctttgtgagt gctacgacgc

4921 aggggctgcg tggtacgatc tcacaccagc ggagaccacc gtcaggctta gagcgtattt

4981 caacacgccc ggcctacccg tgtgtcaaga ccatcttgaa ttttgggagg cagttttcac

5041 cggcctcaca cacatagacg cccacttcct ctcccaaaca aagcaagcgg gggagaactt

5101 cgcgtaccta gtagcctacc aagctacggt gtgcgccaga gccaaggccc ctccccccgtc

5161 ctgggacgcc atgtggaagt gcctggcccg actcaagcct acgcttgcgg gccccacacc

5221 tctcctgtac cgtttgggcc ctattaccaa tgaggtcacc ctcacacacc ctgggacgaa

5281 gtacatcgcc acatgcatgc aagctgacct tgaggtcatg accagcacgt gggtcctagc

5341 tggaggagtc ctggcagccg tcgccgcata ttgcctggcg actggatgcg tttccatcat

5401 cggccgcttg cacgtcaacc agcgagtcgt cgttgcgccg gataaggagg tcctgtatga

5461 ggcttttgat gagatggagg aatgcgcctc tagggcggct ctcatcgaag aggggcagcg

5521 gatagccgag atgttgaagt ccaagatcca aggcttgctg cagcaggcct ctaagcaggc

5581 ccaggacata caacccgcta tgcaggcttc atggcccaaa gtggaacaat tttgggccag

5641 acacatgtgg aacttcatta gcggcatcca atacctcgca ggattgtcaa cactgccagg

5701 gaaccccgcg gtggcttcca tgatggcatt cagtgccgcc ctcaccagtc cgttgtcgac

5761 cagtaccacc atccttctca acatcatggg aggctggtta gcgtcccaga tcgcaccacc

5821 cgcgggggcc accggctttg tcgtcagtgg cctggtgggg gctgccgtgg gcagcatagg

5881 cctgggtaag gtgctggtgg acatcctggc aggatatggt gcgggcattt cgggggccct

5941 cgtcgcattc aagatcatgt ctggcgagaa gccctctatg gaagatgtca tcaatctact

6001 gcctgggatc ctgtctccgg gagccctggt ggtgggggtc atctgcgcgg ccattctgcg

6061 ccgccacgtg ggaccggggg agggcgcggt ccaatggatg aacaggctta ttgcctttgc

6121 ttccagagga aaccacgtcg cccctactca ctacgtgacg gagtcggatg cgtcgcagcg

6181 tgtgacccaa ctacttggct ctcttactat aaccagccta ctcagaagac tccacaattg

6241 gataactgag gactgcccca tcccatgctc cggatcctgg ctccgcgacg tgtgggactg

6301 ggtttgcacc atcttgacag acttcaaaaa ttggctgacc tctaaattgt tccccaagct

6361 gcccggcctc cccttcatct cttgtcaaaa ggggtacaag ggtgtgtggg ccggcactgg

6421 catcatgacc acgcgctgcc cttgcggcgc caacatctct ggcaatgtcc gcctgggctc

6481 tatgaggatc acagggccta aaacctgcat gaacacctgg caggggacct ttcctatcaa

6541 ttgctacacg gagggccagt gcgcgccgaa acccccccacg aactacaaga ccgccatctg

6601 gagggtggcg gcctcggagt acgcggaggt gacgcagcat gggtcgtact cctatgtaac

6661 aggactgacc actgacaatc tgaaaattcc ttgccaacta ccttctccag agtttttctc

6721 ctgggtggac ggtgtgcaga tccataggtt tgcacccaca ccaaagccgt ttttccggga

6781 tgaggtctcg ttctgcgttg ggcttaattc ctatgctgtc gggtcccagc ttccctgtga

6841 acctgagccc gacgcagacg tattgaggtc catgctaaca gatccgcccc acatcacggc

6901 ggagactgcg gcgcggcgct tggcacgggg atcacctcca tctgaggcga gctcctcagt

6961 gagccagcta tcagcaccgt cgctgcgggc cacctgcacc acccacagca acacctatga

7021 cgtggacatg gtcgatgcca acctgctcat ggagggcggt gtggctcaga cagagcctga

7081 gtccagggtg cccgttctgg actttctcga gccaatggcc gaggaagaga gcgaccttga

7141 gccctcaata ccatcggagt gcatgctccc caggagcggg tttccacggg ccttaccggc

7201 ttgggcacgg cctgactaca acccgccgct cgtggaatcg tggaggaggc cagattacca

7261 accgcccacc gttgctggtt gtgctctccc ccccccaag aaggccccga cgcctccccc

7321 aaggagacgc cggacagtgg gtctgagcga gagcaccata tcagaagccc tccagcaact

7381 ggccatcaag acctttggcc agccccctc gagcggtgat gcaggctcgt ccacgggggc

7441 gggcgccgcc gaatccggcg gtccgacgtc ccctggtgag ccggcccct cagagacagg

7501 ttccgcctcc tctatgcccc ccctcgaggg ggagcctgga gatccggacc tggagtctga

7561 tcaggtagag cttcaacctc ccccccaggg gggggggta gctcccggtt cgggctcggg

7621 gtcttggtct acttgctccg aggaggacga taccaccgtg tgctgctcca tgtcatactc

7681 ctggaccggg gctctaataa ctccctgtag ccccgaagag gaaaagttgc caatcaaccc

7741 tttgagtaac tcgctgttgc gataccataa caaggtgtac tgtacaacat caaagagcgc

7801 ctcacagagg gctaaaaagg taacttttga caggacgcaa gtgctcgacg cccattatga

7861 ctcagtctta aaggacatca agctagcggc ttccaaggtc agcgcaaggc tcctcacctt

7921 ggaggaggcg tgccagttga ctccacccca ttctgcaaga tccaagtatg gattcggggc

7981 caaggaggtc cgcagcttgt ccgggagggc cgttaaccac atcaagtccg tgtggaagga

8041 cctcctggaa gacccacaaa caccaattcc cacaaccatc atggccaaaa atgaggtgtt

8101 ctgcgtggac cccgccaagg ggggtaagaa accagctcgc ctcatcgttt accctgacct

8161 cggcgtccgg gtctgcgaga aaatggccct ctatgacatt acacaaaagc ttcctcaggc

8221 ggtaatggga gcttcctatg gcttccagta ctcccctgcc caacgggtgg agtatctctt

8281 gaaagcatgg gcggaaaaga aggaccccat gggtttttcg tatgataccc gatgcttcga

8341 ctcaaccgtc actgagagag acatcaggac cgaggagtcc atataccagg cctgctccct

8401 gcccgaggag gcccgcactg ccatacactc gctgactgag agactttacg taggagggcc

8461 catgttcaac agcaagggtc aaacctgcgg ttacagacgt tgccgcgcca gcggggtgct

8521 aaccactagc atgggtaaca ccatcacatg ctatgtgaaa gccctagcgg cctgcaaggc

8581 tgcggggata gttgcgccca caatgctggt atgcggcgat gacctagtag tcatctcaga

8641 aagccagggg actgaggagg acgagcggaa cctgagagcc ttcacggagg ccatgaccag

8701 gtactctgcc cctcctggtg atccccccag accggaatat gacctggagc taataacatc

8761 ctgttcctca aatgtgtctg tggcgttggg cccgcggggc cgccgcagat actacctgac

8821 cagagaccca accactccac tcgcccgggc tgcctgggaa acagttagac actcccctat

8881 caattcatgg ctgggaaaca tcatccagta tgctccaacc atatgggttc gcatggtcct

8941 aatgacacac ttcttctcca ttctcatggt ccaagacacc ctggaccaga acctcaactt

9001 tgagatgtat ggatcagtat actccgtgaa tcctttggac cttccagcca taattgagag

9061 gttacacggg cttgacgcct tttctatgca cacatactct caccacgaac tgacgcgggt

9121 ggcttcagcc ctcagaaaac ttggggcgcc acccctcagg gtgtggaaga gtcgggctcg

9181 cgcagtcagg gcgtccctca tctcccgtgg agggaaagcg gccgtttgcg gccgatatct

9241 cttcaattgg gcggtgaaga ccaagctcaa actcactcca ttgccggagg cgcgcctact

9301 ggacttatcc agttggttca ccgtcggcgc cggcgggggc gacatttttc acagcgtgtc

9361 gcgcgcccga ccccgctcat tactcttcgg cctactccta cttttcgtag gggtaggcct

9421 cttcctactc cccgctcggt agagcggcac acactaggta cactccatag ctaactgttc

9481 ctttttttt ttttttttt ttttttttt ttttttttt ttttctttt ttttttttc

9541 cctctttctt cccttctcat cttattctac tttctttctt ggtggctcca tcttagccct

9601 agtcacggct agctgtgaaa ggtccgtgag ccgcatgact gcagagagtg ccgtaactgg

9661 tctctctgca gatcatgt…

//

Halaman ini sengaja dikosongkan

LAMPIRAN 2

DAFTAR RIWAYAT HIDUP

Nama	:	Berlian Al Kindhi
Tempat/Tanggal Lahir	:	Ponorogo/04 Desember 1985
Pekerjaan	:	Dosen
Alamat Kantor	:	Departemen Teknik Elektro Otomasi, Fakultas Vokasi, Institut Teknologi Sepuluh Nopember
Alamat Rumah	:	Puri Kencana Karah B16, Ketintang, Surabaya, 60232
Nomor HP	:	08 2131 6232 16
E-mail	:	berlian.alkindhi@gmail.com/berlian@its.ac.id

Riwayat Pendidikan Terakhir

- D4 Jurusan Teknik Informatika – Politeknik Elektronika Negeri Surabaya (PENS), Surabaya, Lulus tahun 2011 (IPK. 3.65)
- S2, Jurusan Teknik Informatika – Sekolah Teknik Elektro dan Informatika (STEI), Institut Teknologi Bandung (ITB), Lulus tahun 2013 (IPK. 3.81)
- S3 Departemen Teknik Elektro - Fakultas Teknologi Elektro – Institut Teknologi Sepuluh Nopember (ITS), Mulai kuliah tahun 2015

Daftar Publikasi selama studi doktor

A. Jurnal Internasional:

1. **Berlian Al Kindhi**, Tri Arief Sardjono, Mauridhi Hery Purnomo, Gijsbertus Jacob Verkerke, "Hybrid K-Means, Fuzzy C-Means, and Hierarchical *Clustering* for DNA Hepatitis C Virus Trend Mutation Analysis", *International Journal Expert System with Application (ESwA)*..Vol. 121, hal. 373-381, 2019, ISSN: 0957-4174
   Quartile: Q1, Impact Factor: 3.768, H Index: 145
   DOI: https://doi.org/10.1016/j.eswa.2018.12.019

B. Jurnal Nasional Terakreditasi:

1. **Berlian Al Kindhi**, Tri Arief Sardjono, Mauridhi Hery Purnomo "Optimasi Support Vector Machine untuk memprediksi adanya mutasi pada DNA Hepatitis C Virus", *Jurnal Nasional Teknik Elektro dan Teknologi Informasi, Volume 7 No. 3, Agustus 2018,* ISSN: 2460-5719

C. Seminar Internasional

1. **Berlian Al Kindhi**, Tri Arief Sardjono, "Pattern Matching Performance Comparisons as Big Data Analysis Recommendations for Hepatitis C Virus (HCV) Sequence DNA", 3rd IEEE Artificial Intelligence and Modelling System, Kota Kinabalu, Malaysia, 2-4 Desember 2015, DOI: 10.1109/AIMS.2015.27
2. **Berlian Al Kindhi**, M. A. Hendrawan, D. Purwitasari, T. A. Sardjono, M. H. Purnomo, "Distance-based pattern matching of DNA sequences for evaluating primary mutation", IEEE 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Jogjakarta, Indonesia, 1-2 November 2017, DOI: 10.1109/ICITISEE.2017.8285518
3. **Berlian Al Kindhi**, Tri Arief Sardjono, Mauridhi Hery Purnomo, "Prototype Infrastructure Cloud Expert System DNA Analysis (CESDA) as the Basis of Sustainability DNA Software Improvement in Indonesia", IEEE European Modelling Symposium (EMS), Manchester, United of Kingdom, 20-21 November 2017, DOI: 10.1109/EMS.2017.43

Hibah Penelitian dan Beasiswa selama studi doktor

1. Beasiswa PKPI, *Enhanching International Publication (EIP),* di University Medical Center Groningen (UMCG), Universitas Groningen, Groningen, Belanda, tahun 2017
2. Hibah Penelitian PUPT Tahun I (2017), judul: "Sistem Pakar Analisis DNA Hepatitis Berdasarkan *Clustering Sequence* terhadap *Centroid Primer* Sebagai Dasar Evaluasi Mutasi Genetik"
3. Hibah Penelitian PTUPT Tahun II (2018) dan III (2019), judul: "Sistem Pakar Analisis DNA Hepatitis Berdasarkan *Clustering Sequence* terhadap *Centroid Primer* Sebagai Dasar Evaluasi Mutasi Genetik"

Sertifikasi/Pelatihan Internasional/Nasional selama studi doktor:

1. **Software Development Fundamentals,** Passing Score 94 (minimum passing score 70), Microsoft International Certification, 2016, International Certiport license code: **hsxP-FVKn**
2. **Deep Learning on Image Classification with DIGITS**, Deep Learning Institute, NVDIA Europe, Groningen, Netherlands, Issued Date 11/21/2017
3. **Asesor Kompetensi**, Badan Nasional Sertifikasi Profesi (BNSP), No. Sertifikat 9300 2419 0074408 2018