KLASIFIKASI POST TRANSLATIONAL MODIFICATION N GLIKOSILASI PADA SEQUENCE PROTEIN MENGGUNAKAN ALGORITMA *RANDOM FOREST*

(Skripsi)

Oleh

SUCI HIKMAWATI
1817051033

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

KLASIFIKASI POST TRANSLATIONAL MODIFICATION N GLIKOSILASI PADA SEQUENCE PROTEIN MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Oleh

Suci Hikmawati

Skripsi

Sebagai Salah Satu Syarat untuk Memperoleh Gelar
SARJANA ILMU KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam

ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRAK

KLASIFIKASI POST TRANSLATIONAL MODIFICATION N GLIKOSILASI PADA SEQUENCE PROTEIN MENGGUNAKAN ALGORITMA *RANDOM FOREST*

Oleh

Suci Hikmawati

Modifikasi protein merupakan protein yang strukturnya mengalami perubahan. Glikosilasi adalah salah satu modifikasi protein yang paling rumit dalam sel eukariotik. Glikosilasi berperan penting dalam fungsi biologis, seperti pengenalan antigen, komunikasi sel-sel, ekspresi gen, dan pelipatan protein. Dalam melakukan identifikasi situs glikosilasi pada sekuens protein dapat dilakukan secara eksperimental, namun merupakan tantangan yang sulit karena memerlukan waktu yang lama dan biaya yang mahal. Beberapa peneliti mengusulkan menggunakan metode komputasi untuk identifikasi situs glikosilasi, sehingga pada penelitian ini dilakukan pengklasifikasian untuk mengidentifikasi situs glikosilasi pada sekuens protein dengan menggunakan metode pendekatan *machine learning* yaitu algoritma *random forest.* Data yang digunakan adalah situs n glikosilasi yang terdiri dari data positif berjumlah 11.601 sekuens dan data negatif 12.160 sekuens. Tahapan yang dilakukan pada penelitian ini, diantaranya : cleaning data yaitu menghapus sekuens yang tidak termasuk ke dalam jenis asam amino, fitur yang berkaitan dengan situs modifikasi pasca translasi di ekstraksi, fitur tersebut terdiri dari *statistical moment* dan fitur *position and composition* dengan total fitur berjumlah 153, pemodelan random forest, hingga pengujian klasifikasi menggunakan 10-fold cross validation dan confusion matrix. Setelah dilakukan beberapa percobaan, diperoleh hasil pengujian klasifikasi random forest dengan akurasi tertinggi sebesar 92,29%, sensitivitas 97,71%, spesifisitas 87,12%, serta MCC sebesar 85,12%.

Kata kunci: klasifikasi, machine learning, glikosilasi, ekstraksi fitur, *random forest.*

ABSTRACT

CLASSIFICATION OF POST-TRANSLATIONAL MODIFICATIONS N-GLYCOSYLATION ON PROTEIN SEQUENCES USING RANDOM FOREST ALGORITHM

By

Suci Hikmawati

Modified protein is a protein whose structure has changed. Glycosylation is one of the most complex protein modifications in eukaryotic cells. Glycosylation plays an important role in biological functions, such as antigen recognition, cell-cell communication, gene expression, and protein folding. In identifying glycosylation sites on protein sequences can be done experimentally, but it is a difficult challenge because it requires a long time and is expensive. Several researchers have proposed using computational methods to identify glycosylation sites, so that in this study a classifier was carried out to identify glycosylation sites in protein sequences using a machine learning approach, namely the random forest algorithm. The data used is site n glycosylation consisting of positive data totaling 11,601 sequences and negative data 12,160 sequences. The stages carried out in this study, including: cleaning data, namely: delete sequences that do not belong to this type of amino acid, features related to post-translation modification sites are extracted, these features consist of statistical moment and position and composition features with a total of 153 features, random forest modeling, to classification testing using 10-fold cross validation and confusion matrix. After several experiments, the results of the random forest classification test were obtained with the highest accuracy of 92.29%, sensitivity of 97.71%, specificity of 87.12%, and MCC of 85.12%.

Keywords: classification, machine learning, glycosylation, feature extraction, random forest.

Judul Skripsi	: **KLASIFIKASI POST TRANSLATIONAL MODIFICATION N GLIKOSILASI PADA SEQUENCE PROTEIN MENGGUNAKAN ALGORITMA *RANDOM FOREST***
Nama Mahasiswa	: Suci Hikmawati
Nomor Pokok Mahasiswa	: 1817051033
Program Studi	: Ilmu Komputer
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Favorisen R. Lumbanraja, Ph.D.
NIP 19830110 200812 1 002

2. Ketua Jurusan Ilmu Komputer

Didik Kurniawan, S.Si., M.T.
NIP 19800419 200501 1 004

MENGESAHKAN

1. Tim Penguji

Ketua : **Favorisen R. Lumbanraja, Ph.D.** .....

Penguji I
Penguji Pembahas : **M. Reza Faisal, S.T., M.T., Ph.D.** .....

Penguji II
Penguji Pembahas : **Dr. Ir. Kurnia Muludi, M.S.Sc.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T.
NIP 19740705 200003 1 001

Tanggal Lulus Ujian Skripsi : **28 September 2022**

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama: Suci Hikmawati
NPM: 1817051033

Dengan ini menyatakan bahwa skripsi saya yang berjudul "KLASIFIKASI POST TRANSLATIONAL MODIFICATION N GLIKOSILASI PADA SEQUENCE PROTEIN MENGGUNAKAN ALGORITMA RANDOM FOREST" adalah benar hasil karya sendiri dan bukan orang lain. Seluruh tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Jika di kemudian hari terbukti skripsi saya adalah hasil penjiplakan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan peraturan yang berlaku.

Bandar Lampung, 9 November 2022

Penulis

Suci Hikmawati
1817051033

RIWAYAT HIDUP

Penulis dilahirkan di Pajar Bulan, Way Tenong, Lampung Barat pada tanggal 30 September 2000, sebagai anak kedua dari dua bersaudara dari pasangan Bapak Yusmadar dan Ibu Darmina. Penulis menempuh pendidikan formal pertama kali di Pendidikan Taman Kanak-kanak (TK) Al-Irsyad Darussalam yang diselesaikan pada tahun 2006, kemudian melanjutkan Sekolah Dasar (SD) di SDN 01 Pajar Bulan dan diselesaikan pada tahun 2012. Penulis melanjutkan Pendidikan Sekolah Pertama (SMP) di SMPN 01 Way Tenong yang diselesaikan pada tahun 2015 dan Pendidikan Sekolah Menengah Kejuruan (SMK) di SMKN 01 Way Tenong yang diselesaikan pada tahun 2018.

Penulis terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung pada tahun 2018 melalui jalur Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN). Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan antara lain.

1. Melaksanakan Karya Wisata Ilmiah (KWI) di Desa Way Bungur, Lampung Selatan, pada bulan Desember tahun 2018.
2. Menjadi anggota Adapter Himpunan Mahasiswa Jurusan Ilmu Komputer pada periode 2018/2019.
3. Menjadi bendahara di organisasi Rois Fmipa Biro Kemuslimahan pada periode 2018/2019.
4. Penulis juga aktif dalam kepanitian yang diadakan dilingkungan

kampus FMIPA UNILA seperti Penerimaan Mahasiswa Baru Ilmu Komputer (PRINTER) 2019 dan Pekan Raya Jurusan (PRJ) HIMAKOM 2019.

5. Pada tahun 2021 penulis melakukan Kerja Praktik di Radar Lampung TV.
6. Melaksanakan Kuliah Kerja Nyata (KKN) pada tahun ajaran 2021/2022 di Desa Pajar Bulan, Kecamatan Way Tenong, Kabupaten Lampung Barat.

MOTTO

“Orang hebat adalah orang yang memiliki kemampuan menyembunyikan kesusahan, sehingga orang lain selalu mengira bahwa ia selalu senang.”

(Imam Syafi`i)

“It’s not always easy, but that’s life. Be strong because
there are better days ahead.”

(Mark Lee)

“You may fail, life is kind of hard, people trip up all the time, but honestly, you're gonna wake up tomorrow and you're gonna be alright. You're gonna be happy.”

(Jae Day6)

“Berfikirlah positif setiap hari meskipun hal-hal buruk terjadi, suatu saat itu akan hilang. Kamu mungkin merasa frustasi, namun suatu saat akan berlalu dan kamu akan tumbuh dari pengalaman itu.”

(Wendy Red Velvet)

“Tetap berusaha, meskipun banyak *unfair* di *universe* ini, ingat ada MAMA yang mendukung. Kita punya POWER untuk sarjana,
DON’T FIGHT THE FEELING.”

(Exo L)

PERSEMBAHAN

Alhamdulillahirobbilalamin

Puji dan syukur saya ucapkan kepada Allah Subhanahu Wa Ta'ala atas segala rahmat dan hidayah-Nya sehingga dapat menyelesaikan penulisan skripsi ini. Sholawat dan salam saya sanjungkan kepada Nabi Muhammad SAW.

Aku persembahkan karya ini kepada:

Papa dan Mama

Sebagai tanda terimakasihku kepada Papa dan Mama yang tercinta dan yang tersayang. Terima kasih telah mendidik dan membesarkanku dengan kasih sayang kalian. Terima kasih selalu mendukungku dan mendoakanku dalam segala pilihanku. Terima kasih atas semua pengorbanan dan perjuangan kalian yang tiada hentinya. Terima kasih Papa dan Mama.

Kepada kakakku serta keluarga besar

Terima kasih telah memberikan semangat, dukungan, dan doa.

Sahabat dan teman-teman

terima kasih telah menemaniku, mendukungku, dan selalu memberikan kebahagiaan dalam hidupku.

Almamater Tercinta Universitas Lampung

SANWACANA

Puji syukur kehadirat Allah SWT, karena telah memberikan rahmat dan hidayahNya kepada saya sehingga saya dapat menyelesaikan skripsi dengan judul "Klasifikasi Post Translational Modification N Glikosilasi Pada Sequence Protein Menggunakan Algoritma Random Forest". Saya berharap skripsi ini dapat menambah pengetahuan bagi pembaca tentang klasifikasi situs glikosilasi, fitur ekstraksi dan algoritma *random forest*.

Proses penulisan skripsi ini tidak terlepas dari dukungan banyak pihak yang telah membimbing, membantu dan mendukung, sehingga pada kesempatan ini saya ingin menyampaikan ungkapan terima kasih kepada:

1. Orangtua, kakak, dan keluarga yang selalu mendoakan, memberi dukungan, kasih sayang, dan semangat baik secara moral maupun material dalam menyelesaikan skripsi ini.
2. Bapak Favorisen R. Lumbanraja, Ph.D. sebagai pembimbing yang telah membimbing saya, memberikan kritik dan saran dalam menyelesaikan skripsi ini sehingga dapat diselesaikan dengan baik.
3. Bapak M. Reza Faisal, S.T., M.T., Ph.D. sebagai pembahas pertama yang telah membimbing saya dalam memberikan ide, kritik, saran sehingga penulisan skripsi ini dapat diselesaikan dengan baik.
4. Bapak Dr. Ir. Kurnia Muludi, M.S.Sc. sebagai pembahas kedua yang telah membimbing saya dalam memberikan ide, kritik, saran sehingga penulisan skripsi ini dapat selesai dengan baik
5. Bapak Tristiyanto, S.Kom., M.I.S., Ph.D sebagai pembimbing akademik saya yang telah membantu dan selalu mendukung peningkatan akademik saya.
6. Bapak Dr. Suripto Dwi Yuwono, M.T. selaku Dekan FMIPA Universitas Lampung.

7. Bapak Didik Kurniawan, S.Si., M.T., selaku Ketua Jurusan Ilmu Komputer Universitas Lampung.
8. Bapak Dr. rer. nat. Akmal Junaidi, M. Sc. selaku sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
9. Bapak dan Ibu Dosen Jurusan Ilmu Komputer FMIPA Universitas Lampung yang telah memberikan ilmu dan pengalaman dalam hidup untuk menjadi lebih baik.
10. Ibu Ade Nora Maela, Bang Zainuddin, dan Mas Nofal yang telah membantu segala urusan administrasi dan segala jenis izin penulis di Jurusan Ilmu Komputer.
11. Teman-teman saya, Aulia Ahmad Nabil, Dewi Lestari, Fikri Mulya Permana, Rahmadila Nurjannah, Ratih Indah Wardani, dan Intania Rahmadhilla yang telah menemani, membantu, memberi semangat dan motivasi, serta mau mendengarkan keluh kesah saya selama masa perkuliahan.
12. Teman seperbimbingan saya, Intan, Ica, Ajeng, Ridho, Fajru, dan Sepryan yang sudah bersama-sama memberikan semangat dan bertukar pikiran serta saling membantu satu sama lain.
13. Sahabat saya, repa, dewi, ramadoni, feri, dan Rosilawati yang telah memberikan semangat dan motivasi.
14. Keluarga Ilmu Komputer 2018 yang tidak bisa penulis sebut satu persatu, yang telah memberikan pengalaman tak ternilai semasa duduk di bangku kuliah.
15. Seluruh kakak tingkat Ilmu Komputer yang tidak bisa disebutkan satu persatu yang telah membantu selama masa perkuliahan.
16. Boy grup BTS, EXO, dan NCT terutama untuk bias saya jungkook, chanyeol, dan markhyuck yang telah menemani, memberikan pengaruh positif, motivasi, dan inspirasi kepada penulis secara tidak langsung melalui karya-karyanya.
17. Semua pihak yang telah berpartisipasi baik secara langsung maupun tidak langsung dalam membantu penyusunan skripsi ini.

Penulis menyadari bahwa dalam penulisan skripsi ini masih terdapat banyak kekurangan. Oleh karena itu, saran dan kritik yang membangun sangat diharapkan

sebagai bahan evaluasi untuk kedepannya. Semoga skripsi ini dapat bermanfaat bagi semua pihak.

Bandar Lampung, 9 November 2022

Penulis

Suci Hikmawati
1817051033

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1. Latar Belakang

Protein atau dalam Bahasa yunani yaitu *proteos* artinya adalah yang utama atau yang didahulukan. Istilah tersebut diperkenalkan oleh Gerardus Mulder (1802-1880) yang merupakan ahli kimia belanda. Protein adalah asam amino yang dirangkai menjadi kesatuan dengan ikatan peptida. Dalam tubuh manusia, tiga per empat zat padatnya terdiri dari protein, yaitu otot, enzim, protein plasma, *antibody*, dan *hormone* (Suprayitno & Sulistiyati, 2017). Molekul besar yang paling banyak ditemukan dalam sel mahluk hidup adalah Protein. Jumlah protein sangat beragam baik dari struktur maupun fungsinya (Campbell et al., 2008).

Pembentukan protein yang matang dapat dilakukan melalui PTM atau *Post-Translational-Modification* dengan disintesisnya protein oleh ribosom dimana mRNA diterjemahkan menjadi rantai polipeptida (Yuliana & Fathurohman, 2020). Struktur protein yang mengalami perubahan disebut dengan modifikasi protein, dengan terjadinya pembentukan senyawa di karbonil, ikatan silang, fluoresensi dan lain sebagainya (E. Suhartono et al., 2004). Penyebab terjadinya modifikasi protein salah satunya adalah karena adanya pembebanan glukosa melalui reaksi glikosilasi (E. Suhartono et al., 2005).

Reaksi glikosilasi merupakan reaksi yang terjadi antara gugus amina protein dengan gugus aldehid dari glukosa yang dapat menciptakan produk-produk reaktif, yang kemudian dapat memodifikasi protein (E. Suhartono et al., 2005). Modifikasi pasca translasi yang paling rumit dalam sel eukariotik salah satunya adalah glikosilasi, pada proteome manusia hampir 50%

terglikosilasi. Glikosilasi berperan penting dalam fungsi biologis seperti pengenalan antigen, komunikasi sel-sel, ekspresi gen dan pelipatan protein (Akmal et al., 2017). Dampak dari glikosilasi pada protein sangat penting untuk perkembangan, pertumbuhan, fungsi atau kelangsungan hidup suatu organisme. Glikosilasi dikelompokkan menjadi empat kategori berdasarkan sifat kimia antara asam amino akseptor spesifik dan glikan, yaitu *N-linked, O-linked, C-Mannosylation dan glikosilfosfatidilinositol* atau GPI (Blom et al., 2004). Terdapat dua jenis glikosilasi yang umum terjadi yaitu glikosilasi pada atom nitrogen atau N glikosilasi yang terjadi pada residu asam amino asparagin dan glikosilasi pada atom oksigen atau O glikosilasi yang terjadi pada residu asam amino serin atau *threonine*. Sedangkan untuk *C-Mannosylation* dan *glikosilfosfatidilinositol* atau GPI jarang terjadi (Chauhan et al., 2013). Diketahui N glikosilasi mempengarui pelipatan protein (Helenius & Aebi, 2004). N glikosilasi memegang 90% bagian dalam total glikosilasi (Van Den Steen et al., 1998). N-glikosilasi adalah modifikasi pasca translasi yang berperan penting dalam pelipatan yang tepat dan fungsi protein (Ruiz-Blanco et al., 2016).

Sulit untuk melakukan identifikasi pada modifikasi setelah protein diisolasi dari sel eukariotik secara eksperimental tanpa menggangu struktur asli protein. Untuk melakukan analisis tersebut dapat dilakukan melalui spektrometri massa, dimana percobaan melibatkan seluruh eksperimen organisme hidup dalam kondisi laboratorium yang terkontrol dan kebutuhan sampel yang digunakan lebih banyak sehingga proses tersebut tidak efisien karena memerlukan biaya dan waktu yang mahal (Akmal et al., 2017). Kemudian berbagai peneliti telah mengusulkan menggunakan metode komputasi yaitu dengan pendekatan *machine learning* sebagai cara untuk menentukan situs glikosilasi pada protein menggunakan struktur utamanya dan juga melalui bidang bioinformatika. Dengan berkembangnya internet saat ini, bidang bioinformatika dapat dilakukan dengan mudah, dimana basis data bioinformatika terhubung ke internet sehingga memudahkan peneliti untuk mengumpulkan atau memperoleh sekuens biologis sebagai bahan analisis (Seprianto, 2017).

Pada penelitian sebelumnya yang dilakukan oleh Akmal, et al. (2017), prediksi N-glikosilasi telah dilakukan menggunakan algoritma *backpropagation* dan validasi silang yang dilakukan sepuluh kali lipat dan kinerja lainnya seperti akurasi, sensitivitas, spesifisitas dan koefisien korelasi *Mathew* dengan menunjukkan hasil bahwa keakuratan model yang diusulkan mengungguli model yang ada seperti *Glyomine, GlycoEP, Ensemble* SVM dan GPP. Selain itu juga memberikan pendekatan biaya dan waktu yang akurat dibandingkan dengan metode *in silico* dan *in vitro* yang ada. Berdasarkan penelitian sebelumnya, maka dalam penelitan ini akan melakukan klasifikasi situs N-glikosilasi pada *sequence* protein dengan menggunakan algoritma *Random Forest.*

1.2. Rumusan Masalah

Berdasarkan pemaparan latar belakang diatas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Berapa tingkat akurasi algoritma *random forest* dalam mengklasifikasikan dataset situs N-glikosilasi.
2. Apakah hasil akurasi algoritma *random forest* lebih tinggi dibandingkan dengan penelitian oleh Akmal, et al. (2017) yang menggunakan algoritma *Neural Network*

1.3. Batasan Masalah

Batasan Masalah dalam penelitian ini adalah sebagai berikut.

1. Penelitian ini menggunakan algoritma *Random Forest* untuk melakukan klasifikasi situs N-glikosilasi.
2. Data yang digunakan dalam penelitian ini diperoleh dari jurnal penelitian sebelumnya yang dilakukan oleh Akmal, et al. (2017).

1.4. Tujuan

Tujuan penelitian ini adalah sebagai berikut.

1. Tujuan penelitian ini adalah untuk mengetahui keakuratan algoritma *Random Forest* dalam melakukan klasifikasi situs N-glikosilasi.
2. Membandingkan hasil kinerja dari klasifikasi *random forest* dengan penelitian sebelumnya, yaitu penelitian yang dilakukan oleh Akmal, et al. (2017) yang menggunakan algoritma *Neural Network.*

1.5. Manfaat

Manfaat penelitian ini adalah sebagai berikut.

1. Sebagai wawasan dan pengetahuan mengenai klasifikasi situs N-glikosilasi menggunakan algoritma *Random Forest* yang dapat digunakan oleh penelitian selanjutnya.
2. Hasil dari tingkat keakuratan algoritma *Random Forest* dalam mengklasifikasikan situs N-glikosilasi dapat dijadikan sebagai informasi bagi penelitian selanjutnya.

II. TINJAUAN PUSTAKA

2.1. Penelitan Terdahulu

Penelitian ini tidak terlepas dari beberapa penelitian terdahulu yang telah dilakukan. Sehingga penelitian ini mempunyai kesamaan dan perbedaan objek yang diteliti. Tabel 1 merupakan beberapa ringkasan dari penelitian terdahulu.

Tabel 1. Ringkasan Penelitian Terdahulu Sebagai Acuan Penelitian ini

No	Nama Peneliti	Judul	Data	Metode	Hasil
1	(Akmal et al., 2017)	Prediction of N-linked glycosylation sites using position relative features and statistical moments	Dataset situs N glikosilasi Positif 11.601 Negatif 12.160	*Neural Network Back Propagation*	Rata-rata akurasi = 99,9%
2	(Taherzadeh et al., 2019)	SPRINT-Gly: predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties	Dataset protein manusia *N*- 2369 *O*- 211 Dataset protein tikus *N*- 2096 *O*- 398	*Deep neural network dan* SVM	Nilai AUC dari protein manusia *N*- 0,98 *O*- 0,82 Nilai AUC dari protein tikus *N*- 0,99 *O*- 0,79

Tabel 1. Ringkasan Penelitian Terdahulu Sebagai Acuan Penelitian ini (lanjutan)

No	Nama Peneliti	Judul	Data	Metode	Hasil
3	(Chien et al., 2020)	N-GlycoGo: Predicting Protein N-Glycosylation Sites on Imbalanced Data Sets by Using Heterogeneous and Comprehensive Strategy	*Human training set* *Positif* 3.836 *Negative* 18.277 *Mouse training set* *Positif* 57 *Negative* 948 *Human independent set* *Positif* *57* *Negative* *948* *Mouse independent set* *Positif* *13* *Negative* *145*	Metode *XGBoost*	*Independent test of humans* Sn 0,828 Sp 0,848 Acc 0,847 MCC 0,397 *Independent test of mouse* Sn 0,923 Sp 0,948 Acc 0,946 MCC 0,719

Penelitian terdahulu digunakan sebagai acuan dan perbandingan dalam penelitian ini. Beberapa penelitian yang digunakan adalah sebagai berikut:

2.1.1. Prediction of N-linked glycosylation sites using position relative features and statistical moments (2017)

Penelitian ini dilakukan oleh Akmal, et al. (2017) yang melakukan prediksi terkait N-glikosilasi. Dalam penelitian ini menggunakan algoritma *back propagation neural network.* Terdiri dari empat fase yaitu pengumpulan data, penyaringan data, ekstraksi fitur dan pelatihan. Data diperoleh dari *database uniprot* dalam bentuk *text xml* kemudian data diekstraksi menggunakan skrip *parsing* yang kemudian total data menjadi 23761, yang mana 11601 merupakan situs positif N-glikosilasi dan 12160 situs negatif. Fitur ekstraksi yang digunakan penelitian ini adalah fitur varian posisi dan komposisi, *raw*, *Hahn* dan *central moment*. Akurasi model menggunakan metrik akurasi *benchmark* seperti *matthew correlation*

coefficient, sensitifitas, spesifisitas dan akurasi. Model yang diusulkan berusaha untuk memprediksi situs glikosilasi n-*linked* dalam molekul protein yang berada pada sel prokariotik dan eukariotik. Model yang dilatih kemudian divalidasi menggunakan *cross-validation, jacknife testing,* dan *self-consistency testing*.

Dalam penelitian ini sepuluh kali lipat validasi silang telah dilakukan baik data negatif maupun positif yang mana masing masing data dibagi menjadi data uji dan data latih. Rata-rata nilai akurasi model prediksi adalah 99,9%. Hasil yang dihasilkan oleh model tersebut menunjukkan bahwa keakuratan alat yang diusulkan adalah jauh lebih baik dari sistem yang ada seperti *Glyomine, GlycoEP, Ensemble* SVM dan GPP.

2.1.2. SPRINT-Gly: predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties (2019)

Penelitian yang dilakukan oleh Taherzadeh, et al. (2019), yaitu melatih *deep learning neural network dan support vector machine* untuk memprediksi N dan O situs glikosilasi pada protein manusia dan tikus. Data yang digunakan berasal dari *uniprot, N-linked glycoproteins* dari dbPTM, *GlycoProtDB,* informasi situs N-glikosilasi protein manusia dari *Unipep* dan *UniCarbKB. Dataset* berisi 7023 glikoprotein yang merupakan kumpulan data khusus situs untuk model *machine learning.* Dalam mempersiapkan *benchmark* ini, *dataset* dipecah menjadi *N-linked* dan *O-linked* dan model prediktif yang dihasilkan untuk setiap spesies dan jenis glikosilasi. Jumlah *dataset* protein manusia pada *linked-N* adalah 2369 dan *linked-O* 211, sedangkan *dataset* protein tikus adalah 2096 untuk data *linked-N* dan 398 untuk *linked-O*. Disetiap set data dipilih 90% secara acak untuk digunakan sebagai data latih dengan melakukan 10 kali validasi silang dan 10% sisanya digunakan untuk

set uji independent. Situs terkait-N, protein manusia dan tikus diprediksi dengan sensitivitas dan presisi tinggi.

Untuk set uji independen, metode sensitivitas mencapai 98% dan presisi 93% untuk protein manusia, peningkatan yang signifikan jika dibandingkan dengan *NetNGlyc* (sensitivitas 75% dan presisi 93%), atau *GlycoPP* (sensitivitas 91% dan presisi 68%). Metode untuk protein tikus lebih akurat dengan sensitivitas 99% dan presisi 99%, dibandingkan dengan *NetNGlyc* (sensitivitas 72% dan presisi 96%) dan *GlycoPP* (sensitivitas 100% dan presisi 73%). Ini menunjukkan bahwa prediksi tentang situs terkait-N mendekati akurasi eksperimental. Untuk situs terkait-O akurasi *Sprint-Gly* tidak tinggi yang disebabkan oleh ukuran kumpulan data dan pola urutan yang lebih lemah. Namun demikian, ada peningkatan yang signifikan dibandingkan dengan metode sebelumnya, meskipun sensitivitas set tes independent manusia 33% dan presisi uji independent 14% tetap rendah.

2.1.3. N-GlycoGo: Predicting Protein N-Glycosylation Sites on Imbalanced Data Sets by Using Heterogeneous and Comprehensive Strategy (2020)

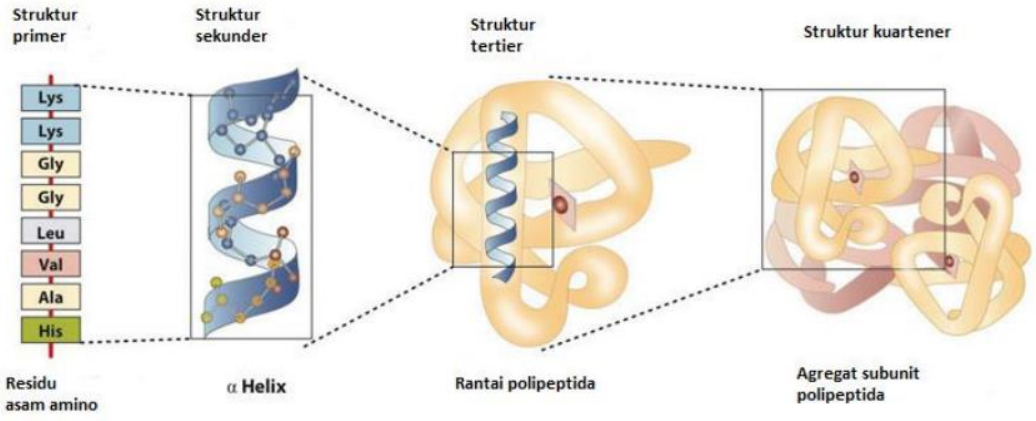
Penelitian ini dilakukan oleh Chien, et al. (2020). Alat prediksi glikosilasi terkait-N saat ini tidak memperhitungkan ketidakseimbangan yang serius antara data positif dan negatif. Dalam penelitian ini, menggunakan urutan protein dan karakteristik asam amino untuk membangun model prediksi glikosilasi terkait-N yang disebut *N-GlycoGo*. Peneliti membangun *N-GlycoGo* untuk meningkatkan metode prediksi situs glikosilasi melalui integrasi model, menggunakan semua data ketidakseimbangan positif dan negatif, memecahkan masalah data yang tidak seimbang, dan mengembangkan lebih banyak model yang akurat untuk memprediksi glikosilasi. Sumber data glikosilasi yang digunakan

oleh *N-GlycoGo* adalah Sumber Data Protein Universal (UniProt), dbPTM, dan O-GlycBase v6.00. Algoritma yang dibangun menggunakan algoritma *XGBoost*.

Hasil dari penelitian ini adalah pengujian independen manusia dan model prediksi tikus menunjukkan bahwa *N-GlycoGo* lebih unggul dari alat lain dengan korelasi *Matthews* nilai koefisien (MCC) masing-masing 0,397 dan 0,719, yang lebih tinggi dari alat prediksi situs glikosilasi lainnya. Selain itu nilai akurasi MCC dari metode *XGBoost* mencapai 0,981 yang mana lebih tinggi dari metode SVM dan *random forest* dengan masing-masing nilai akurasi MCC yaitu 0,96 dan 0,961.

2.2. Protein

Protein adalah molekul besar yang ditemukan sangat banyak di dalam sel makhluk hidup. Jumlah protein sangat bervariasi baik dari struktur maupun fungsinya (Campbell et al., 2008). Proses-proses dalam metabolisme diatur oleh protein berbentuk enzim dan hormon, salah satu sumber utama energi adalah protein. Protein dalam bentuk *khromosom* memiliki peran untuk menyimpan dan melanjutkan sifat-sifat keturunan gen (Thomy & Harnelly, 2018). Protein merupakan polimer asam amino yang memiliki tanggung jawab dalam melaksanakan perintah yang berada pada kode genetic 20 asam amino berbeda yang digunakan untuk melakukan sintesis protein. Protein dapat ditemukan dalam sel-sel, yaitu di sitoplasma dan membran sel yang berada pada sel prokariot dan dapat ditemukan pada sel eukariot, yaitu di sitoplasma, membran sel dan organel beserta membrannya diantaranya adalah mitokondria, inti sel, lisosom, badan golgi, kloroplas, plastida, dan vakuola (Suhartono, 2017).



Gambar 1. Tingkatan keempat struktur protein (Simamora, 2015).

Protein mempunyai berbagai struktur khusus yang tersusun dari rangkaian asam amino. Menurut Yuliana & Fathurohman (2020). Tingkat struktur protein ada 4 yang dapat dilihat pada Gambar 1, diantaranya adalah:

a. Struktur Primer

   Struktur primer adalah struktur yang tidak rumit dengan urutan asam amino yang terangkai secara linear yaitu seperti susunan huruf dalam sebuah kata dan tidak adanya percabangan rantai. Struktur primer terbentuk Melalui ikatan peptida atau ikatan amida dimana struktur ini dapat menetapkan urutan asam amino dari suatu polipeptida.

b. Struktur Sekunder

   Struktur sekunder merupakan campuran dari struktur primer yang linear dinormalkan oleh ikatan hidrogen amina dan oksigen karbonil pada polipeptida. Terdapat 2 macam struktur sekunder, yaitu *alpha helix* dan *beta sheet.*

c. Struktur Tersier

   Struktur tersier adalah struktur protein yang membentuk struktur 3 dimensi tertentu dari keseluruhan lipatan rantai polipeptida.

d. Struktur Kuartener

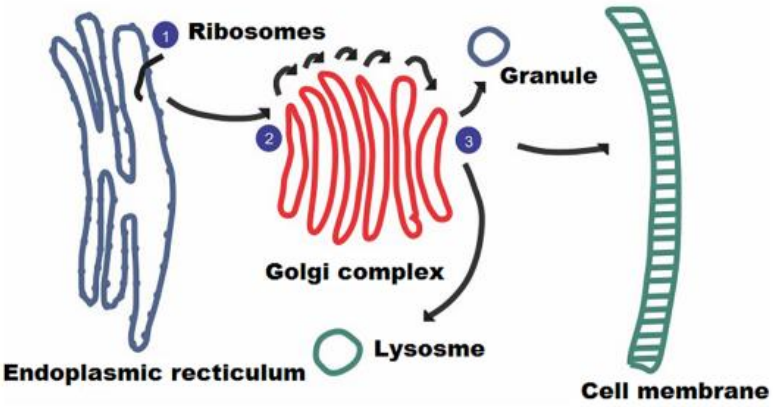
   Terdapat dua atau lebih sub unit protein pada struktur kuartener yang akan membentuk protein lengkap yang fungsional melalui struktur tersier. Protein multimerik adalah sebutan protein dengan struktur kuartener.

2.3. *Post Translational Modification* (PTM)

Modifikasi pascatranslasi atau PTM dilakukan untuk membentuk produk protein yang matang dengan disintesisnya protein oleh ribosom yang menerjemahkan mRNA menjadi rantai polipeptida. Modifikasi ini dapat terjadi di rantai samping asam amino atau pada terminal C atau N protein (Yuliana & Fathurohman, 2020). Sebelum menjadi protein fungsional, beberapa polipeptida akan mengalami suatu proses lebih lanjut setelah proses translasi terjadi. Yang pertama, polipeptida akan ditujukan ke berbagai unsur selular. Kedua, melalui reaksi kimia beberapa polipeptida akan mengalami substitusi sebelum protein aktif terbentuk. Ketiga, akan terjadinya mekanisme degradasi terprogram pada protein. Polipeptida membutuhkan struktur yang tepat supaya dapat berfungsi. Modifikasi pasca translasi juga terjadi pada beberapa protein prokariot dan eukariot, seperti glikosilasi, fosforilasi, asetilasi, metilasi dan pemotongan peptida sinyal. Untuk menjadikan protein aktif dalam sel, maka protein harus melalui minimal satu proses dari empat tipe pemrosesan, diantaranya adalah *protein folding* atau pelipatan protein, *proteolytic cleavage* atau pemotongan proteolitik, modifikasi kimia dan pembuangan intein (Agus, 2018).

2.4. Glikosilasi

Glikosilasi adalah salah satu modifikasi protein pasca translasi yang paling kompleks. Adapun dampak dari adanya glikosilasi adalah sangat penting untuk pertumbuhan, perkembangan, fungsi, atau kelangsungan hidup suatu organisme. Terdapat 5 jenis glikosilasi pada eukariota, yaitu *N-linked, O-linked, C-linked, P-linked,* dan *G-linked.* Jenis glikosilasi yang paling umum adalah N-glikosilasi dan *O-linked,* untuk *G-linked* dan *P-linked* jarang terjadi. Penambahan bagian glikan ke atom nitrogen merupakan tanda dari glikosilasi *N-linked.* Aparatus golgi merupakan tempat terjadinya glikosilasi *O-linked* setelah glikosilasi *N-linked.* Untuk glikosilasi *C-linked* merupakan glikosilasi yang jarang terjadi, dalam hal ini ditemukannya glikan melekat di karbon triptofan pertama (Chauhan et al., 2013).



Gambar 2. Proses glikosilasi (Akmal et al., 2017).

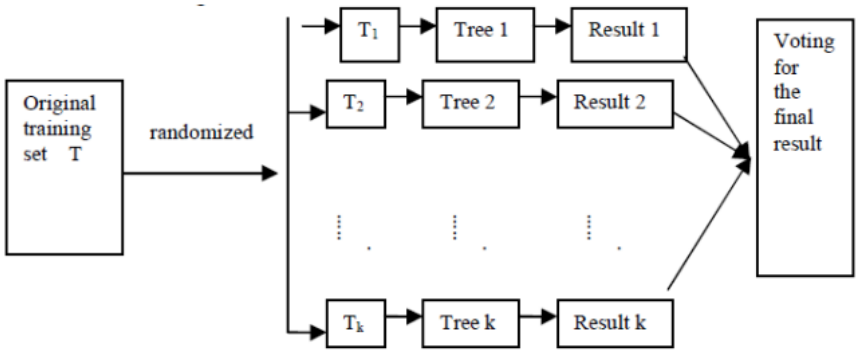
Proses glikosilasi bermula pada reticulum endoplasma, yang kemudian protein disintesis oleh ribosom dan masuk ke dalam apparatus golgi. Proses glikosilasi dapat dilihat pada Gambar 2. Glikosilasi berperan penting dalam modifikasi protein, dimana karbohidrat melekat pada molekul protein. Asam amino spesifik yang ditambahkan gula dari protein akan memperoleh heterogenitas protein yang dapat melakukan berbagai fungsi seluler (Akmal et al., 2017).

2.5. *Random Forest*

Metode *random forest* merupakan pengembangan dari *decision tree*, yang mana telah dilakukannya proses pelatihan pada setiap *decision tree* dengan menggunakan sampel individu (Putra et al., 2016). *Decision tree* merupakan metode klasifikasi yang menggunakan representasi struktur pohon, di mana setiap node merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut dan daun merepresentasikan kelas (Saifullah et al., 2017). Teknik *decision tree* termasuk ke dalam kelompok teknik klasifikasi tunggal, dimana kelemahan dari teknik klasifikasi tunggal adalah sangat rentan terhadap data yang tidak seimbang. Hal tersebut dikarenakan teknik klasifikasi tunggal lebih fokus pada kelas mayoritas dan cenderung mengabaikan informasi yang ada pada kelas minoritas. Selain itu pohon yang dihasilkan kurang stabil dikarenakan hasil prediksi hanya berpacu pada satu pohon saja, sehingga jika terjadi sedikit perubahan pada data training

maka akan mengalami perubahan yang signifikan pada pohon yang dihasilkan. Melihat permasalahan pada *decision tree,* beberapa peneliti menemukan teknik baru, salah satunya adalah teknik *ensemble* yang dipercaya dapat menangani berbagai macam tipe data (Marie et al., 2019). Teknik *ensemble* adalah cara untuk meningkatkan akurasi klasifikasi melalui kombinasi dari metode klasifikasi, Salah satu metode teknik *ensemble* adalah *Random forest* (Han et al., 2012). Kemampuan metode *random forest* dengan *decision tree* tidak jauh berbeda (Ali et al., 2012). Namun *random forest* lebih unggul dalam memprediksi kelas dengan jumlah yang sedikit atau minoritas dan jumlah data yang banyak dibandingkan dengan *decision tree* (Marie et al., 2019).

Random forest merupakan algoritma yang digunakan pada klasifikasi data dengan jumlah yang besar. Pada *random forest* banyak pohon ditumbuhkan sehingga terbentuk hutan *(forest)* yang kemudian dilakukan analisis pada kumpulan pohon tersebut (Binarwati et al., 2017). Metode ini memiliki konsep dasar, yaitu beberapa *tree* dibangun untuk melakukan klasifikasi suatu objek menurut atributnya (Tahyudin et al., 2021). Prediksi klasifikasi didapatkan dari proses *vote* pohon-pohon klasifikasi yang terbentuk (Jin et al., 2020).



Gambar 3. *Random forest schematic* (Liu et al., 2012).

Gambar 3. Merupakan *schema* dari algoritma *random forest. Training* data akan dilakukan terhadap sampel data, masing masing pohon keputusan memprediksi dengan prediktor acak, kemudian hasil dari setiap pohon

dikombinasikan dan melakukan *vote* untuk menghasilkan klasifikasi berdasarkan *majority vote* (Liu et al., 2012). Berikut merupakan algoritma *random forest* :

1. Tahapan *bootstrap,* yaitu dengan mengambil n data sampel acak dari dataset awal dengan pengembalian (*replacement).*
2. Setiap dataset hasil dari tahapan bootstrap dilakukan penyusunan pohon klasifikasi dengan penentuan pemilah terbaik didasarkan pada variabel *predictor* yang dipilih secara acak. Jumlah variabel tersebut dapat ditentukan melalui perhitungan (L Breiman, 2003) yang dapat dilihat pada Persamaan 1.

$$mtry = \begin{cases} mtry_1 = \sqrt{p} \\ mtry_2 = \frac{1}{2}\sqrt{p} \\ mtry_3 = 2\sqrt{p} \end{cases} \quad (1)$$

   dimana *p* adalah jumlah variable *predictor*.
3. Setelah pohon klasifikasi terbentuk, maka melakukan prediksi klasifikasi terhadap data sampel.
4. Ulangi langkah 1 sampai 3 sebanyak k kali sehingga terbentuk sebuah hutan yang terdiri atas k pohon acak.
5. Melakukan prediksi klasifikasi data sampel akhir dengan mengombinasikan hasil prediksi pohon klasifikasi yang diperoleh berdasarkan aturan *majority vote.*

Terdapat dua parameter utama *random forest*, yaitu *mtry* yang merupakan jumlah variabel input yang dipilih secara acak pada setiap *split* dengan nilai *default mtry* = $\sqrt{p}$ dimana *p* adalah jumlah variabel prediktor dan parameter *ntree* merupakan jumlah pohon yang akan dibangun dengan nilai *default ntree* = 500 (Genuer et al., 2008). Menurut Breiman, (2001) penggunaan nilai *mtry* yang tepat akan menghasilkan hubungan atau korelasi antar pohon kecil namun memiliki kekuatan setiap pohon yang cukup besar.

Pembentukan pohon pada *random forest* menggunakan algoritma CART (*Classification And Regression Tree*), namun tidak ada proses pemangkasan

(*pruning*) prosesnya terdiri dari tiga hal, yaitu pemilihan pemilah atau *split*, penentuan simpul terminal dan penandaan label kelas (L. Breiman et al., 1984). Untuk melakukan proses pemilihan pemilah atau *splitting* pada *random forest* menggunakan perhitungan *gini index* yang dapat dilihat pada Persamaan 2 ini.

$$Gini = 1 - \sum_{i=1}^{n}(pi)^2 \quad \text{.....(2)}$$

Dimana *n* adalah jumlah kelas variable Y, *i* adalah kelas atribut dan *pi* adalah proporsi jumlah kelas dalam atribut *i* terhadap jumlah kelas *n* dalam atribut. Berikut merupakan contoh perhitungan *gini index* menggunakan potongan *dataset* sederhana yang dapat dilihat pada Tabel 2.

Tabel 2. Dataset sederhana untuk contoh perhitungan gini index

Alamat	**Pekerjaan**	**Status Kredit**
Kota Agung	Karyawan Swasta	Macet
Metro	Guru	Lancar
Natar	PNS	Lancar
Metro	Pensiunan	Lancar
Metro	PNS	Lancar
Bandar Lampung	Karyawan BUMN	Lancar
Metro	Karyawan Swasta	Macet
Kedaton	Karyawan Swasta	Macet
Metro	Karyawan BUMN	Macet
Metro	PNS	Lancar

Jika dilihat pada Tabel 2 di atas terdapat dua *feature* yang terdiri dari alamat, pekerjaan dan label status kredit yang memiliki 10 *row* data, kemudian akan dilakukan perhitungan *gini* dari masing masing *feature*, dengan menghitung proporsi data dari masing-masing *sample* dalam *feature* terlebih dahulu. Sebagai contoh dilakukan perhitungan proporsi data pada *feature* alamat yang dinotasikan dalam Tabel 3.

Tabel 3. Proporsi data dari atribut alamat

Alamat	Proporsi Data
Kota Agung	$\frac{1}{10}$
Metro	$\frac{6}{10}$
Natar	$\frac{1}{10}$
Bandar Lampung	$\frac{1}{10}$
Kedaton	$\frac{1}{10}$

Setelah mendapatkan proporsi data dari masing-masing *sample* atau *row* pada *feature* alamat, selanjutnya mencari probabilitas *feature* alamat dimana memiliki kemungkinan untuk mendapatkan label dari status kredit yang terdiri dari dua status, yakni lancar dan macet. Hasil dari probabilitas tersebut dimuat dalam bentuk Tabel 4.

Tabel 4. Probabilitas Lancar dan Macet

Alamat	Probabilitas Lancar	Probabilitas Macet
Kota Agung	$\frac{0}{6}$	$\frac{1}{4}$
Metro	$\frac{4}{6}$	$\frac{2}{4}$
Natar	$\frac{1}{6}$	$\frac{0}{4}$
Bandar Lampung	$\frac{1}{6}$	$\frac{0}{4}$
Kedaton	$\frac{0}{6}$	$\frac{1}{4}$

Setelah memperoleh proporsi data dan probabilitas dari masing-masing *sample* pada *feature* alamat, maka selanjutnya adalah memasukkan hasil tersebut kedalam perhitungan *gini* yang dapat dilihat pada Tabel 5.

Tabel 5. Hasil Perhitungan Gini Index

Alamat	***Gini Index Formula***	***Hasil Gini Index***
Kota Agung	$1 - \left(\left(\frac{0}{6}\right)^2 + \left(\frac{1}{4}\right)^2\right)$	0,93
Metro	$1 - \left(\left(\frac{4}{6}\right)^2 + \left(\frac{2}{4}\right)^2\right)$	0,31
Natar	$1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{0}{4}\right)^2\right)$	0,97
Bandar Lampung	$1 - \left(\left(\frac{1}{6}\right)^2 + \left(\frac{0}{4}\right)^2\right)$	0,97
Kedaton	$1 - \left(\left(\frac{0}{6}\right)^2 + \left(\frac{1}{4}\right)^2\right)$	0,93

Dapat dilihat pada Tabel 5 yang merupakan hasil *gini index* dari masing-masing *sample* pada *feature* alamat, yang selanjutnya akan dilakukan perhitungan untuk menentukan *gini impurity*. Perhitungan tersebut dinotasikan pada Persamaan 3.

$$\left(\frac{1}{10}\right)0{,}93 + \left(\frac{6}{10}\right)0{,}31 + \left(\frac{1}{10}\right)0{,}97 + \left(\frac{1}{10}\right)0{,}97 + \left(\frac{1}{10}\right)0{,}93 = 0{,}56 \text{.....} (3)$$

Dari perhitungan diatas diperoleh hasil *gini index* dari *feature* alamat sebesar 0,56. Begitupun dengan *feature* pekerjaan akan dilakukan juga perhitungan *gini index* seperti yang dilakukan pada *feature* alamat. Berikut merupakan Hasil perhitungan *gini index* dari *feature* pekerjaan yang dapat dilihat pada Persamaan 4.

$$\left(\frac{3}{10}\right)0{,}43 + \left(\frac{1}{10}\right)0{,}97 + \left(\frac{3}{10}\right)0{,}75 + \left(\frac{1}{10}\right)0{,}97 + \left(\frac{2}{10}\right)0{,}91 = 0{,}73 \text{.....} (4)$$

Berdasarkan hasil *gini index* yang diperoleh dari *feature* pekerjaan, maka *feature* alamat akan menjadi *node* teratas untuk dijadikan variable pemisah atau *split* dari salah satu *tree* karena memiliki *gini index* terkecil. Setelah memperoleh *node* teratas maka selanjutnya menentukan variable pemisah berikutnya yang disebut sebagai *sub-node.* Tahap ini akan dilakukan sampai

stopping criteria tercapai yaitu minimum sampel per simpul terminal terpenuhi. Sehingga jika percabangan berhenti maka akan menghasilkan simpul terminal sebagai hasil prediksi satu pohon.

Pendugaan kesalahan dalam klasifikasi *random forest* dapat dilakukan dengan memperoleh nilai *error* OOB (*out of bag).* Data OOB merupakan data yang tidak tertera pada saat melakukan pembuatan *bootstrap* dan digunakan untuk memvalidasi pohon yang bersesuaian. Tingkat *Error* OOB berdasarkan data pelatihan dapat diperoleh dengan cara sebagai berikut:

1. Melakukan prediksi data yang tidak tertera di dalam *sample bootstrap* atau OOB pada pohon yang terbentuk.
2. Secara rata-rata, Setiap gugus data asli sekitar 36,8% akan menjadi *out of bag* untuk setiap pohon. Sehingga pada tahap 1 masing-masing amatan gugus data asli sekitar sepertiga kali dari banyaknya pohon akan mengalami prediksi.
3. Menghitung tingkat akurasi *error* OOB dari proporsi data misklasifikasi yang terdapat pada seluruh amatan gugus data asli.

Menurut Hastie et al., (2008) secara umum *error* OOB pada *random forest* tergantung pada dua hal yaitu kekuatan (*strength)* yang dilambangkan dengan μ pada masing-masing pohon tunggal dalam *random forets,* yang mana apabila nilai μ semakin besar maka nilai salah klasifikasi akan semakin kecil. Korelasi antar pohon dilambangkan dengan $\bar{p}$, nilai korelasi yang kecil mengakibatkan ragam dugaan hasil *random forest* menjadi kecil. Batasan besarnya kesalahan prediksi *random forest* telah dibuktikan oleh breiman (2001) pada Persamaan 5.

$$\varepsilon_{RF} \leq \bar{p}\left(\frac{1-\mu^2}{\mu^2}\right) \quad \text{.....(5)}$$

Dimana $\bar{p}$ adalah rata-rata korelasi atau *error* OOB *(out of bag)* antar pasangan dugaan dari dua pohon tunggal dan μ adalah rata-rata kekuatan atau *strength* akurasi pohon tunggal (Leo Breiman, 2001a). Apabila nilai μ semakin besar maka menunjukkan bahwa akurasi prediksinya semakin baik, sehingga jika ingin memiliki akurasi *random forest* yang memuaskan maka

harus diperoleh banyak pohon tunggal dengan $\bar{p}$ yang kecil dan μ yang besar (Sartono & Syafitri, 2010).

Kinerja *random forest* dilihat berdasarkan jumlah pohon dan *mtry* yang memiliki nilai *error* OOB yang terkecil untuk mendapatkan parameter yang optimal. Nilai *error* OOB diperoleh menggunakan data *training* dengan perhitungan pada Persamaan 6.

$$\frac{\text{FN+FP}}{\text{jumlah data training}} \times 100\% \quad (6)$$

2.6. *Variable importance*

Variable importance merupakan ukuran yang merepresentasikan kepentingan variabel prediktor dalam memprediksi. Pengukuran kepentingan variabel dapat menggunakan *Mean Decrease Impurity* (MDI) atau biasa disebut dengan *Mean Decrease Gini* (MDG) dan *Mean Decrease Accuracy* (MDA) yang diusulkan oleh Breiman, (2001). MDG merupakan ukuran tingkat kepentingan suatu variabel yang dihitung dengan rata-rata gini pada sebuah variabel ketika variabel tersebut terpilih sebagai *splitting node,* ukuran tersebut digunakan untuk mengetahui tingkat kestabilan dari tiap variabel independen dalam *random forest*. Semakin besar nilainya maka akan semakin baik (Leo Breiman, 2001). Misalkan terdapat *p* peubah penjelas dengan *h* = (1, 2,…, *p)* maka MDG akan mengukur tingkat kepentingan peubah penjelas x_h dengan Persamaan 7.

$$MDG_{(x_h)} = \frac{1}{k}[1 - \sum_k Gini\ (h)^k] \quad (7)$$

Dimana :

Gini $(h)^k$: merupakan indeks gini untuk peubah penjelas x_h pada pohon ke –k.

k : banyaknya pohon pada *random forest.*

Begitupun dengan MDA merupakan metode yang menghitung tingkat kepentingan dengan permutasi dan menggunakan OOB untuk membagi data sampelnya (Christy et al., 2021). Nilai MDA diukur dengan menggunakan Persamaan 8 berikut ini.

$$MDA_{(X_h)} = \frac{1}{k}\sum_{t=1}^{k} \frac{\sum_{i\epsilon OOB} I(y_i = f(x_i)) - \sum_{i\epsilon OOB} I\left(y_i = f\left(x_j^i\right)\right)}{|OOB|} \quad \text{.....(8)}$$

Dengan t : $\in \{1, 2, 3, \ldots, k\}$

$f(x_i)$: nilai rata-rata dari perbedaan antara kelas prediksi sebelum permutasi $x_{h.}$

$f\left(x_j^i\right)$: nilai rata-rata dari perbedaan antara kelas prediksi setelah permutasi variabel x_h.

2.7. *Feature Extraction*

Feature extraction adalah suatu proses pengambilan ciri atau karakteristik dari suatu data yang dapat menunjukkan informasi penting untuk kemudian dimanfaatkan dalam kebutuhan proses analisa data maupun klasifikasi (Baniya et al., 2015). *Feature extraction* akan melakukan beberapa perubahan fitur asli untuk memperoleh fitur lain yang lebih relevan. (Khalid et al., 2014). *Feature extraction sequence* protein disebut dengan protein *descriptor*. Panjang fitur yang dihasilkan disebut dengan *dynamic length* dan *fixed length*. *Fixed length* menghasilkan jumlah fitur yang sama atau tidak berubah sepanjang apapun *sequence* yang digunakan, sebaliknya jika jumlah fitur yang dihasilkan berubah-ubah atau tergantung dengan panjang *sequence* yang digunakan maka disebut dengan *dynamic length.* Pada penelitian ini fitur yang berkaitan dengan situs modifikasi pasca translasi di ekstraksi, fitur tersebut terdiri dari *statistical moment* dan fitur *position and composition.*

2.7.1. *Statistical Moment*

Momen statistik adalah ukuran kuantitatif yang pada dasarnya digunakan untuk mewakili akumulasi data. Momen statistik dihitung untuk deskripsi numerik sampel dalam kumpulan data dan berbagai urutan momen menggambarkan berbagai property data. Momen bergantung pada fungsi *polynomial* dan distribusi yang dihitung seperti *raw, hahn,* dan *central moment* (Awais et al., 2021). Dalam penelitian ini *central moment, raw* dan *hahn* dihitung untuk tujuan ekstraksi fitur dari sampel dalam kumpulan data.

Raw moment digunakan untuk menghitung *mean, varians* dan *asymmetry* dari *probability distribution* dalam kumpulan datanya. Begitupun dengan *central moment* digunakan untuk menghitung *mean, varians* dan *asymmetry*, keduanya merupakan varian skala tetapi *raw moment* merupakan lokasi invarian sedangkan central moment tidak. Kemudian *hahn* dihitung berdasarkan *polynomial hahn* dan merupakan varian skala dan lokasi (Khan et al., 2019). Moment tersebut dihitung dengan membentuk matriks persegi berdimensi dan semua residu asam amino dari urutan tersebut digabungkan kedalamnya. Matriks tersebut di representasikan pada Persamaan 10.

$$P' = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{n1} & \cdots & \beta_{nn} \end{pmatrix} \quad \text{.....(10)}$$

setiap elemen matriks ini merupakan residu dari *sequence*. Untuk menghitung moments dua dimensi struktur primer satu dimensi direpresentasikan ke dalam struktur dua dimensi dengan menggunakan skema baris mayor. Dimensi matriks dua dimensi dapat dihitung pada Persamaan 11 dengan mengambil akar kuadrat dari panjang protein,

$$n = \lceil \sqrt{k} \rceil \quad \text{.....(11)}$$

Dimana n merupakan dimensi matrik persegi dua dimensi dan k adalah panjang rantai polipeptida. Untuk perhitungan *raw moments*, ekspresi yang digunakan dapat dilihat pada Persamaan 12.

$$M_{ij} = \sum_{p=1}^{n} \sum_{q=1}^{n} p^i q^j \, \beta_{pq} \quad \text{.....(12)}$$

Dimana orde momen dideskripsikan oleh i dan j. Pada *central moment* menggunakan data *centroid* sebagai titik referensi dan dihitung dari Persamaan 13.

$$\mu_{ij} = \sum_{p=1}^{n} \sum_{q=1}^{n} (p - \bar{x})^i (q - \bar{y})^j \, \beta_{pq} \quad \text{.....(13)}$$

Dimana $\bar{x}$ dan $\bar{y}$ membentuk *centroid* yang dihitung dari Persamaan 14.

$$\bar{x} = \frac{M_{10}}{M_{00}} \; dan \; \bar{y} = \frac{M_{01}}{M_{00}} \quad \text{.....(14)}$$

Kemudian pada momen *hahn* dapat dihitung setelah notasi satu dimensi diubah menjadi notasi matriks persegi. Momen Hahn diskrit dua dimensi merupakan momen ortogonal yang membutuhkan matriks bujur sangkar sebagai data masukan dua dimensi. *Polynomial hahn* dengan ordo n direpresentasikan pada Persamaan 15.

$$h_n^{u.v}(r, N)$$
$$= (N + V - 1)_n (N - 1)_n \times \sum_{k=0}^{n} (-1)^k \frac{(-n)_k (-r)_k (2N+u+v-n-1)_k}{(N+v-1)_k (N-1)_K} \frac{1}{k!} \quad \text{...(15)}$$

Persamaan diatas menggunakan *symbol pochhammer* yang digeneralisasi pada Persamaan 16 sebagai berikut.

$$(a)_k = a(a+1) \dots (a+k-1) \quad \text{.....(16)}$$

Dan juga disederhanakan melalui *operator gamma* yang dapat dilihat pada Persamaan 17.

$$(a)_k = \frac{\Gamma(a+k)}{\Gamma(a)} \quad \text{.....(17)}$$

Moment hahn yang ternormalisasi *orthogonal* untuk data diskrit dua dimensi dihitung menggunakan Persamaan 18.

$$H_{ij} = \sum_{q=0}^{N-1} \sum_{p=0}^{N-1} \beta_{ij} h_i^{\widetilde{\mu,\nu}}(q, N) \, h_j^{\widetilde{\mu,\nu}}(p, N), m, n = 0,1, \dots, N-1 \quad \text{.....(18)}$$

Raw, *Central moment* dan *hahn moment* dihitung hingga orde 3.

2.7.2. *Position and Composition Variant Features*

Faktor yang mempengaruhi perilaku protein bukan hanya komposisi asam amino saja, tetapi juga posisi relatif residu asam amino sebagai faktor penting. Diketahui bahwa perubahan kecil dalam posisi relatif asam amino dapat mengubah karakteristik protein. Penempatan relatif residu asam amino adalah salah satu paradigma inti yang mengatur atribut fisik protein.

Site Vicinity Vector (SVV) diturunkan sebagai sub-struktur dari urutan primer yang berisi situs potensial bersama dengan tetangganya yang diberikan pada Persamaan 9.

$$\alpha_{q-r} \dots \alpha_{q-2}, \alpha_{q-1}, \alpha_{q+1}, \alpha_{q+2}, \dots . \alpha_{q+r} \quad \text{.....(9)}$$

Dimana r adalah nilai *integer* kecil yang dipilih secara optimal melalui *probing* dan eksperimen. SVV membentuk komponen vektor fitur inklusif, Dalam sintesis protein hanya 20 asam amino yang signifikan, untuk mengekstrak vektor fitur setiap asam amino diberi nilai integer yang unik. Selama nilainya unik, integral dan ditetapkan secara konsisten, tidak masalah nilai mana yang diberikan pada asam amino.

Position Relative Incident Matrix (PRIM) mengutip informasi posisi relatif komponen asam amino dalam rantai polipeptida. PRIM dibentuk sebagai matriks dimensi elemen 20x20 seperti yang terlihat pada Persamaan 19.

$$S_{prim} = \begin{bmatrix} A_{1\to1} & A_{1\to2} \cdots & A_{1\to n} \cdots A_{1\to20} \\ A_{2\to1} & A_{2\to2} \cdots & A_{2\to n} \cdots A_{2\to20} \\ A_{i\to1} & A_{i\to2} \cdots & A_{i\to j} \cdots A_{i\to20} \\ A_{N\to1} & A_{N\to2} \cdots & A_{N\to j} \cdots A_{N\to20} \end{bmatrix} \quad \text{.....(19)}$$

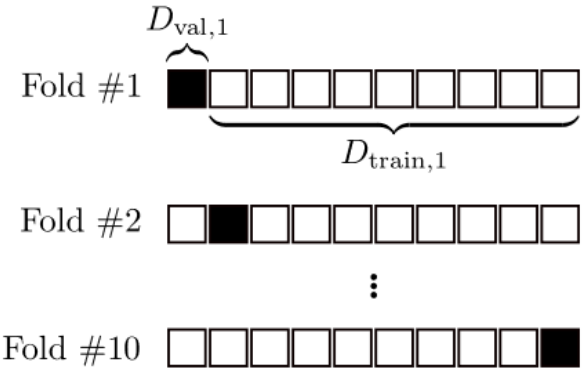
Matriks PRIM mengekstrak informasi mengenai posisi relatif asam amino dalam rantai polipeptida. Matriks lain, yaitu Reverse Position Relative Incident Matrix (RPRIM) dibentuk yang bekerja dengan cara yang sama seperti PRIM tetapi pada urutan primer terbalik. Pengenalan RPRIM membantu mengungkap pola tersembunyi lebih lanjut dan mengurangi ambiguitas di antara protein dengan urutan polipeptida yang tampak sama.

Frequency matrix merupakan distribusi kemunculan setiap residu asam amino dalam struktur primer. *Frequency matrix* berisi informasi mengenai komposisi protein. Selain itu, Accumulative

Absolute Position Incidence Vector atau AAPIV memberikan informasi mengenai komposisi protein. AAPIV dibentuk untuk mengekstrak informasi mengenai posisi residu asam amino dalam rantai polipeptida. Sebuah vektor dibentuk dengan 20 elemen sedemikian rupa sehingga setiap elemen menampung jumlah semua nilai ordinal di mana residu yang sesuai terjadi dalam struktur primer. Sebuah RAAPIV juga dibentuk untuk melakukan hal yang sama. RAAPIV dibangun dengan membalikkan string struktur utama dan kemudian mengekstraksi AAPIV dari string yang dibalik.

2.8. *k-fold cross-validation*

Cross-validation adalah metode yang digunakan untuk melakukan evaluasi kinerja model prediktif suatu model (Yadav & Shukla, 2016), dalam *cross-validation* data dibagi menjadi dua bagian, yaitu *training* dan *testing*. (Refaeilzadeh et al., 2016). Dalam proses pembangunan model menggunakan data *training* sedangkan untuk memvalidasi model menggunakan data *testing* (Novianti & Santosa, 2016). Pada penelitian ini, pembagian data akan menggunakan *k-fold cross-validation*, dimana k adalah bilangan bulat yang digunakan untuk membagi data (Faisal & Nugrahadi, 2019).



Gambar 4. 10-*fold cross-validation* (Berrar, 2018).

Gambar 4 merupakan contoh *k-fold cross-validation* dengan k=10, yang mana pada *fold* pertama, *subset* pertama digunakan sebagai set validasi dan

sisanya sebagai set pelatihan atau *training*. Pada *fold* kedua, *subset* kedua digunakan untuk set validasi dan sisanya untuk set pelatihan. Begitupun seterusnya sampai dengan *fold* sepuluh (Berrar, 2018).

2.9. *Confusion matrix*

Confusion matrix adalah sebuah tabel untuk mendeskripsikan performa dari sebuah model klasifikasi, yang digunakan untuk menghitung akurasi pada konsep data *mining* (Rosandy, 2016). Tabel *confusion matrix* dapat dilihat pada Tabel 6.

Tabel 6. Confusion matrix (Ting, 2017)

	Actual Positive	***Actual Negative***
Predicted positive	TP	FP
Predicted negative	FN	TN

Ada empat istilah yang merepresentasikan hasil proses klasifikasi pada *confusion matrix,* yaitu:

1. *True Positive* (TP)
   Nilai sebenarnya *positive* dan model memprediksi nilai *positive.*
2. *False Positive* (FP)
   Nilai sebenarnya *negative* dan model memprediksi nilai *positive*.
3. *True Negative* (TN)
   Nilai sebenarnya *negative* dan model memprediksi nilai *negative*.
4. *False Negative* (FN)
   Nilai sebenarnya *positive* dan model memprediksi nilai *negative*.

Parameter akurasi yang digunakan pada penelitian ini adalah sebagai berikut:

2.9.1. Akurasi

Akurasi adalah ukuran untuk berapa banyak prediksi yang diidentifikasi benar dengan jumlah semua kasus. Akurasi adalah metrik dasar yang digunakan untuk mengukur kinerja model (Ting, 2017). Akurasi mengukur kualitas prediksi untuk kejadian positif dan negatif dan direpresentasikan pada Persamaan 20, dimana TP (*True Postifive*), FP (*False Positive*), FN (*False Negative*), dan TN (*True Negative*) (Alkuhlani et al., 2020).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(20)}$$

2.9.2. Sensitivitas

Sensitivitas mewakili proporsi semua positif yang diprediksi dengan benar yang menggambarkan kemampuan model dalam melakukan prediksi positif (Chien et al., 2020). Untuk mendapatkan nilai evaluasi matrik sensitivitas dapat dihitung dengan Persamaan 21.

$$Sn = \frac{TP}{TP+FN} \quad \text{.....(21)}$$

2.9.3. Spesifisitas

Merupakan rasio prediksi negatif terhadap keseluruhan data negatif (Jahangiri et al., 2020). Rasio ini menunjukkan kemampuan model untuk memprediksi negatif dengan benar (Chien et al., 2020), dapat dihitung pada Persamaan 22.

$$SP = \frac{TN}{TN+FP} \quad \text{.....(22)}$$

2.9.4. *Matthew Correlation Coefficient*

Kinerja algoritma klasifikasi dapat diukur menggunakan metode yang disebut *Matthew Correlation Coefficient* (MCC). *Confusion matrix* dapat digunakan untuk klasifikasi Perhitungan awal MCC. Pada dasarnya *confusion matrix* berisi informasi dari perbandingan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi

yang seharusnya. Rentang *Matthew Correlation Coefficient* yaitu -1 hingga 1, yang mana -1 menyatakan klasifikasi biner yang sepenuhnya salah sedangkan 1 menyatakan klasifikasi biner yang sebenarnya (Hasdyna & Dinata, 2020). Ketika rasio positif terhadap negatif tidak genap maka indeks evaluasi yang cocok adalah MCC (Chien et al., 2020). Rumus dari MCC dapat dilihat pada Persamaan 23.

$$MCC = \frac{((TN*TP)-(FN*FP))}{\sqrt{(FP+TP)(FP+TN)(FN+TP)(FN+TN)}} \quad \text{.....(23)}$$

III. METODOLOGI PENELITIAN

3.1. Tempat dan Waktu Penelitian

Tempat dan waktu penelitian adalah sebagai berikut:

3.1.1. Tempat Penelitian

Lab RPL jurusan ilmu komputer fakultas matematika dan ilmu pengetahuan alam digunakan sebagai tempat penelitian ini, yang menggunakan beberapa komputer pada lab tersebut.

3.1.2. Waktu dan Jadwal Penelitian

Penelitian dilakukan pada bulan November 2021 hingga penyelesaian pada bulan mei 2022. Adapun beberapa kegiatan yang akan dilaksanakan dalam penelitian ini, diantaranya, pengumpulan dan pemahaman studi literatur serta pengumpulan *dataset*, yang selanjutnya melakukan penyusunan *draft* proposal bab 1-3 tahapan tersebut dilaksanakan dalam waktu pengerjaan kurang lebih 8 minggu. Kemudian melakukan tahapan pengerjaan program yang dimulai dari melakukan fitur ekstraksi data, melakukan pemodelan dan prediksi menggunakan *random forest*, evaluasi *confusion matrix*, dan yang terakhir yaitu melakukan penyusunan *draft* proposal bab 4-5. Untuk lebih jelasnya alur waktu pengerjaan penelitian dapat dilihat pada Tabel 7.

Tabel 7. Alur waktu pengerjaan penelitian

No	Kegiatan	2021																2022																							
		November				Desember				Januari				Februari				Maret				April				Mei				Juni				Juli							
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Penentuan tema dan pengumpulan studi literatur																																								
2	Pengumpulan *dataset*																																								
3	Penyusunan *draft* proposal (BAB I - III)																																								
4	*Feature extraction*																																								
5	*k-fold cross-validation*																																								
6	Pemodelan dan prediksi *random forest*																																								
7	Evaluasi *confusion matrix*																																								
8	Penyusunan draft proposal (BAB IV - V)																																								

Alur waktu pengerjaan penelitian memiliki 3 tahapan, yaitu.

1. Perancangan penelitian

   Pada tahapan ini dilakukan pengumpulan *dataset*, yang mana data diperoleh dari penelitian sebelumnya yang dilakukan oleh Akmal (2017). Data tersebut memiliki 2 jenis data, yaitu data negatif situs glikosilasi *N-linked* dan data positif situs glikosilasi *N-linked*.

2. Implementasi atau pelaksanaan penelitian

   Pada tahapan ini melakukan fitur ekstraksi, yang mana beberapa fitur akan diekstraksi, yaitu *raw, Hahn and central moment,* setelah itu melakukan pemodelan dan prediksi menggunakan *random forest,* serta melakukan validasi pengujian menggunakan *k-fold cross-validation.*

3. Evaluasi penelitian

   Pada tahap evaluasi penelitian menggunakan *confusion matrix,* dimana dalam penelitian ini untuk mengukur kinerja model menggunakan akurasi, sensitivitas, spesifisitas, dan *matthew correlation coefficient.*

3.2. Data dan Alat

Berikut merupakan penjelasan mengenai data dan alat yang digunakan dalam penelitian ini:

3.2.1. Data

Pada penelitian ini data yang digunakan berasal dari penelitian yang dilakukan oleh Akmal et al (2017) yang terdiri dari 2 *dataset,* yaitu *dataset* negatif dan *dataset* positif dengan atribut tipenya adalah situs glikosilasi *N-linked*. Terdapat 11.601 situs positif glikosilasi *N-linked* dan 12.160 situs negatif glikosilasi *N-linked*, sehingga jumlah *sequence* situs glikosilasi yang digunakan adalah 23.761 *sequence.* Dalam *dataset* tersebut panjang *sequence* adalah 41. Detail dari data yang akan digunakan dapat dilihat pada Tabel 8.

.

Tabel 8. Detail data situs glikosilasi *N-linked*

Tipe Atribut	Jenis Data	Jumlah *Sequence* situs glikosilasi
Situs glikosilasi *N-linked*	Positif	11.601
Situs glikosilasi *N-linked*	Negatif	12.160

3.2.2. Alat

Peralatan yang digunakan dalam penelitian ini adalah sebagai berikut:

3.2.2.1. *Hardware* (perangkat keras)

penelitian ini menggunakan perangkat keras laptop dengan spesifikasi sebagai berikut.

a) *Processor* : AMD A9-9425 Radeon R5, 5 Compute Cores 2C+3G 3.10 Ghz.
b) *Installed* RAM : 8.00 GB DDR4.
c) *Harddisk* : 1 TB.
d) *Network Interface* : Realtek RTL8723DE 802.11b/g/n PCIe Adapter.
e) *Video Graphics Array*: AMD Radeon™ R5 *Graphics.*

3.2.2.2. *Software* (perangkat lunak)

Penelitian ini menggunakan beberapa perangkat lunak, yaitu sebagai berikut.

a. Sistem Operasi : Windows 10 Home Single Language 64 bit.
b. *Tools* : R *programming version* 3.6.1 dan R studio *version* 1.2.5001.
c. *Library*

   Adapun beberapa *library* yang akan digunakan pada penelitian ini, yaitu sebagai berikut:

1. *Library* Caret 6.0-86

   Library Caret merupakan singkatan dari *classification and regression training* yang digunakan untuk membangun model prediksi dan mengukur hasil dari klasifikasi menggunakan *confusion matrix* (Kuhn, 2008).

2. *Library* MCCR 0.4.4

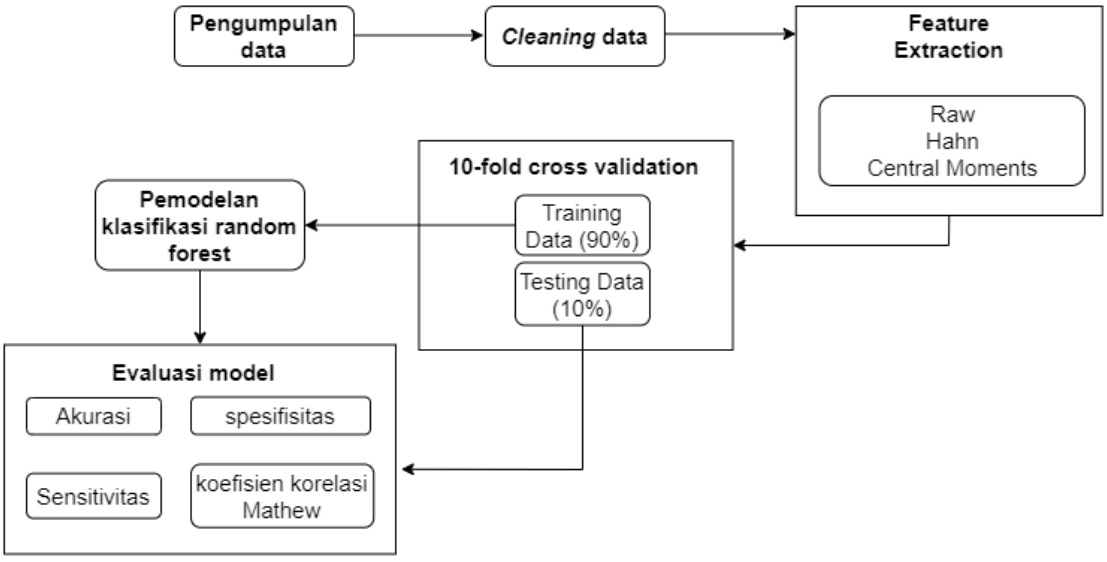
   Digunakan untuk menghitung nilai akurasi *matthew correlation coefficient* (MCC) yang menampilkan kualitas suatu metode klasifikasi tertentu.

3. Library RandomForest 4.6-14

   Package RandomForest digunakan pada klasifikasi dan regresi untuk melakukan pemodelan prediksi berdasarkan algoritma *random forest*. *Library* ini dikembangkan oleh Breiman and Adele Cutler.

3.3. Metodologi

Alur kerja penelitian ini didasari oleh penelitian Akmal et al (2017). *Flowchart* alur rencana penelitian dapat dilihat pada Gambar 5.



Gambar 5. Alur pengerjaan penelitian klasifikasi situs n-glikosilasi menggunakan random forest.

Berikut merupakan penjelasan mengenai alur penelitian pada Gambar 5 sebagai berikut:

3.3.1. Pengumpulan Data

Pada tahapan ini dilakukannya pengumpulan data, dimana data diperoleh dari penelitian sebelumnya yang dilakukan oleh Akmal et al (2017), yang berjudul *Prediction of N-linked glycosylation sites using position relative features and statistical moments.* Terdapat dua jenis data, yaitu *dataset* positif dengan jumlah 11.601 *sequence* dan *dataset* negatif dengan jumlah 12.160 *sequence.*

3.3.2. *Cleaning* Data

Tahap *cleaning* data pada penelitian ini dilakukan dengan menghapus *sequence* yang tidak termasuk kedalam jenis asam amino yang ditandai dengan adanya huruf B, J, O, U, X, dan Z. hal ini dilakukan supaya data dapat digunakan pada tahap selanjutnya.

3.3.3. *Feature Extraction*

Tahap ini adalah mengubah data *string* menjadi *numeric* atau angka. *Feature extraction* yang digunakan dalam penelitian ini adalah *statistical moment*s yang terdiri dari *raw, hahn* dan *central moment,* serta *position and composition variant features.*

3.3.4. *10-fold cross-validation*

Pada tahap ini pemisahan data dilakukan menggunakan 10-*fold cross-validation* dengan membagi data menjadi data *training* dan *testing*. Dimana, data *training* sebesar 90% digunakan untuk membangun model prediksi dan *testing* sebesar 10% digunakan untuk memvalidasi model prediksi.

3.3.5. Pemodelan Klasifikasi

Setelah melakukan tahap pemisahan data, pada data *training* selanjutnya dilakukan klasifikasi data, yang mana pada penelitian ini model klasifikasi yang akan digunakan adalah *random forest.*

3.3.6. Evaluasi Model

Keakuratan model prediksi digambarkan melalui *confusion matrix* diantaranya adalah *Koefisien Korelasi Matthew*, sensitivitas, spesifisitas dan akurasi.

V. SIMPULAN DAN SARAN

5.1. Simpulan

Pada penelitian klasifikasi situs n-glikosilasi yang telah dilakukan menggunakan metode *random forest* memperoleh beberapa kesimpulan sebagai berikut.

1. Dari percobaan seluruh ntree yang telah dilakukan memperoleh akurasi tertinggi pada *ntree* 750 dengan *mtry* 25, yakni akurasi sebesar 92,29%, sensitivitas 97,71%, spesifisitas 87,12%, dan MCC sebesar 85,12%.
2. Hasil klasifikasi *random forest* pada penelitian ini memperoleh akurasi yang lebih rendah dibandingkan dengan hasil dari penelitian sebelumnya yang dilakukan oleh Akmal et al., (2017) menggunakan metode *Neural Network Back Propagation* dengan akurasi sebesar 99,9%.

5.2. Saran

Adapun saran pada penelitian ini adalah sebagai berikut.

1. Pada penelitian ini dapat menggunakan metode klasifikasi lain seperti SVM, *Xgboost* ataupun K-*Nearest Neighbors* (KNN) untuk dapat digunakan sebagai pembanding dengan penelitian ini.
2. Penelitian ini dapat menggunakan fitur esktraksi lain seperti *Pseudo Amino Acid Composition* (PseAAC), *Composition, Transition, and Distribution* (CTD), *hydrophobicity* dan lain sebagainya.
3. Pada penelitian selanjutnya dapat menggunakan *feature selection* untuk memilih fitur terbaik dari kumpulan fitur yang digunakan.

.

DAFTAR PUSTAKA

Agus, R. 2018. *Dasar-Dasar Biologi Molekuler: Basics of Molecular Biology (IND SUB)*. CELEBES MEDIA PERKASA.

Akmal, M. A., Rasool, N., & Khan, Y. D. 2017. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PLoS ONE*, *12*(8), 1–21.

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. 2012. Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–278.

Alkuhlani, A., Gad, W., Roushdy, M., & Salem, A. M. 2020. *Artificial Intelligence for Glycation Site Prediction*. *September*.

Awais, M., Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., & Chou, K. C. 2021. IPhosH-PseAAC: Identify Phosphohistidine Sites in Proteins by Blending Statistical Moments and Position Relative Features According to the Chou's 5-Step Rule and General Pseudo Amino Acid Composition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 596–610.

Baniya, B. K., Ghimire, D., & Lee, J. 2015. Automatic music genre classification using timbral texture and rhythmic content features. *International Conference on Advanced Communication Technology, ICACT*, *2015-Augus*(3), 434–443.

Berrar, D. 2018. Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1–3*(January 2018), 542–545.

Binarwati, L., Mukhlash, I., & Soetrisno, S. 2017. Implementasi Algoritma Genetika untuk Optimalisasi Random Forest dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru : Studi Kasus PT.XYZ. *Jurnal Sains Dan Seni ITS*, *6*(2), 2–6.

Blom, N., Sicheritz-Pontén, T., Gupta, R., Gammeltoft, S., & Brunak, S. 2004. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, *4*(6), 1633–1649.

Breiman, L., Friedman, J., Olshen., R., & Stone, C. . 1984. Classification and Regression Trees. *Biometrics*, *40*(3), 874.

Breiman, L. 2003. Manual–Setting Up, Using, and Understanding Random Forests, v 4.0.

Breiman, Leo. 2001. Random Forests. *Machine Learning*, *12343 LNCS*, 503–515.

Breiman, Leo. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231.

Campbell, N. A., Reece, J. B., Urry, L. A., Wasserman, S. A., & Cain, M. L. 2008. *Biology*. Pearson, Benjamin Cummings.

Chauhan, J. S., Rao, A., & Raghava, G. P. S. 2013. In silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences. *PLoS ONE*, *8*(6), 1–10.

Chien, C. H., Chang, C. C., Lin, S. H., Chen, C. W., Chang, Z. H., & Chu, Y. W. 2020. N-GlycoGo: Predicting protein N-glycosylation sites on imbalanced data sets by using heterogeneous and comprehensive strategy. *IEEE Access*, *8*, 165944–165950.

Christy, E., Suryowati, K., Statistika, J., Sains Terapan, F., & AKPRIND Yogyakarta, I. 2021. Analisis Klasifikasi Status Bekerja Penduduk Daerah Istimewa Yogyakarta Menggunakan Metode Random Forest. *Jurnal Statistika Industri Dan Komputasi*, *6*(1), 69–76.

Faisal, R. M., & Nugrahadi, D. T. 2019. *Belajar Data Science dengan Pemrograman R* (Issue February).

Genuer, R., Poggi, J.-M., & Tuleau, C. 2008. *Random Forests: some methodological insights*.

Han, J. ;, Kamber, M., & Pei, J. 2012. Data mining: Data mining concepts and techniques. In *Data Mining Concepts and Techniques Third Edition*.

Hasdyna, N., & Dinata, R. K. 2020. Analisis Matthew Correlation Coefficient pada K-Nearest Neighbor dalam Klasifikasi Ikan Hias. *INFORMAL: Informatics Journal*, *5*(2), 57.

Hastie, T., Tibshirani, R., & Friedman, J. 2008. The Elements of Statistical Learning: Data-mining, Inference and Prediction. In *springer verlag* (Vol. 26, Issue 4).

Helenius, A., & Aebi, M. 2004. Roles of N-linked glycans in the endoplasmic reticulum. *Annual Review of Biochemistry*, *73*, 1019–1049.

Jahangiri, M., Jahangiri, M., & Najafgholipour, M. 2020. The sensitivity and specificity analyses of ambient temperature and population size on the transmission rate of the novel coronavirus (COVID-19) in different provinces of Iran. *Science of the Total Environment*, *728*, 138872.

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. 2020. RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *International Conference on Web Information Systems Engineering*, 503–515.

Khalid, S., Khalil, T., & Nasreen, S. 2014. A survey of feature selection and feature extraction techniques in machine learning. *Proceedings of 2014 Science and Information Conference, SAI 2014*, 372–378.

Khan, Y. D., Jamil, M., Hussain, W., Rasool, N., Khan, S. A., & Chou, K. C. 2019. pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *Journal of Theoretical Biology*, *463*, 47–55.

Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

Liu, Y., Wang, Y., & Zhang, J. 2012. New machine learning algorithm: Random forest. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7473 LNCS*, 246–252.

Marie, I. A., Hakim, L., Sugiarto, D., & Septiani, W. 2019. Perbandingan Performansi Teknik Klasifikasi Breakdown Mesin pada Proses Produksi Pembuatan Battery Mobil. *Jurnal Ilmiah Teknik Industri*, *18*(1), 33–41.

Novianti, T., & Santosa, I. 2016. PENENTUAN JADWAL KERJA BERDASARKAN KLASIFIKASI DATA KARYAWAN MENGGUNAKAN METODE DECISION TREE C4 . 5 (Studi Kasus Universitas Muhammadiyah Surabaya). *Komunikasi, Media Dan Informatika*, *5*(1).

Putra, D. S., Wibawa, A. D., & Purnomo, M. H. 2016. *KLASIFIKASI SINYAL EMG PADA OTOT TUNGKAI SELAMA BERJALAN MENGGUNAKAN RANDOM FOREST*. *1*(1), 51–56.

Refaeilzadeh, P., Tang, L., Liu, H., Angeles, L., & Scientist, C. D. 2016. Encyclopedia of Database Systems. *Encyclopedia of Database Systems*.

Rosandy, T. 2016. PERBANDINGAN METODE NAIVE BAYES CLASSIFIER DENGAN METODE DECISION TREE (C4.5) UNTUK MENGANALISA KELANCARAN PEMBIAYAAN (Study Kasus : KSPPS / BMT AL-FADHILA. *Jurnal Teknologi Informasi Magister Darmajaya*, *2*(01), 52–62.

Ruiz-Blanco, Y. B., Marrero-Ponce, Y., García-Hernández, E., & Green, J. 2016. Novel “extended sequons” of human N-glycosylation sites improve the precision of qualitative predictions: an alignment-free study of pattern recognition using ProtDCal protein features. *Amino Acids*, *49*(2), 317–325.

Saifullah, S., Zarlis, M., Zakaria, Z., & Sembiring, R. W. 2017. Analisa Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Processing Data. *J-SAKTI (Jurnal Sains Komputer Dan Informatika)*, *1*(2), 180.

Sartono, B., & Syafitri, U. D. 2010. Metode Pohon Gabungan: Solusi Pilihan Untuk Mengatasi Kelemahan Pohon Regresi Dan Klasifikasi Tunggal. *Forum Statistika Dan Komputasi*, *15*(1), 1–7.

Seprianto. 2017. *Modul Mata Kuliah Pengantar Bioinformatika*. *Ibt 431*, 1–86.

Simamora, A. (2015). Asam amino, Peptida, dan Protein. *Buku Ajar Blok 3 Biologi Sel*, *1*, 43.

Suhartono, E., Rohman, T., Setiawan, B., & Primasari, A. A. 2005. Model pembentukan Advance Glycation End Products dan modifikasi protein akibat reaksi glikosilasi. *Maj. Kedokt. Indon*, *55*(11), 681–685.

Suhartono, E., Setiawan, B., & Edyson, M. 2004. Modifikasi protein akibat reaksi Maillard dan pengaruhnya terhadap kadar tirosin. *Jurnal Profesi Medika*, *4*(2), 20–28.

Suhartono, M. T. 2017. *Protein - Serial Biokimia Mudah dan Menggugah*. Gramedia Widiasarana Indonesia.

Suprayitno, E., & Sulistiyati, T. D. 2017. *Metabolisme Protein*. Universitas Brawijaya Press.

Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y., & Campbell, M. P. 2019. SPRINT-Gly: Predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics*, *35*(20), 4140–4146.

Tahyudin, I., Putra, I. M., & Syafa’at, A. Y. 2021. *Data Mining Dan Data Warehouse Menggunakan Aplikasi KNIME*. Zahira Media Publisher.

Thomy, Z., & Harnelly, E. 2018. *Buku Ajar Dasar-Dasar Biologi Sel dan Molekuler : Buku untuk mahasiswa*. Syiah Kuala University Press.

Ting, K. M. 2017. Confusion Matrix. *Encyclopedia of Machine Learning and Data Mining*, *October*, 260–260.

Van Den Steen, P., Rudd, P. M., Dwek, R. A., & Opdenakker, G. 1998. Concepts and principles of O-linked glycosylation. *Critical Reviews in Biochemistry and Molecular Biology*, *33*(3), 151–208.

Yadav, S., & Shukla, S. 2016. Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification. *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, *Cv*, 78–83.

Yuliana, A., & Fathurohman, M. 2020. *TEORI DASAR DAN IMPLEMENTASI PERKEMBANGAN BIOLOGI SEL DAN MOLEKULER*. Jakad Media Publishing.