IMPLEMENTASI PENGGABUNGAN EKSTRAKSI FITUR DALAM MENINGKATKAN KLASIFIKASI *POST-TRANSLATIONAL MODIFICATION* (PTM) GLIKOSILASI PADA PROTEIN *SEQUENCE* N, O DAN C DENGAN METODE *EXTREME GRADIENT BOOSTING* (XGBOOST)

Disertasi

Oleh:

DAMAYANTI
NPM 1737061012

PROGRAM STUDI DOKTOR (S3) MIPA
PASCASARJANA
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

IMPLEMENTASI PENGGABUNGAN EKSTRAKSI FITUR DALAM MENINGKATKAN KLASIFIKASI *POST-TRANSLATIONAL MODIFICATION* (PTM) GLIKOSILASI PADA PROTEIN *SEQUENCE* N, O DAN C DENGAN METODE *EXTREME GRADIENT BOOSTING* (XGBOOST)

Oleh:

DAMAYANTI

Disertasi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
DOKTOR

Pada

Program Studi Doktor (S3) MIPA
Fakultas Matematika dan Ilmu Pengetahuan Alam

PROGRAM STUDI DOKTOR (S3) MIPA
PASCASARJANA
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2024

ABSTRAK

IMPLEMENTASI PENGGABUNGAN EKSTRAKSI FITUR DALAM MENINGKATKAN KLASIFIKASI *POST-TRANSLATIONAL MODIFICATION* (PTM) GLIKOSILASI PADA PROTEIN *SEQUENCE* N, O DAN C DENGAN METODE *EXTREME GRADIENT BOOSTING* (XGBOOST)

Oleh

DAMAYANTI

Post-translational modification (PTM) adalah salah satu mekanisme penting dalam mengatur fugsi protein. Modifikasi pasca translasi mengacu pada penambahan modifikasi protein kovalen dan enzimatik dalam biosintesis protein, yang memiliki peran penting dalam memodifikasi fungsi protein dan mengatur ekspresi gen. Salah satu modifikasi pasca translasi adalah glikosilasi. Glikosilasi adalah penambahan gugus gula ke struktur protein. Glikosilasi telah terkait dengan beberapa penyakit diantaranya diabetes, kanker, dan alzheimer. Oleh karena itu, penting untuk mengantisipasi terjadinya glikosilasi dengan melakukan prediksi glikosilasi.

Permasalahan dalam prediksi glikosilasi saat ini masih bergantung pada teknik laboratorium manual, yang menyebabkan proses prediksi menjadi lambat dan memerlukan biaya peralatan laboratorium yang tinggi. Untuk mengatasi hal tersebut, diperlukan pendekatan *machine learning* sehingga prediksi dapat dilakukan lebih cepat dan tidak membutuhkan biaya yang mahal.

Data yang digunakan dalam penelitian ini adalah data PTM glikosilasi-N, glikosilasi-O, dan glikosilasi-C yang diperoleh dari website UniProt yang tersedia secara terbuka. Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi modifikasi pasca translasi glikosilasi-N, glikosilasi-O, dan glikosilasi-C dengan menggabungkan 5 (lima) Ekstraksi fitur dan menggunakan algoritma Extreme Gradient Boosting (XGBoost).

Ekstraksi fitur terdiri dari: AAIndex, *Hydrophobicity*, SABLE, CTD, dan PseAAC. Seleksi fitur dilakukan dengan pendekatan MRMR. Masing-masing fitur memberikan kontribusi terhadap peningkatan prediksi glikosilasi. Fitur AAIndex memberikan kontribusi terbesar pada peningkatan prediksi glikosilasi-N secara keseluruhan sebesar 24%. Sedangkan, fitur SABLE memberikan kontribusi terbesar pada peningkatan prediksi glikosilasi-O sebesar 44%. Fitur *Hydrophobicity* dan PseAAC masing-masing berkontribusi sebesar 27% untuk pningkatan prediksi glikosilasi-C

Hasil penelitian ini menunjukkan kinerja prediksi modifikasi pasca translasi glikosilasi-N, glikosilasi-O dan glikosilasi-C dengan masing-masing nilai akurasi 100%. Pendekatan menggunakan XGBoost dalam penelitian ini berhasil meningkatkan akurasi sebesar 5% dibandingkan dengan penelitian sebelumnya.

Kata Kunci: *Post-translational modifications*, glikosilasi, *sequence*, xgboost, klasifikasi

ABSTRACT

THE IMPLEMENTATION OF FEATURE EXTRACTION FUSION IN ENHANCING THE CLASSIFICATION OF N-, O-, AND C-LINKED GLYCOSYLATION POST-TRANSLATIONAL MODIFICATIONS (PTMS) IN PROTEIN SEQUENCES USING EXTREME GRADIENT BOOSTING (XGBOOST) METHOD

By

DAMAYANTI

Post-translational modification (PTM) is one of the important mechanisms in regulating protein function. Post-translational modifications refer to the addition of covalent and enzymatic protein modifications in protein biosynthesis, which play a crucial role in modifying protein function and regulating gene expression. One of the post-translational modifications is glycosylation, which involves adding sugar groups to protein structures. Glycosylation has been associated with several diseases including diabetes, cancer, and Alzheimer's. Therefore, it is important to anticipate glycosylation by predicting it.

The current issue in glycosylation prediction still relies on manual laboratory techniques, resulting in slow prediction processes and requiring expensive laboratory equipment. To address this, a machine learning approach is needed so that predictions can be made faster and at a lower cost.

The data used in this study are PTM glycosylation-N, glycosylation-O, and glycosylation-C data obtained from the publicly accessible UniProt website. This research aims to improve the accuracy of classification of post-translational modifications glycosylation-N, glycosylation-O, and glycosylation-C by combining 5 feature extractions and using the Extreme Gradient Boosting (XGBoost) algorithm.

Feature extraction consists of: AAIndex, Hydrophobicity, SABLE, CTD, and PseAAC. Feature selection is performed using the MRMR approach. Each feature contributes to improving glycosylation prediction. The AAIndex feature contributes the most to the overall improvement in glycosylation prediction by 24%. Meanwhile, the SABLE feature contributes the most to the improvement in glycosylation-O prediction by 44%. The Hydrophobicity and PseAAC features contribute 27% each to the improvement of C-glycosylation prediction accuracy.

The results of this study show the performance of predicting post-translational modifications of glycosylation-N, glycosylation-O, and glycosylation-C, with each having an accuracy value of 100%. The approach using XGBoost in this study successfully increased the accuracy by 5% compared to previous research.

Keywords: Post-translational modification, glycosylation, site sequence, xgboost, classification

Judul Disertasi	:	Implementasi Penggabungan Ekstraksi Fitur dalam Meningkatkan Klasifikasi *Post-Ranslational Modification* (PTM) Glikosilasi Pada Protein *Sequence* N, O Dan C dengan Metode *Extreme Gradient Boosting* (XGBoost)
Nama Mahasiswa	:	Damayanti
No Pokok Mahasiswa	:	1737061012
Program Studi	:	Doktor MIPA
Fakultas	:	Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

Komisaris Promotor

Promotor

Prof. Dr. Satyaru[illegible], M.Biomed.
NIP. 195704241987031001

Ko-Promotor 1

Dr. rer. nat. Akmal Junaidi, M.Sc.
NIP. 197101291997021001

Ko-Promotor 2

Favorisen R. Lumbanraja S.Kom., M.Si., Ph.D.
NIP. 198301102008121002

Direktur Pascasarjana,

Prof. Dr. Ir. Murhadi, M.Si.
NIP. 196[illegible]1989021001

Ketua Program Studi S3 MIPA

Dr. G. Nugroho Susanto, M.Sc.
NIP. 196103111988[illegible]

MENGESAHKAN

1. Tim Penguji

Promotor	:	Prof. Dr. Sutyarso, M.Biomed.	(.....)
Ko-Promotor 1	:	Dr. rer. nat. Akmal Junaidi, M.Sc.	(.....)
Ko-Promotor 2	:	Favorisen R. Lumbanraja S.Kom., M.Si., Ph.D	(.....)
Sekretaris Prodi	:	Dr. Khoirin Nisa, S.Si., M.Si.	(.....)
Penguji Internal	:	Dr. G. Nugroho Susanto, M.Sc.	(.....)
Penguji Eksternal	:	Dr. Eng. Wisnu Ananta Kusuma, ST., MT.	(.....)

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian : 18 April 2024

LEMBAR IDENTITAS DAN PENGESAHAN

6. Judul Disertasi	:	Implementasi Penggabungan Ekstraksi Fitur dalam Meningkatkan Klasifikasi *Post-Ranslational Modification* (PTM) Glikosilasi Pada Protein *Sequence* N, O Dan C dengan Metode *Extreme Gradient Boosting* (XGBoost)
8. Bidang Ilmu	:	Ilmu Komputer
1. Ketua		
a. Nama	:	Damayanti
b. Gol/Pangkat/NPM	:	IIId/Penata Tk.I/1737061012
c. Jabatan Fungsional	:	Lektor
d. Fakultas/Bagian	:	Matematika dan Ilmu Pengetahuan Alam / Ilmu Komputer
e. Universitas	:	Universitas Lampung
2. Anggota Peneliti	:	4 (empat) orang
3. Lokasi Penelitian	:	Kota Bandar Lampung

Bandar Lampung, 18 April 2024

Mengetahui,
Dekan Fakultas MIPA Unila

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Peneliti,

Damayanti
NPM. 1737061012

RIWAYAT HIDUP

Penulis dilahirkan di Cahaya Negeri, Kecamatan Abung Barat Lampung Utara pada tanggal 4 Agustus 1977 sebagai anak pertama dari ayah yang bernama Darman dan ibu bernama Bahmawati. Penulis saat ini bertempat tinggal di Perumahan Cahaya Alam Permai Blok F No.15 Jalan Padat Karya, Rajabasa Jaya Bandar Lampung. Pendidikan yang telah ditempuh penulis Sekolah Dasar di SD N 05 Cahaya Negeri Lampung Utara diselesaikan pada tahun 1990. Selanjutnya penulis melanjutkan pendidikan tingkat menengah pertama di SMPN Ogan Lima Lampung Utara selesai pada tahun 1993. Kemudian penulis menyelesaikan pendidikan tingkat menengah atas di SMA N Gedong Tataan Lampung Selatan selesai pada tahun 1996. Pada tahun 2008 penulis menyelesaikan pendidikan strata satu Sarjana Komputer di STMIK Teknokrat Bandar Lampung. Pada tahun 2016 penulis menyelesaikan pendidikan Magister Komputer di Institut Teknologi Sepuluh Nopember Surabaya.

Pada tahun 2009 sampai dengan sekarang penulis bekerja sebagai dosen di Universitas Teknokrat Indonesia Bandar Lampung.

Pada tahun 2017 penulis terdaftar sebagai mahasiswa Doktor MIPA Universitas Lampung.

Selama menjadi mahasiswa penulis melakukan kegiatan pembelajaran, penelitian dan pengabdian kepada masyarakat. Prestasi hibah yang diperoleh penulis di bidang Pembelajaran, Penelitian dan PKM diantaranya sebagai berikut:

1. Tahun 2017 penulis mendapatkan hibah penelitian DRPM dengan judul penelitian Analisis Pemanfaatan Teknologi Informasi dan Komunikasi Terhadap Minat dan Hasil Belajar Mahasiswa pada skema Penelitian Dosen Pemula (PDP) untuk tahun pelaksanaan 2018.
2. Tahun 2018 penulis memperoleh hibah Pengabdian Kepada Masyarakat DRPM dengan judul *E-Marketing* Perjalanan Wisata Kelompok Nelayan

Wisata Desa Batu Menyan Kabupaten Pesawaran Lampung dengan skema PKMS tahun pelaksanaan 2019.

3. Tahun 2019 penulis mendapatkan hibah penelitian DRPM pada skema Penelitian Dosen Pemula (PDP) dengan judul Pengembangan Teknologi Web Sebagai Alat Bantu Auditor dalam Pengukuran Penyelarasan Teknologi dan Bisnis tahun pelaksanaan 2020.
4. Tahun 2019 penulis mendapatkan hibah PKM DRTPM pada skema PKMS dengan judul Penerapan Teknologi Tabungan untuk Siswa di SD Ar-Raudah Suka Jawa Tanjung Karang Barat Bandar Lampung Tahun pelaksanaan Tahun 2020.
5. Tahun 2021 penulis mendapatkan hibah penelitian dengan skema Penelitian Disertasi Doktor (PDD) dengan judul Pengembangan Model *Extreme Gradient Boosting* Untuk Peningkatan Kinerja Prediksi Protein Glikosilasi Tahun pelaksanaan 2022.
6. Tahun 2022 penulis mendapatkan hibah pelaksanaan Kurikulum Pembelajaran Daring Kolaboratif (PDK) dari Kemendikbud berimitra dengan Universitas Hamzanwadi Nusa Tenggara Barat.

Selanjutnya penulis juga melakukan publikasi karya ilmiah diantaranya sebagai berikut:

1 Publikasi artikel ilmiah pada prosiding Internasional *Scopus* penerbit IOP *Publishin*g, dengan judul *E-CRM Information System for Tapis Lampung SMEs* tahun 2019.

2 Publikasi artikel ilmiah pada prosiding internasional *Scopus* penerbit IEEE, dengan judul *Assessment of The Alignment Maturity Level of Business and Information Technology at CV Jaya Technology* tahun 2019.

3 Publikasi artikel ilmiah pada prosiding Nasional penerbit Universitas Trisakti, dengan judul Analisis Pemanfaatan Teknologi Informasi dan Komunikasi Terhadap Minat Belajar Mahasiswa tahun 2018.

4 Publikasi artikel ilmiah pada Prosiding Seminar Nasional Darmajaya, dengan judul Penerapan Teknologi Tabungan untuk Siswa di SD Ar-Raudah Bandar Lampung tahun 2020.

5 Publikasi artikel ilmiah pada jurnal Nasional SINTA 2 dengan judul Sistem Informasi Manajemen Penggajian dan Penilaian Kinerja Pegawai pada SMK Taman Siswa Lampung, Jurnal Teknologi Informasi dan Ilmu Komputer, tahun 2019

6 Publikasi artikel ilmiah pada jurnal Nasional SINTA 2, dengan judul Game Edukasi Pengenalan Hewan Langka Berbasis *Android* Menggunakan *Construct 2*, Jurnal Teknologi Informasi dan Ilmu Komputer, tahun 2020

7 Publikasi artikel ilmiah pada jurnal Nasional SINTA 4 dengan judul Rancang Bangun Sistem Pengukuran Keselarasan Teknologi Dan Bisnis Untuk Proses Auditing, Jurnal Tekno Kompak, tahun 2020.

8 Publikasi artikel ilmiah pada jurnal Nasional SINTA 4 dengan judul Aplikasi Permainan Sebagai Media Pembelajaran Peta Dan Budaya Sumatera Untuk Siswa Sekolah Dasar, Jurnal Komputasi, tahun 2021.

9 Publikasi artikel ilmiah pada jurnal Nasional SINTA 4 dengan judul Analisis dan Perancangan Sistem Informasi Akuntansi Pengelolaan Tabungan Siswa pada SD Ar-Raudah Bandarlampung, Jurnal Teknologi dan Informasi, tahun 2021.

10 Publikasi artikel ilmiah pada jurnal PKM dengan judul Pelatihan Digital Marketing Bagi Pemuda-Pemudi Karang Taruna Di Desa Kunjir Lampung Selatan, Journal of Social Sciences and Technology for Community Service (JSSTCS), tahun 2022.

11 Publikasi artikel ilmiah pada Jurnal Internasional dengan judul *Model Classification for Predicting the Post-Translational Modification (PTM) Glycosylation in Sequence O Using an Extreme Gradient Boosting Algorithm, Journal of Computer Science* dengan status *Accepted*, tahun 2024.

12 Publikasi artikel ilmiah pada Jurnal Internasional dengan judul *Performance Evaluation of Feature Extraction to Improve the Classification of Post-*

Translational Modification in C-Glycosylation Using XGBoost, *Journal Bulletin of Electrical Engineering and Informatics*, tahun 2024 dengan status in *Review*.

13 Publikasi artikel ilmiah pada Jurnal Internasional dengan judul *A New Feature Extraction Approach in Classification for Improving the Accuracy of Protein, International Journal on Informatics Visualization*, tahun 2024 dengan status *submite*.

HALAMAN PERSEMBAHAN

Alhamdulillah ya Allah, atas rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan laporan disertasi ini. Sholawat dan salam sanjungkan kepada Nabi Muhammad SAW.

Karya ini saya persembahkan kepada:

Yayasan Pendidikan Teknokrat
Universitas Teknokrat Indonesia
Terima kasih yang sebesar-besarnya saya ucapkan kepada Yayasan Pendidikan Teknokrat dan Universitas Teknokrat Indonesia yang telah memberikan dukungan penuh baik secara *financial* maupun moril terhadap penyelesaian karya ini.

Suamiku Mirhansyah
Terima kasih yang sebesar-besarnya kepada suamiku yang sangat memotivasi dan mendukung dalam penyelesaian karya ini

Anakku Eka Putri Adamansyah dan Mutiara Ahadia Syifa
Terima kasih telah memberikan dukungan penuh terhadap penyelesaian karya ini.

Ayahku Darman dan Ibuku Bahmati
Terima kasih yang setingi-tingginya kepada orang tuaku yang senanatiasa mendo'akan dan mendukungku

Naurah Nazhifah adik tingkatku
Terima kasih atas bantuan dan dukungan terhadap penyelesaian karya ini

Terima kasih kepada seluruh keluarga besarku, dosen dan staf Universitas Teknokrat Indonesia, sahabat dan rekan-rekan semua.

Terima kasih atas dukungan dan bantuan dalam penyelesaian karya ini.

Almamater Tercinta, Universitas Lampung

KATA PENGANTAR

Alhamdulilah puji syukur kehadirat Allah SWT karena atas rahmatnya penulis dapat menyelesaikan disertasi ini. Penulisan disertasi ini dilakukan dalam rangka memenuhi salah satu syarat untuk mendapatkan predikat Doktor di FMIPA Universitas Lampung. Penulis menyadari bahwa, tanpa bantuan dan bimbingan dari berbagai pihak, sangatlah sulit bagi penulis untuk menyelesaikan disertasi ini. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Ibu Prof. Dr. Ir. Lusmelia Afriani, D.E.A., IPM. selaku Rektor Universitas Lampung.
2. Bapak Dr. Eng. Heri Satria, S.Si., M.Si., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
3. Bapak Dr. G. Nugroho Susanto M.Sc. selaku Ketua Program Studi Doktor MIPA Universitas Lampung dan pembahas I yang telah meluangkan waktu membahas dan membimbing untuk penyempurnaan disertasi ini.
4. Ibu Dr. Khoirin Nisa, S.Si., M.Si., selaku Sekretaris Prodi Doktor MIPA Universitas Lampung yang telah memotivasi dan mengarahkan penulis dengan sangat baik. Sehingga disertasi ini dapat diselesasikan dengan baik.
5. Prof. Dr. Sutyarso, M.Biomed., selaku promotor I yang telah memotivasi dan membimbing penulis dengan sangat baik. Sehingga disertasi ini dapat diselesasikan dengan baik.
6. Bapak Dr. rer. nat. Akmal Junaidi, M.Sc. selaku Kopromotor I yang telah meluangkan waktu untuk membimbing dan mengarahkan penulis menyelesaikan disertasi ini.
7. Bapak Favorisen R. Lumbanraja S.Kom., M.Si., Ph.D., selaku Kopromotor II yang telah meluangkan waktu untuk membimbing, memotovasi dan memberi banyak masukan kepada penulis dalam penyelesaikan disertasi ini.
8. Ibu Hj. Hernaini, SS., M.Pd. selaku Pembina Yayasan Pendidikan Teknokrat yang telah memberikan dukungan dan motivasi hingga terselesaikannya disertasi ini.

9. Ibu Dewi Sukmasari, SE., MSA., CA., Akt. Selaku Ketua Yayasan Pendidikan Teknokrat yang telah memberikan dukungan dan motivasi hingga terselesaikannya disertasi ini
10. Bapak Dr. H.M. Nasrullah Yusuf, S.E., M.B.A., selaku Rektor Universitas Teknokrat Indonesia yang telah memberikan motivasi dan dukungan kepada penulis dalam penyelesaian disertasi ini.
11. Bapak Dr. H. Mahathir Muhammad, S.E., M.M., selaku Dekan Fakultas Teknik dan Ilmu Komputer Universitas Teknokrat Indonesia yang telah memberikan motivasi, dukungan dan arahan kepada penulis dalam penyelesaian disertasi ini.
12. Bapak/Ibu Dosen Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang telah mendukung penulis menyelesaikan disertasi ini.
13. Rekan-rekan mahasiswa Doktor MIPA Universitas Lampung yang telah banyak memberikan dukungan dan semangat dalam penyelesaian disertasi ini.
14. Bapak/Ibu Tenaga Kependidikan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung yang telah memberikan layanan terbaiknya kepada penulis untuk dokumen administrasi.
15. Bapak/Ibu Dosen Universitas Teknokrat Indonesia yang telah banyak memberikan dukungan dan semangat kepada penulis dalam menyelesaikan disertasi ini.
16. Kepada rekan-rekan mahasiswa/mahasiswi Universitas Teknokrat Indonesia dan rekan-rekan mahasiswa/mahasiswi Universitas Lampung yang telah memberi bantuan dan dukungan dalam penyelesaian disertasi ini.

Disertasi ini tentunya masih jauh dari kata sempurna karena keterbatasan pengetahuan dan pengalaman penulis. Untuk itu penulis sangat mengharapkan kritik dan saran untuk perbaikan disertasi ini. Terima kasih kepada semua pihak yang telah membantu dalam penyelesaian disertasi ini. Semoga disertasi yang telah

disusun ini dapat bermanfaat untuk pengembangangan ilmu pengetahuan selanjutnya.

Bandar Lampung, 18 April 2024

Damayanti
NPM. 1737061012

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

Gambar

DAFTAR TABEL

Halaman

Tabel

DAFTAR KODE PROGRAM

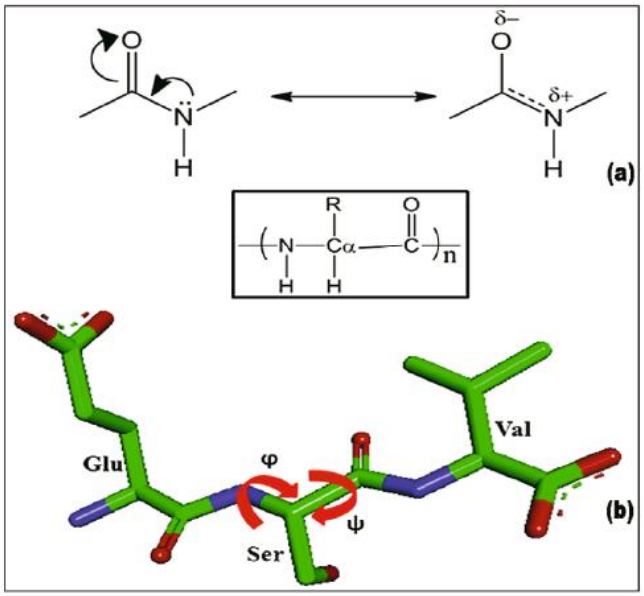
Halaman

Kode Program

I. PENDAHULUAN

1.1 Latar Belakang

Protein merupakan makromolekul yang penting dalam kehidupan. Protein adalah biopolimer heteromerik linier yang terdiri dari asam amino yang terhubung secara kovalen melalui ikatan amida (Guruprasad, 2019). Protein merupakan ikatan rantai-rantai asam amino yang mengikat antara satu sama lain dalam ikatan peptide. Struktur dasar asam amino terdiri dari atom karbon (C) yang terikat pada gugus amino (-NH2), gugus karboksil (-COOH), atom hidrogen (-H) dan R (bervariasi) tergantung dari jenis asam aminonya (Fujii *et al.*, 2018). Struktur asam amino ditunjukkan pada Gambar 1.



Gambar 1. Struktur Dasar Asam Amino dan Struktur Primer Protein (Guruprasad, 2019).

Protein memiliki fungsi dalam tubuh untuk membantu perkembangan sel dan menjaga pertahanan tubuh. Tubuh manusia terdiri dari 42% protein. Protein dapat memiliki banyak fungsi seperti sirkulasi oksigen ke seluruh tubuh, melawan infeksi,

memindahkan zat masuk dan keluar sel, mengendalikan reaksi kimia dan mengirimkan pesan dari satu bagian tubuh ke bagian tubuh lainnya. Protein juga berfungsi dalam organisme dengan mengkatalisis reaksi metabolisme, replikasi DNA, dan pergerakan molekul dari satu tempat ke tempat lain (Kadakeri *et al.*, 2020).

Protein memainkan peran penting dalam semua proses biologis. Protein adalah polimer linier yang terdiri dari unit monomer yang disebut asam amino. Rantai polipeptida memiliki empat tingkat struktur protein berbeda yang berbeda satu sama lain. Struktur protein terdiri dari struktur primer, sekunder, tersier dan kuaterner (Kadakeri *et al.,* 2020). Proses protein terdiri dari transkripsi dan translasi. Selama translasi, protein disintesis berdasarkan pola informasi genetik (RNA) dalam ribosom dan organel intraseluler, dan secara spontan terlipat menjadi struktur tiga dimensi (3D) (Kadakeri *et al.,* 2020). Pelipatan protein yang stabil sering terjadi bersamaan dengan translasi protein. Beberapa protein diangkut ke organel yang berbeda, atau setidaknya terlepas dari ribosom dan diangkut ke sitoplasma sebelum dilipat. Pelipatan protein terkadang dibantu oleh pendamping protein dan enzim yang membentuk ikatan. Hanya struktur protein asli yang terlipat dengan benar yang menunjukkan aktivitas biologis. Protein berfungsi sebagai enzim, hormon, dan antibodi. Banyak penyakit seperti alzheimer, parkinson, katarak, fibrosis kistik, dan lain-lain yang disebabkan oleh protein yang salah lipatan (Kadakeri *et al.,* 2020).

Polipeptida hasil translasi tidak langsung aktif, untuk menjadi protein aktif atau berfungsi dalam sel, maka harus melalui tahapan *Post-Translational Modification* (PTM). PTM adalah perubahan yang terjadi pada struktur protein setelah menyelesaikan dan pelepasan polipeptida dari ribosom. PTM adalah mekanisme penting yang terlibat dalam pengaturan fungsi protein (Caragea *et al.*, 2007). *Modifikasi pasca-translasi* mengacu pada penambahan kovalen dan enzimatik modifikasi protein selama atau setelah biosintesis protein, yang memainkan peran

penting dalam memodifikasi fungsi protein dan mengatur ekspresi gen (Minguez *et al.*, 2013).

PTM terdiri dari berbagai macam seperti fosforilasi, glikosilasi, ubiquitinasi, nitrosilasi, metilasi, asetilasi, dan lipidasi (Qiu *et al.*, 2016). Salah satu *modifikasi pasca-translasi* yaitu glikosilasi. Glikosilasi adalah modifikasi protein pasca-translasi dalam sel eukariotik (Yang dan Han 2017) yang mempengaruhi berbagai proses biologis seperti pelipatan protein, interaksi sel-sel, dan respon imun (Pitti *et al.*, 2019). Glikosilasi adalah salah satu modifikasi pasca-translasi yang paling umum dalam pemrosesan protein eukariotik (Zhang dan Sun, 2020).

Glikosilasi merupakan modifikasi pasca-translasi yang melibatkan penambahan kovalen karbohidrat ke protein. Karbohidrat berupa glukosa, gula atau sakarida yang memiliki struktur linier dan bercabang yang mengandung monosakarida dihubungkan secara kovalen (Mazola *et al.*, 2011). Glikosilasi merupakan salah satu modifikasi translasi yang paling beragam dibandingkan modifikasi pasca translasi lainnya Weerapana dan Imperiali, (2006) atau paling komplek diantara modifikasi pasca translasi yang lain karena melibatkan penempelan glikan pada situs spesifik pada protein (Zhang *et al.*, 2016). Glikosilasi juga berperan penting dalam berbagai proses biologis, termasuk pemrosesan ligan protein di dalam sel, termasuk sekresi, antigenisitas, metabolisme glikoprotein, pelipatan protein, dan imunitas. Glikosilasi dikelompokkan menjadi empat jenis yaitu; glikosilasi-N, glikosilasi-O, glikosilasi-C, dan glikosilasi GPI (*glycosylphosphatidylinositol*) (Caragea *et al.*, 2007), (Yang dan Han, 2017), (Pitti *et al.*, 2019). Glikosilasi-N adalah glikan yang terikat pada nitrogen tengah asparagin (N), yang secara khusus terjadi pada urutan N. Asam amino merupakan unit dasar protein dan berperan penting dalam pengaturan berbagai proses yang berkaitan dengan ekspresi gen dan berperan penting dalam pembentukan protein. Protein terdiri dari rantai panjang asam amino yang dihubungkan oleh ikatan peptide (Akram *et al.*, 2011). Glikosilasi-O berkontribusi besar dalam pencegahan berbagai penyakit seperti penyakit tulang dan saraf (Ohtsubo dan Marth, 2006). Pada penelitian Everest-Dass

et al. (2018) mengenai pengamatan perubahan struktur parsial rantai gula pada berbagai penyakit diantaranya *glycomics* kanker, penyakit saraf dalam konteks *glyco*, fitur *glycan* diabetes, fitur struktural *glycan* antibodi. Perubahan glikosilasi yang diamati pada berbagai penyakit penting untuk memahami perkembangan penyakit dan kemajuan terapi (Everest-Dass *et al.*, 2018). Telah banyak pendekatan yang digunakan untuk mendukung penelitian prediksi glikosilasi diantaranya pada penelitian Zhang *et al.*, (2016) yang menganalisis *glycan* dalam menentukan fitur *glycan* seperti situs glikosilasi, struktur *glycan* dan konten. Hasil penelitian menunjukkan bahwa *platform* lektin *microarray* memungkinkan untuk mengetahui hasil yang tinggi dalam menentukan dari ada atau tidaknya varian g*lycan* dalam *sample lipoprotein* (Zhang *et al.*, 2016).

Permasalahan prediksi glikosilasi adalah seperti metode eksperimen, sehingga membutuhkan waktu yang cukup lama serta membutuhkan biaya pemeliharaan alat yang mahal. Dengan demikian untuk mengatasi permasalahan tersebut dibutuhkan pembangunan model *machine learning* untuk memprediksi glikosilasi otomatis. Salah satu metode *machine learning* yaitu *Extreme Gradient Boosting* (XGBoost).

Saat ini, penelitian tentang modifikasi pasca translasi protein sedang menarik banyak perhatian. Ketersediaan data komputasi menggunakan pembelajaran mesin membuka kemungkinan untuk memprediksi glikosilasi dengan lebih cepat, memberikan metode yang dapat diandalkan untuk mengurangi biaya pemeliharaan peralatan yang mahal. *Machine learning* adalah pendekatan kecerdasan buatan yang memungkinkan komputer untuk belajar dan meningkatkan kinerjanya dalam melakukan tugas tertentu secara otomatis, tanpa perlu diprogram secara eksplisit (Reddy, *et al*, 2018). *Machine learning* memiliki beberapa fungsi dalam penerapannya antara lain klasifikasi dan prediksi. Proses yang dilakukan oleh pembelajaran mesin meliputi pelatihan. *Machine learning* didefenisikan sebagai metode komputasi berdasarkan pengalaman untuk meningkatkan kinerja atau prediksi yang akurat. *Machine learning* disebut dengan pelatihan, sehingga

proses *machine learning* memerlukan data untuk dipelajari disebut data pelatihan (Tanaka dan Okutomi, 2014).

Beberapa penelitian telah menggunakan data komputasi termasuk penelitian yang dilakukan oleh Ma *et al.* (2020) menggunakan pendekatan XGBoost untuk mengklasifikasikan kanker awal dan kanker akhir. Hasil penelitian menunjukkan bahwa metode yang diterapkan lebih kompetitif dibandingkan metode klasifikasi sebelumnya, sehingga model XGBoost direkomendasikan sebagai pendekatan klasifikasi yang baik untuk mencapai kinerja tinggi (Ma *et al.,* 2020). XGBoost adalah kerangka peningkatan gradien yang dikembangkan oleh Friedman *et al.* (2000) dan cenderung lebih efisien (Chen dan He, 2014). XGBoost mencakup penyelesaian model linier dan *tree learning* (Chen *et al.,* 2018). Algoritma XGBoost mendukung berbagai macam fungsi diantaranya regresi, klasifikasi dan perangkingan. Saat ini XGBoost merupakan algoritma yang populer dalam menjawab tantangan *machine learning*. Metode *extreme gradient boosting* memiliki kinerja yang tinggi dalam melakukan prediksi (Ma *et al.,* 2020).

Penelitian ini membangun model XGBoost untuk meningkatkan hasil akurasi prediksi glikosilasi. Beberapa penelitian yang relevan yang telah dikembangkan sebelumnya adalah penelitian Li *et al.* (2015) memprediksi glikosilasi-N, Glikosilasi-O pada data *benchmark* dan *independent* menggunakan metode *random forest.* Hasil penelitian menunjukkan bahwa akurasi dari masing-masing data tersebut sebesar 95.00%. Penelitian Pitti *et al.,* (2019) memprediksi glikosilasi-N menggunakan metode *Support Vector Machine* (SVM) menunjukkan hasil akurasi sebesar 74.00%. Lebih lanjut, dalam penelitian Chien *et al.* (2020) memprediksi situs N-glikosilasi sehubungan dengan data positif dan negatif. Dalam penelitian ini, sekuens protein dan sifat asam amino digunakan untuk memprediksi glikosilasi-N dengan akurasi 94.60% menggunakan pemodelan XGBoost. Selanjutnya, penelitian oleh Alkuhlani *et al.* (2022) menghasilkan akurasi 95.11% dalam memprediksi glikosilasi-N menggunakan metode PUStackNGly. Berdasarkan data

dari beberapa penelitian yang telah dikembangkan masih memiliki peluang untuk peningkatan kinerja prediksi glikosilasi.

Penelitian ini bertujuan untuk mengolah data glikosilasi-N, glikosilasi-O, dan glikosilasi-C menggunakan fitur baru dan pendekatan agoritma XGBoost untuk memprediksi apakah data tersebut terglikosilasi atau tidak terglikosilasi. Model XGBoost diharapkan dapat meningkatkan kinerja prediksi *Post-Translational Modification* glikosilasi-N, glikosilasi-O, dan glikosilasi-C. Selanjutnya hasil kinerja penelitian ini akan dibandingkan dengan penelitian yang telah dikembangkan sebelumnya. Berdasarkan sepengetahuan penelitian belum ada penelitian yang membahas prediksi glikosilasi menggunakan XGBoost dengan akurasi yang melebihi akurasi dari penelitian sebelumnya. Penelitian ini penting dilakukan sebagai dasar yang dapat digunakan untuk pengembangan obat pada bidang klinis. Hampir semua protein pada sel manusia dan mamalia lainnya terglikosilasi, namun bila terjadi penyimpangan proses glikosilasi dapat menyebabkan berbagai penyakit diantaranya kanker, alzheimer's, penyakit menular patobiologi, dan lain-lain. Penyakit alzheimer adalah penyakit pada otak dimana terjadi penurunan daya ingat pada penderita Regan *et al.* (2019). Beberapa bakteri dipengaruhi oleh glikosilasi-N dan glikosilasi-O untuk fungsi biologis. Masing-masing mengandung banyak glikosilasi yang harus diproses dan dimodifikasi agar aktif. Proses glikosilasi memainkan peran penting untuk kekebalan tubuh. Dengan memahami pentingnya glikosilasi dalam kelangsungan hidup maka dapat digunakan dalam pengembangan obat (Vigerust, 2011).

1.2 Rumusan Masalah

Adapun rumusan masalah pada penelitian ini adalah:

1. Apakah teknik ekstraksi fitur menggunakan AAIndex, *Hydrophobicity*, SABLE, CTD, PseAAC dan seleksi fitur menggunakan pendekatan MRMR mempengaruhi tingkat akurasi prediksi *post-translational modification* glikosilasi-N, glikosilasi-O, dan glikosilasi-C ?

2. Berapa perbandingan kinerja prediksi *post-translational modification* (PTM) glikosilasi-N, glikosilasi-O, dan glikosilasi-C menggunakan pendekatan algoritma *Extreme Gradient Boosting* (XGBoost) apabila dibandingkan dengan penelitian yang telah dikembangkan?

1.3 Tujuan Penelitian

Penelitian ini menggunakan fitur baru dan membangun algoritma *Extreme Gradient Boosting* (XGBoost) untuk meningkatkan akurasi prediksi *post-translational modification* PTM glikosilasi-N, glikosilasi-O, dan glikosilasi-C. Tujuan penelitian ini adalah sebagai berikut:

1. Menganalisis hasil ekstraksi fitur menggunakan AAIndex, *Hydrophobicity*, SABLE, CTD, dan PseAAC untuk peningkatan prediksi PTM glikosilasi-N, glikosilasi-O, dan glikosilasi-C.
2. Menganalisis hasil seleksi fitur menggunakan pendekatan *Minimum Redundancy Maximum Relevance* (MRMR)
3. Menganalisis performa prediksi PTM glikosilasi-N, glikosilasi-O, glikosilasi-C menggunakan pendekatan algoritma XGBoost
4. Membandingkan hasil kinerja predikasi glikosilasi-N, glikosilasi-O, glikosilasi-C dengan metode sebelumnya.

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut;

1 Penggunaan metode *Extreme Gradient Boosting* (XGBoost).

2 Batasan data yang digunakan yaitu data *post-translational modification* glikosilasi yang diperoleh dari web UniProt https://www.uniprot.org/ yang terdiri dari:

   a. Glikosilas-N yaitu gula yang terikat pada residu *asparagine* dalam rantai polipeptida

b. Glikosilasi-O yaitu gula yang terikat pada residu serin atau treonin dalam rantai polipeptida.
c. Glikosilasi-C yaitu gula yang ditambahkan ke karbon pada rantai samping triptofan

3 Data glikosilasi terdiri dari data *benchmark* dan *independent*,masing-masing data terdiri dari data:
a. *Benchmark* N berjumlah 873
b. *Benchmark* O berjumlah 1.362
c. *Benchmark* C berjumlah 142
d. *Independent* N berjumlah 218
e. *Independent* O berjumlah 334
f. *Independent* C berjumlah 38

1.5 Kontribusi Penelitian

Penelitian ini memiliki kontribusi secara keilmuan atau ilmu pengetahuan berupa:

1. Ekstraksi fitur menggunakan 5 macam protein deskriptor sebagai berikut:
   a *Amino Acid Index* (AAIndex)
   b *Hydrophobicity*
   c *Solvent AccessiBiLitiEs* (SABLE)
   d *Composition, Transition, and Distribution* (CTD)
   e *Pseudo Amino Acid Composition* (PseAAC)
2. Seleksi fitur menggunakan pendekatan *Minimum Redundancy Maximum Relevance* (MRMR).
3. Membangun model dengan algoritma *Extreme Gradient Boosting*
4. Pengembangan ilmu pengetahuan terkait teori tentang algoritma *Extreme Gradient Boosting* untuk memprediksi glikosilasi pada penelitian selanjutnya.
5. Pengembangan pengetahuan untuk penelitian selanjutnya dalam bidang klinis untuk terapi gen dan pengembangan obat.

1.6 Keterbaruan (*Novelty*)

Kebaruan utama pada penelitian ini adalah sebagai berikut:

1. Fitur baru SABLE dan *Hydrophobicity* untuk peningkatan akurasi prediksi PTM glikosilasi-N, glikosilasi-O, dan glikosilasi-C.
2. Pendekatan seleksi fitur *Minimum Redundancy Maximum Relevance* (MRMR).
3. Pendekatan menggunakan agoritma XGBoost untuk peningkatan akurasi prediksi PTM glikosilasi-N, glikosilasi-O, dan glikosilasi-C

Kebaruan di atas diharapkan dapat meningkatkan akurasi prediksi glikosilasi sehingga dapat memiliki manfaat sebagai berikut:

a. Fitur dan model yang digunakan memiliki kemampuan yang lebih baik untuk mengidentifikasi lokasi dan jenis glikosilasi yang mungkin terikat pada protein dengan tepat. Hal ini memungkinkan peneliti untuk mendapatkan informasi yang lebih akurat tentang posisi dan sifat glikosilasi pada protein.
b. Akurasi yang lebih tinggi dalam prediksi glikosilasi memungkinkan peneliti untuk menghindari eksperimen penggunaan laboratorium manual. Dengan memanfaatkan model prediksi yang akurat, peneliti dapat memilih kandidat protein yang berpotensi terjadi glikosilasi dengan lebih cepat. Hal ini memungkinkan dapat meningkatkan efisiensi dalam prediksi glikosilasi.
c. Tingkat akurasi yang lebih tinggi dalam prediksi glikosilasi memungkinkan pemahaman yang lebih mendalam tentang dampak glikosilasi pada struktur dan fungsi protein sehingga dapat digunakan untuk bidang klinis untuk terapi gen dan pengembangan obat.

II. TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Penelitian ini akan memberikan kontribusi pengembangan ilmu pengetahuan khususnya algoritma XGBoost dengan prediksi protein Glikosilasi. Beberapa penelitian telah dilakukan dan relevan dengan penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Penelitian Terdahulu Terkait Studi Klasifikasi Glikosilasi

No	Penelitian	Data	Metode	Hasil
1.	GlycoMine: *a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome* (Li *et al.*, 2015)	Data Glikosilasi-N, Glikosilasi-O dan Glikosilasi-C *Benchmark* *Dataset* Jumlah: 2701 Negatif: 1793 Positif: 908 *Independent Dataset* Jumlah: 667 Negatif: 452 Positif: 225 Panjang sequence: 15 Redundansi menggunakan CTD (30%)	Ekstraksi Fitur 1. AAIndex 2. *Physicochemical properties of proteins position-specific* 3. *Scoring matrices* (PSSMs) 4. *Residue conversation score* Seleksi Fitur *Minimum redundancy Maximum Relevance* (MRMR) Metode: *Random Forest* Evaluasi: *Cross validation k-fold* 5 kali	Hasil penelitian : *Benchmark* C: 100% N: 95.00% O: 86.61% *Independent* C: 100% N: 95.58% O: 90.70%
2.	N-GlyDE: *a two-stage N-linked*	Data Glikosilasi-N *Independent*	Metode *Support Vector Machine* (SVM)	Hasil *Accuracy*: 74.00%

	glycosylation site prediction incorporating gapped dipeptides and pattern-based encoding (Pitti *et al.* 2019)	Positif glikosilasi: 2050 Negatif Glikosilasi: 1030		MCC 49.00% *Precision*: 61.00% *Sensitivity*: 82.00% *Respectively*: 68.00%
3.	N-GlycoGo: *Predicting Protein N-Glycosylation Sites on Imbalanced Data Sets by Using Heterogeneous and Comprehensive Strategy* (Chien *et al.* 2020)	Data *Independent* Glikosilasi-N: Positif: 3836 Negatif: 18277 Panjang *sequence*: 21 Redundansi menggunakan CTD (30%)	Ekstraksi Fitur Binary, AAIndex, AAC, CKSAAP, Kmer, PC-PseAAC, SC-PseAAC, Motif, RSA/ASA, SS, and Signal Seleksi Fitur *Minimum Redundancy Maximum Relevance* (MRMR) Metode: *Random Forest* Evaluasi: Validation k-fold 5 kali	Hasil *Accuracy*: 94.60%
4.	*PUStackNGly: Positive-Unlabeled and Stacking Learning for N-Linked Glycosylation Site Prediction* (Alkuhlani *et al.* 2022)	Data Glikosilasi-N Data Positif: 1989 Negatif: 13737 Panjang Sequence: 25 Redundansi menggunakan CTD (30%),	Ekstraksi Fitur AAC, EAAC, CKSAAP, DDE, GEAAC, KSAAGP, GDPC, GTPC, AAC, APAAC, and PseKRAAC Seleksi Fitur *Minimum Redundancy*	Hasil menunjukkan bahwa data *independent* dengan akurasi: 95.11%

			Maximum Relevance (MRMR) Metode: PUStackNGly Uji validasi *cross validation*	
5	*Prediction of O-Glycosylation Site Using Pre-Trained Language Model And Machine Learning* (Alkuhlani *et al.* 2023)	Data sequence O diambil di O-*lycoprotein repository* (OGP) Panjang Sequence: 31 Redundansi menggunakan CTD (30%)	Ekstraksi Fitur menggunakan *Support Vector Machine* (SVM Seleksi Fitur TAPE (*Tasks Assessing Protein Embeddings*) Metode: XGBoost Evaluasi: Uji validasi *cross validation k-fold* 10 kali	Hasil menunjukkan bahwa nilai akurasi 77.86%

Beberapa penelitian terdahulu yang relevan dengan penelitian yang dilakukan yaitu penelitian Li *et al.* (2015) yang membahas prediksi glikosilasi dengan menggunakan metode *random forest.* Hasil penelitian menunjukkan bahwa prediksi glikosilasi-N untuk data *benchmark* dan *independent* masing-masing bernilai 95.00.0%. Selanjutnya penelitian Pitti *et al. (*2019) memprediksi glikosilasi-N dengan menggunakan metode SVM dengan hasil *accuracy* sebesar 74.00%. Penelitian Chien *et al.* (2020) memprediksi glikosilasi-N dengan menggunakan metode XGBoost dengan hasil akurasi sebesar 94.60%. Kemudian penelitian (Alkuhlani *et al.,* 2022) memprediksi glikosilasi-N dengan menggunakan metode PUStackNGly. Hasil penelitian menunjukkan bahwa nilai akurasi 95.11%

Penelitian Ahkulani *et al*. (2023) menggunakan algoritma XGBoost dengan nilai akurasi sebesar 77,86%. Berdasarkan beberapa penelitian terdahulu yang telah dilakukan untuk tingkat akurasi prediksi glikosilasi masih dapat ditingkatkan kembali. Kemudian untuk prediksi glikosilasi-N, glikosilasi-O, glikosilasi-C masih tergolong minim sehingga masih membuka peluang bagi peneliti untuk melakukan penelitian lebih lanjut dengan tingkat akurasi yang lebih tinggi dibandingkan penelitian sebelumnya.

2.2 Biosintesis Protein

Biosintesis protein oleh ribosom pertama kali menghasilkan polipeptida yang belum selesai. Dalam banyak kasus, beberapa rantai polipeptida harus berkumpul menjadi konformasi tiga dimensi melalui lipatan yang tepat untuk menjadi protein fungsional. Pelipatan protein yang tepat di dalam sel dimediasi oleh protein khusus yang disebut pendamping. Pengawal tampaknya berfungsi dengan mengikat dan menstabilkan polipeptida yang tidak terlipat atau sebagian terlipat untuk mencegah lipatan yang salah dan membiarkan rantai *polipeptida* terlipat ke dalam konformasi yang benar. Pengikatan pendamping dapat terjadi selama translasi di mana rantai polipeptida yang baru lahir sedang diterjemahkan pada ribosom, sehingga mencegah lipatan atau agregasi yang salah dari bagian terminal-amino dari polipeptida sebelum sintesis rantai selesai (Misran dan HaniffJaafar, 2018). Biosintesis protein merupakan proses biologis dimana asam amino dirakit oleh ikatan peptida menjadi urutan polipeptida tertentu sesuai yang dikodekan oleh asam deoksiribonukleat (DNA) (Brostrom dan Brostrom, 2007). Biosintesis memiliki tahapan yang sangat kompleks diantaranya:

2.2.1 Replikasi DNA

DNA merupakan materi genetik yang membawa informasi biologis dari setiap makhluk hidup. DNA merupakan molekul yang sangat penting bagi makhluk hidup, DNA berisi informasi genetik makhluk hidup yang menentukan jenis apa makhluk hidup tersebut. DNA

menentukan jenis makhluk hidup berupa bakteri, hewan, tumbuhan atau manusia. DNA digandakan untuk diwariskan pada keturunannya dari sel induk ke sel anakan dari individu induk ke keturunannya. Replikasi DNA bertujuan agar sel anak mengandung DNA yang identik dengan DNA induknya. Proses penggandaan DNA ini dinamakan dengan replikasi DNA. Model replikasi DNA terdapat tiga model yaitu: Model konservatif merupakan model kedua untai DNA induk berfungsi sebagai cetakan, kedua model semi konservatif yaitu dimana kedua untai DNA induk berpisah, masing-masing membuat untai baru sebagai pelengkapnya. Model ketiga, yaitu model dispersif yaitu barru sebagai beberapa bagian DNA induk secara tersebar berfungsi sebagai cetakan kemudian masing-masing bagian tersebut membuat bagian DNA baru sebagai pelengkap. Enzim yang digunakan dalam replikasi DNA (Brostrom dan Brostrom, 2007) adalah :

1 Enzim helikase yang berfungsi memutuskan ikatan hidrogen untuk membuka rantai ganda heliks DNA hingga menjadi 2 rantai tunggal
2 RNA primase yang berfungsi membentuk RNA primer sebagai awal pembentukan rantai baru DNA
3 DNA polimerase yang berfungsi menggabungkan nukleotida menjadi polimer DNA yang panjang
4 DNA ligase yang berfungsi menyambungkan fragmen DNA (fragmen okazaki) menjadi untaian DNA lengkap (Brostrom dan Brostrom, 2007)

2.2.2 Transkripsi DNA

Transkripsi adalah proses dimana DNA membentuk molekul RNA. Transkripsi adalah pembentukan RNA dari DNA dan merupakan proses pertama dalam sintesis protein (ekspresi gen). Proses transkripsi mengikuti proses penerjemahan. Enzim RNA polimerase terlibat dalam proses transkripsi. Transkripsi adalah proses replikasi DNA menjadi mRNA, atau proses pencetakan mRNA oleh DNA (Hamidah *et al.*, 2020).

2.2.3 Translasi

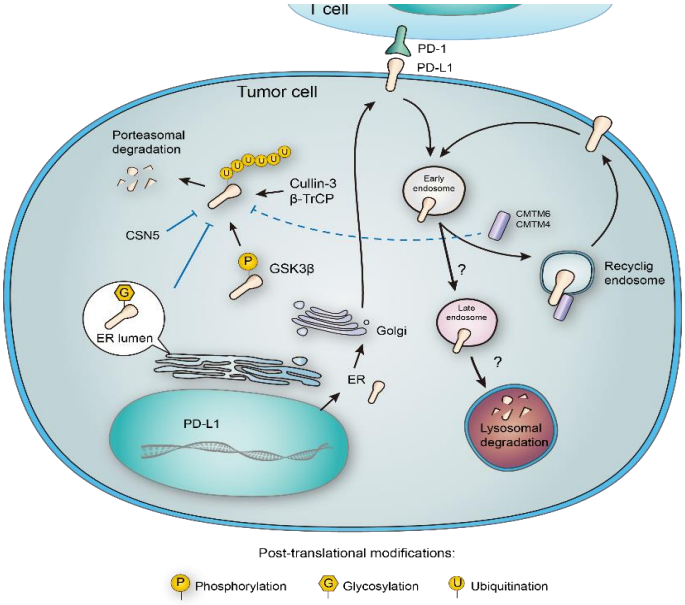
Translasi adalah proses pembentukan mRNA dari protein. Translasi adalah penerjemahan rangkaian nukleotida mRNA menjadi rangkaian asam amino, yang kemudian membentuk polipeptida/protein. Terjemahan adalah urutan dasar mRNA yang memberikan informasi yang diperlukan untuk tRNA (Hamidah *et al.,* 2020).

2.3 *Post-Translational Modification* (PTM)

Modifikasi pasca-translasi merupakan mekanisme penting yang terlibat dalam pengaturan fungsi protein (Caragea *et al.,* 2007). *Modifikasi pasca-translasi* mengacu pada penambahan kovalen dan enzimatik modifikasi protein selama atau setelah *biosintesis* protein, yang memainkan peran penting dalam memodifikasi fungsi protein dan mengatur ekspresi gen (Minguez *et al.,* 2013). PTM terdiri dari berbagai macam seperti: fosforilasi, glikosilasi, *ubiquitinasi*, nitrosilasi, metilasi, asetilasi, dan lipidasi (Qiu *et al.,* 2016). Sebagai salah satu *modifikasi pasca-translasi* disebut dengan glikosilasi.

PTM memainkan peran yang sangat penting dalam proses seluler karena beberapa organisme eukariotik menjalani glikosilasi (Mann dan Jensen, 2003). Sebagian besar protein dalam *modifikasi pasca-translasi* ini memainkan peran penting dalam sintesis dan pergantian protein, metabolisme nitrogen, siklus sel, dan lain-lain (Macek *et al.*, 2019). PTM sangat penting untuk memahami biologi sel, diagnosis dan pencegahan penyakit. Rantai *polipeptida* disintesis dalam sitoplasma sel melalui proses yang disebut translasi. Pengembangan membutuhkan polipeptida yang dapat digunakan setelah translasi seperti pelipatan yang disederhanakan, pembentukan kompleks, pengiriman ke berbagai bagian sel, atau modifikasi kovalen gugus fungsi. Beberapa modifikasi terjadi di sitoplasma dan penambahan residu gula atau polisakarida (Acar *et al.,* 2015). *Modifikasi pasca-translasi* (PTM) adalah biomekanisme kimia di mana residu asam amino dalam protein dimodifikasi

secara kovalen (Prabakaran *et al.*, 2012). Ilustrasi PTM dapat dilihat pada Gambar 2.

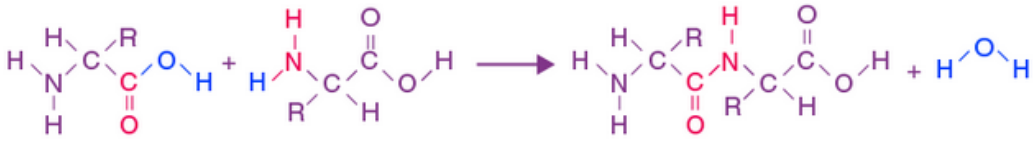


Gambar 2 Ilustrasi *Post-Translational Modification (PTM)* (Wong et al., 2018).

2.4 Protein

Protein adalah rantai asam amino yang disatukan melalui ikatan peptide (Fujii *et al.*, 2018). Protein memiliki fungsi menyusun protoplasma dan struktur tubuh yang berbentuk enzim atau hormon. Mekanisme sintesis pada protein terbagi menjadi dua tahapan, tahap transkripsi dan translasi. Transkripsi adalah pencetakan mRNA oleh DNA, sedangkan translasi adalah penerjemahan kode oleh tRNA berupa urutan yang dikehendaki. Translasi pada sistesis protein mengacu pada perakitan protein dalam sel yang melibatkan ribosom dimana RNA diterjemahkan untuk menghasilkan rantai asam amino (Fujii *et al.*, 2018). Polipeptida hasil translasi tidak langsung aktif, untuk menjadi protein aktif atau fungsional dalam sel dapat dilakukan *modifikasi pasca-translasi* atau yang sering disebut dengan *Post Translational Modification* (PTM). *Modifikasi pasca-translasi* adalah perubahan yang terjadi pada struktur protein setelah menyelesaikan dan pelepasan polipeptida dari ribosom. Protein merupakan makromolekul hal yang sangat penting dalam

proses biologis (Kadakeri *et al.*, 2020). Ilustrasi pembentukan ikatan peptide dapat dilihat pada Gambar 3.



Gambar 3. Pembentukan Ikatan Peptida (Rodnina et al, 2006)

2.5 Asam Amino

Asam amino adalah dasar dari protein yang mengandung gugus amino dan gugus karboksil. Asam amino berperan penting dalam mengatur proses ekspresi gen, termasuk mengatur fungsi protein. Asam amino diperlukan untuk pembentukan protein. Kekurangan asam amino dapat menghambat sintesis protein, sehingga menyebabkan defisiensi protein (Kimball dan Jefferson, 2006). Asam amino terdapat kurang lebih 300 jenis asam amino. Namun hanya dua puluh asam amino yang merupakan bahan untuk penguat protein. Berikut daftar asam amino dalam penyusun protein (Sumardjo, 2006). Nama-nama asam amino ini akan digunakan untuk melihat urutan sekuen protein untuk mengidentifikasi pola-pola dataset yang digunakan. Nama-nama asam amino nampak pada Tabel 2.

Tabel 2. Daftar Asam Amino dalam Penyusun Protein

No.	Nama Asam Amino	Singkatan	Lambang
1	Glisin	Gly	G
2	Alanin	Ala	A
3	Valin	Val	V
4	Leusin	Leu	L
5	Isoleusin	lle	I
6	Fenilalanin	Phe	F
7	**Triptofan**	**Try**	**W**
8	Tirosin	Tyr	Y
9	Serin	Ser	S
10	**Tronin**	**Thr**	**T**

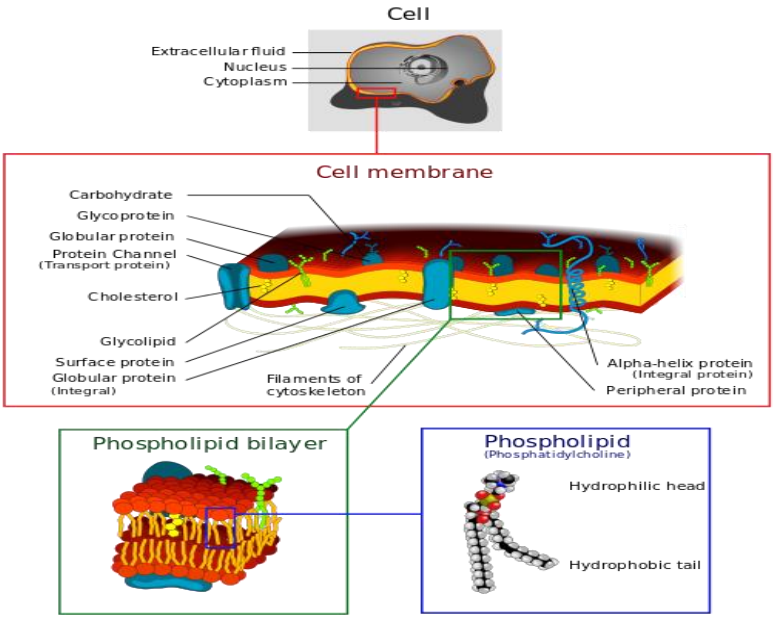
11	Proline	Pro	P
12	Sistein	Cys	C
13	Metionin	Met	M
14	Asam aspartat	Asp	D
15	Asam Glutamat	Glu	E
16	Lisin	Lys	K
17	Arigin	Arg	R
18	Histidin	His	H
19	**Asparagin**	**Asn**	**N**
20	Glutamin	Gln	Q

Asam amino adalah senyawa organik dengan gugus fungsi karboksilat (-COOH) dan amina (biasanya -NH2) dan rantai samping (gugus R) yang spesifik untuk setiap jenis asam amino. gugus karboksil dan amina yang terikat pada atom karbon (C). Gugus karboksil memberikan sifat asam dan gugus amina memberi sifat basa (El-Sayed, 2020).

2.6 Glikosilasi

Glikosilasi adalah salah satu modifikasi translasi protein yang paling kompleks. Glikosilasi polipeptida dipengaruhi oleh glikoprotein glikan seperti glikosilasi-N, glikosilasi-O, dan glikosilasi-C. Glikosilasi merupakan salah satu modifikasi protein pasca translasi pada sel eukariotik (Yang dan Han, 2017). Glikosilasi mempengaruhi berbagai proses biologis seperti pelipatan protein, interaksi sel, dan respon imun (Pitti *et al.,* 2019).

Glikosilasi adalah proses dalam penambahan gugus-gugus gula pada struktur protein. Glikosilasi adalah proses dimana karbohidrat melekat pada makromolekul seperti protein. Glikoprotein adalah molekul protein yang melekat pada rantai pendek karbohidrat yang terjadi pada membran sel dan juga dalam darah. Glikoprotein berfungsi sebagai antigen (Ohtsubo dan Marth, 2006). Membran sel ditunjukkan pada Gambar 4.



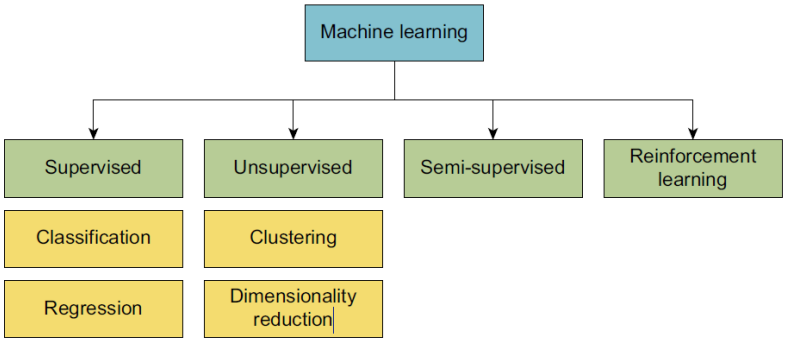
Gambar 4. Membran Sel (Ohtsubo dan Marth, 2006).

Glikosilasi merupakan salah satu translasi yang memiliki keragaman paling banyak dibandingkan dengan modifikasi pasca translasi yang lainnya (Weerapana dan Imperiali, 2006) atau paling komplek diantara modifikasi pasca translasi yang lain karena melibatkan penempelan glikan pada situs spesifik pada protein (Zhang *et al.*, 2016). Glikosilasi juga memainkan peran penting dalam berbagai proses biologis termasuk antigenisitas, metabolisme glikoprotein, pelipatan protein dan imunitas, termasuk pemrosesan ligan protein di dalam sel. Glikosilasi dibagi menjadi empat jenis: glikosilasi-N, glikosilasi,O glikosilasi-C, dan GPI-glikosilasi (glikosilfosfatidilinositol) (Caragea *et al.*, 2007), (Yang dan Han, 2017), (Pitti *et al.*, 2019).

2.7 *Machine Learning*

Machine learning merupakan bagian dari kecerdasan buatan (AI) yang sering digunakan untuk meniru perilaku manusia dan memecahkan masalah secara otomatis. *Machine learning* bertujuan membuat program komputer dan belajar data yang diinputkan (Vieira *et al.* 2019). Pembelajaran mesin termasuk area kecerdasan buatan yang berkaitan dengan mengidentifikasi pola dari data dan menggunakan

pola tersebut untuk memprediksi data yang ada. Berbagai cara atau pendekatan yang digunakan untuk menghasilkan algoritma pembelajaran mesin. Metode pembelajaran mesin dikelompokkan menjadi empat jenis pembelajaran yaitu: *supervised, unsupervised, semi supervised dan reinforcement learning*, terlihat pada Gambar 5. (Vieira *et al.*, 2019).



Gambar 5. Bagan Jenis Pembelajaran dalam *Machine Learning* (Vieira et al. 2019).

Jenis pembelajaran dalam pembelajaran mesin terbagi menjadi 4 jenis:

1. Pembelajaran *Supervised*

   Pembelajaran *supervised* merupakan model dilatih pada data yang tidak berlabel dan model secara otomatis belajar dari data (Dridi *et al.*, 2015). Dalam pendekatan *supervised learning* merupakan algoritma yang memiliki akses untuk memprediksi target. Misalnya ada atau tidak penyakit, keparahan gejala. Tujuan algoritma ini untuk mempelajari fungsi yang paling optimal untuk memprediksi fitur target. Alasan disebut dengan *supervised* karena algoritma ini memiliki pengetahuan sebelumnya tentang keluaran yang seharusnya. Pembelajaran *supervised* terdapat dua bagian yaitu klasifikasi dan regresi (Dridi *et al.*, 2015).

   a. Klasifikasi

      Algoritma klasifikasi bertujuan untuk memprediksi keanggotaan pada kelompok yang dikenal sebagai label atau kelas, untuk serangkaian pengamatan. Algoritma klasifikasi merupakan algoritma paling umum

yang digunakan dalam penelitian *machine learning*. Jenis algoritma ini sebagian besar digunakan untuk penyelesaian masalah klinis untuk diagnosis penyakit. Algoritma klasifikasi untuk mengelompokkan pasien yang sakit atau yang sehat (Vieira *et al.,* 2019). Model klasifikasi menggunakan fungsi pemetaan yang ditemukan oleh model dari dataset latihan untuk memprediksi kelas. Dalam klasifikasi, setiap contoh dalam dataset dikategorikan ke dalam kelas-kelas yang telah ditentukan sebelumnya (Dridi *et al.*, 2015).

b. Regresi

Metode regresi merupakan metode analisis statistik yang dapat digunakan untuk mencari hubungan antara dua fitur atau lebih. Metode regresi adalah teknik analisis statistik yang tujuannya untuk menghitung hubungan sebab akibat antara dua atau lebih fitur yang berbeda (Yang, 2020). Selain itu definisi regresi adalah teknik *supervised* yang dapat digunakan untuk menemukan korelasi antar fitur dan memprediksi nilai kontinu berdasarkan fitur yang ada. Dalam regresi, X (fitur *input*) dipetakan ke Y (*output continue*) (Dridi *et al.*, 2015).

2. Pembelajaran *Unsupervised*

Pembelajaran *unsupervised* tidak ada nilai target atau label yang ditetapkan pada algoritma pembelajaran sehingga dapat menemukan sendiri struktur yang mendasari dalam data. Pembelajaran *unsupervised* terdapat dua kategori yaitu pengelompokan dan pengurangan dimensi. Tujuan utama *unsupervised* adalah penemuan pola tersembunyi dan menarik dalam data yang tidak berlabel. *Clustering* adalah konsep penting dalam pembelajaran *unsupervised* yang menemukan struktur atau pola dalam kumpulan data yang tidak terklasifikasi. Algoritma *clustering* merupakan *unsupervised* mengelompokkan objek data ke dalam kelompok (*cluster*) berdasarkan karakteristik atau pola yang mirip dari objek tersebut (Dridi, 2015).

3. Pembelajaran Semi-*Supervised*

   Pembelajaran semi *supervised* digunakan dalam mengatasi masalah yang memungkinkan model untuk mengintegrasikan data tak berlabel yang tersedia dalam *supervised.* Semi *supervised* adalah salah kombinasi dari pembelajaran *supervised* dan *unsupervised.* Pembelajaran *supervised* membutuhkan jumlah data yang besar untuk mengklasifikasikan data uji dan memakan waktu. Di sisi lain, pembelajaran *unsupervised* tidak memerlukan data berlabel yang mengelompokkan data berdasarkan kesamaan titik data. Pembelajaran semi *supervised* untuk mengatasi masalah yang bisa dipelajari dengan sedikit data latih dan dapat diberikan label. Pembelajaran semi *supervised* membuat model dengan beberapa pola yang berlabel sebagai data pelatihan dan sisanya sebagai data uji (C A Padmanabha Reddy *et al.,* 2018).

4. Pembelajaran *Reinforcement Learning*

   Reinforcement Learning adalah pembelajaran mesin yang membangun sistem untuk dapat belajar guna memperoleh pengetahuan dan membuat keputusan melalui interaksi dengan lingkungan yang dinamis. Pada algoritma *reinforcement learning* sistem bertindak dalam lingkungan untuk menerima umpan balik berupa *reward* atau pinalti dan menggunakan pengalaman tersebut untuk mempelajari kebijakan (*policy*) yang optimal untuk mencapai tujuan yang ditentukan (Vieira *et al.,* 2019).

Machine learning setidaknya memiliki dua kemampuan dalam aplikasinya yaitu klasifikasi dan prediksi. Proses yang dilakukan *machine learning* meliputi pelatihan, pembelajaran sering disebut training sehingga pada proses machine learning ini membutuhkan data untuk dipelajari atau yang sering dikenal dengan data training (Tanaka dan Okutomi, 2014). Klasifikasi merupakan metode yang dilakukan oleh mesin dalam memilih dalam mengelompokkan objek berdasarkan ciri tertentu seperti saat manusia mampu membedakan benda yang satu dengan yang lainnya. Sedangkan prediksi adalah kemampuan yang dilakukan oleh mesin

untuk menerka dari suatu data yang telah diinputkan dan berdasarkan data yang sudah dipelajari oleh mesin. Beberapa metode *machine learning* yang digunakan diantaranya sistem pengambilan keputusan, *support vector machine* dan *neural network* (Tanaka dan Okutomi, 2014). Penelitian ini direkomendasikan menggunakan *Extreme Gradient Boosting* yang diharapkan mampu memperoleh kinerja tinggi pada analisis *machine learning*.

2.9 Ekstraksi Fitur

Salah satu langkah terpenting dalam penelitian pembelajaran mesin adalah tahap ekstraksi fitur. Ekstraksi fitur proses pengidentifikasian dan pemilihan fitur yang relevan dari data yang terkait dengan glikosilasi protein. Glikosilasi adalah proses penambahan gugus gula ke protein dan penelitian tentang glikosilasi sering kali melibatkan analisis urutan asam amino protein. Ekstraksi fitur bertujuan untuk meningkatkan kinerja akurasi dalam prediksi protein glikosilasi (Guo *et al.,* 2011). Ada lima ekstraksi fitur yang digunakan yaitu:

1 *Amino Acid Index (AAIndex)*

Paket yang digunakan untuk ekstraksi ciri menggunakan Amino Acid Index (AAIndex) yaitu paket *BioSeqClass* dengan fungsi *featureAAIndex().*

AAIndex dihitung dengan cara menghitung atribut fisikokimia dan biokimia pada asam amino berdasarkan basis data AAIndex dalam bentuk matriks yang menginformasikan setiap asam amino terhadap berbagai indeks atau parameter. Misalnya parameter *AAIndex.name="all",* maka parameter tersebut akan menghitung semua atribut AAIndex (Kawashima *et al.,* 2008). AAIndex terdapat 21 dimensi.

2 *Hydrophobicity*

Paket yang digunakan untuk ekstraksi fitur menggunakan hidrofobik adalah paket *BioSeqClass* dengan *fiturHydro(). Hydrophobicity* terdapat 21 dimensi (Kyte dan Doolittle, 1982)

3 *Solvent AccessiBiLitiEs* (SABLE)

SABLE adalah situs yang digunakan untuk prediksi struktur umum dalam mengidentifikasi lipatan yang paling cocok dari urutan tertentu. SABLE bertujuan untuk prediksi fitur struktural protein, seperti struktur sekunder dan aksesibilitas pelarut yang relatif. SABLE memiliki 21 dimensi. Rumus yang digunakan untuk menghitung SABLE dapat menggunakan Persamaan (1).

$$RSA_i = 100 \frac{SA_i}{MSA_i} = [\%] \quad (1)$$

Fungsi SA_i menunjukkan luas permukaan residu asam amino yang terpapar oleh pelarut dalam suatu struktur protein, sedangkan MSA_i menunjukkan luas permukaan maksimal yang bisa terpapar oleh pelarut untuk residu asam amino tertentu (Wagner *et al.,* 2005).

4 *Composition, Transition, and Distribution* (CTD)

Paket yang digunakan untuk ekstraksi fitur menggunakan *Composition, Transition, and Distribution* (CTD) yaitu paket *BioSeqClass* dengan fungsi *featureCTD().* CTD adalah fitur yang digunakan untuk memprediksi lokasi protein yang akan ditargetkan. Fungsi CTD telah berhasil digunakan dalam banyak penelitian tentang struktur dan fungsi protein (Govindan dan Nair, 2011). Terdapat 21 dimensi untuk ekstraksi fitur CTD. Rumus menghitung *composition* dapat menggunakan Persamaan (2).

$$Composition = \frac{Ne}{N} \quad (2)$$

Fungsi Ne menunjukkan jumlah asam amino sifat tertentu sedangkan N menunjukkan jumlah asam amino (Govindan dan Nair, 2011). Sedangkan rumus perhitungan *transition* dapat menggunakan Persamaan (3).

$$Transition = \frac{Nnm+Nmn}{N-1} \quad (3)$$

Fungsi $N_{nm}+N_{mn}$ menunjukkan jumlah peptide, sedangkan N merupakan panjang asam amino (Govindan dan Nair, 2011).

5 *Pseudo Amino Acid Composition* (PseAAC)

Pada ekstraksi fitur ini, paket yang digunakan adalah paket *BioSeqClass* dengan fungsi *featurePseudoAAComp().*

PseAAC terdapat serangkaian informasi lebih dari 20 atribut, dimana 20 atribut pertama merepresentasikan urutan asam amino. Atribut tambahan lainnya adalah gabungan informasi urutan komponen asam amino semu (Zare *et al.*, 2015). Dimensi untuk ekstraksi fitur PseAAC berjumlah 24.

Lima ekstraksi fitur diatas terdapat tiga pemilihan ekstraksi fitur diantaranya PseAAC, CTD, AAIndex didasarkan pada penelitian sebelumnya (Li *et al.*, 2015). SABLE dan *Hydrophobicity* merupakan keterbaruan dari penelitian ini.

2.10 Data Duplikat

Data merupakan hal yang sangat penting dalam sebuah penelitian, agar dapat diproses maka data harus dibersihkan terlebih dahulu sebelum masuk ke tahap pemrosesan. Pembersihan data berfungsi untuk mendapatkan data yang berkualitas tinggi dan mendapat data yang yang andal dan tidak bias (Kapil Kumar, 2018). Salah satu cara membersihkan data diantaranya menghilangkan data duplikat atau redudansi data. Data *sequence* protein berpotensi memiliki kesamaan, sehingga data yang sama harus dihapus. *Software* yang digunakan untuk menghapus data duplikat pada data set *sequence* yaitu menggunakan *skiperedudant.*

2.11 Seleksi Fitur

Tahap seleksi fitur bertujuan untuk mendapatkan hasil akurasi yang lebih baik. Seleksi fitur merupakan tahapan pemilihan fitur yang paling relevan dan mengurangi redudansi data. Sehingga fitur yang digunakan merupakan hasil seleksi yang optimal. Pada penelitian ini seleksi fitur menggunakan *library Minimum Redundancy Maximum Relevance* (MRMR). Penggunaan MRMR dipilih karena fitur seleksi MRMR cenderung meningkatkan akurasi seperti yang terjadi pada penelitian sebelumnya yang juga menggunakan MRMR (Huang *et al.,* 2016).

2.12 *Extreme Gradient Boosting* (XGBoost)

Extreme Gradient Boosting (XGBoost) adalah *framework* yang bertujuan untuk meningkatkan *gradient* yang pertama kali dikenalkan oleh Friedman *et al.,* (2000) yang cenderung lebih cepat dan akurat. Paket XGBoost dapat memecahkan pada model linier dan *algoritma tree learning* Chen dan Guestrin. (2016), Chen *et al.* (2018), serta mendukung berbagai macam fungsi diantaranya regresi, klasifikasi dan perangkingan. XGBoost merupakan metode *machine learning* yang dikembangkan dari pohon keputusan *gradient* yang lebih cepat (Chen & Guestrin, 2016). XGBoost dikembangkan dari klasifikasi dan regresi yang disebut dangan *Classification and Regression Trees* (CART) atau gabungan dari metode klasifikasi dengan regresi. Alasan pemilihan algoritma XGBoost karena fitur yang digunakan diyakini dapat mempercepat perhitungan dan menghindari *overfitting*. *Overfitting* adalah perilaku yang tidak diinginkan dalam pembelajaran mesin yang terjadi ketika model pembelajaran mesin menghasilkan prediksi akurat pada data pelatihan, namun tidak pada data baru. Algoritma XGBoost termasuk dalam kelompok metode pembelajaran ansambel yang menciptakan beberapa model yang lemah dan menggabungkannya untuk membuat model yang lebih kuat. XGBoost menjadi salah satu algoritma yang paling banyak digunakan karena performanya yang konsisten dan kemampuannya untuk membangun model dengan hasil akurasi yang tinggi. XGBoost termasuk metode yang digunakan untuk menyelesaikan

permasalahan *supervised learning* dimana XGBoost terdiri dari data latih (x_i) dapat memprediksi data target *(y_i).* Kinerja XGBoost dapat dilihat pada persamaan berikut:

Fungsi tujuan terdiri dari *training loss* dan *regularization term* Chen dan Guestrin, (2016) dapat dilihat pada Persamaan (4).

$$Obj\ (\theta) = \mathcal{L}(\theta) + \Omega(\theta) \quad (4)$$

Fungsi $\mathcal{L}$ *function* menunjukkan data latih sedangkan Ω adalah parameter yang digunakan (Zhang dan Zhan, 2017). Fungsi untuk mendefinisikan *training* dapat dilihat pada Persamaan (5).

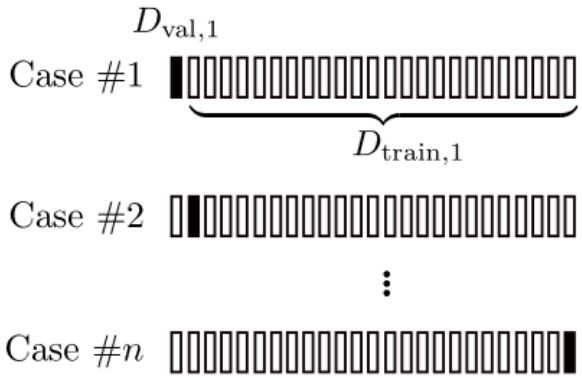
$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \iota(yi, \hat{y}i) \quad (5)$$

2.12 *Cross-Validation*

Cross-validation adalah metode penghitungan ulang data untuk menilai generalisasi model prediksi dan mencegah *overfitting* (Berrar, 2018). *Cross validation* adalah sebagai berikut:

1 *Leave-One-Out Cross Validation*

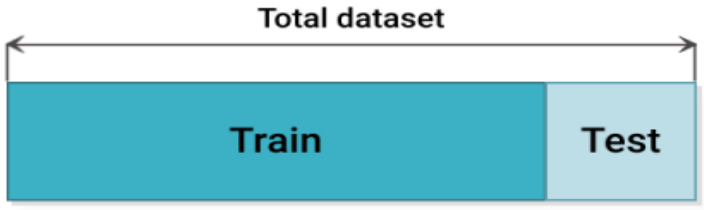
LOOCV adalah bentuk pengujian khusus dari validasi silang. Jumlah *fold* sesuai dengan jumlah data pelatihan. LOOCV sering digunakan untuk menilai secara generalisasi dari pengelompokan statistik (Cawley dan Talbot, 2003).



Gambar 6. Ilustrasi *Leave-One-Out Cross Validation* (Berrar, 2018).

2 *Holdout Cross-Validation*

Holdout Cross-Validation merupakan teknik pengujian yang paling sederhana dan umum digunakan. *Holdout cross-validation* data dibagi menjadi dua dataset terpisah yang disebut dataset pelatihan dan pengujian, biasanya menggunakan 80% dataset untuk pelatihan dan 20% dataset untuk pengujian. Keuntungan dari metode ini adalah rasio dari ketiga kumpulan data ini tidak dibatasi secara ketat. Dalam hal ini, ada kemungkinan perbedaan kategori data yang ditemukan dalam materi pelatihan dan materi ujian tidak terdistribusi secara merata. Untuk mengatasi hal ini, dataset *training* dan *testing* dibuat dengan distribusi kelas data yang berbeda (Šafránková *et al.*, 2010).

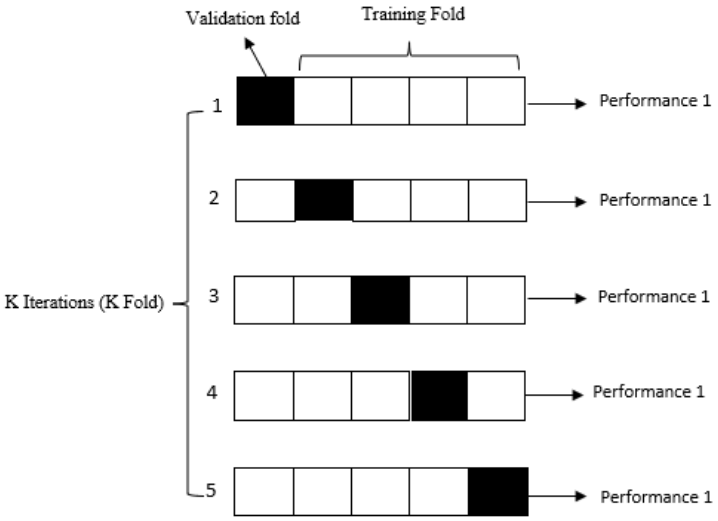


Gambar 7. Ilustrasi *Holdout Cross-Validation* (Vladimir Lyashenko, 2023).

3 *K-Fold Cross-Validation*

Pendekatan yang digunakan dalam validasi silang menggunakan metode *K-Fold Cross Validation* dimana data untuk proses pembuatan model disebut data

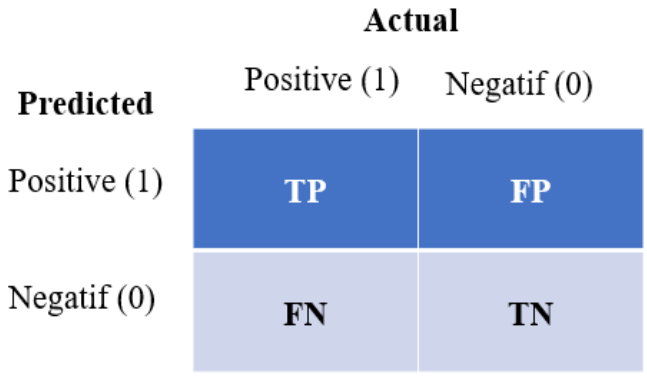
training/latih. Kinerja *K-Fold Cross-Validation* dimana data latih dibagi menjadi K subset yang biasa disebut *fold* yang memiliki ukuran yang sama. Selama proses validasi satu *fold* digunakan sebagai data validasi, sementara sisanya digunakan sebagai data latih. Proses ini diulangi K kali, sehingga setiap *fold* menjadi data validasi satu kali. Pada penelitian ini menggunakan validasi silang sebanyak 5 kali. Berikut simulasi tahap *cross-validation* dilihat pada Gambar 8.



Gambar 8. Simulasi *Cross-Validation* (Berrar, 2018).

2.13 *Confusion Matrix*

Tahap ini adalah tahap melakukan pengukuran kinerja dari klasifikasi dalam *machine learning* (Ohsaki *et al.*, 2017). *Confusion matrix* ini penting dilakukan untuk mengetahui seberapa baik model yang telah dibuat. Pengukuran kinerja menggunakan *confusion matrix* untuk mengetahui kinerja model yang diterapkan dengan kesesuaian data yang disajikan. Terdapat matrik perhitungan pada *confusion matrix* terlihat pada Gambar 9.



Gambar 9. *Confusion Matrix* (Ohsaki et al. 2017).

Keterangan:

1. *TP* (*True Positive*) artinya memprediksi data yang positif dan sistem dapat memprediksi dengan benar
2. *TN* (*True Negative)* artinya memprediksi data yang negatif dan sistem dapat memprediksi dengan benar
3. *FP* (*False Positive*) artinya memprediksi data positif namun sistem memprediksi salah (Data Negatif)
4. *FN* (*False Negative*) artinya memprediksi data negatif namun sistem memprediksi positif

Kemudian matrik perhitungan digunakan untuk menentukan nilai *Accuracy (ACC), Recall* atau *Sensitivity* (SN), *Specificity* (SP) dan *Matthews Correlation Coefficient* (MCC). Berikut penjelasan dari masing-masing perhitungan matrix:

1. *Accuracy (ACC)*

   Akurasi adalah metode pengujian yang digunakan untuk mengukur kedekatan nilai prediksi dengan nilai sebenarnya, untuk menghitung nilai akurasi dapat menggunakan persamaan 6. Akurasi dari pengukuran kedekatan kesepakatan antara kuantitas nilai yang diperoleh dengan pengukuran dan yang sebenarnya (Menditto *et al.*, 2007). Perhitungan nilai akurasi dapat menggunakan Persamaan (6).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (6)$$

Contoh:

$$Accuracy = \frac{23+20}{23+8+20+9}$$

$$= \frac{43}{60} = 0{,}71$$

$$= 0{,}71\text{x } 100\%$$

$$= 71\%$$

2. *Recall* atau *Sensitivity* (SN)

   Metode pengujian yang membandingkan jumlah informasi relevan yang diterima oleh sistem dengan jumlah total informasi relevan yang tersedia atau dapat dikatakan bahwa akurasi menggambarkan seberapa akurat model dapat memprediksi dengan benar. Nilai *recall* atau *sensitivity* adalah proporsi jumlah kasus positif sebenarnya yang diprediksi dengan benar menjadi positif (Powers, 2020). Perhitungan nilai *recall* dapat menggunakan Persamaan (7).

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

Contoh:

$$Recall = \frac{23}{23+9} = 0{,}71$$

$$= 0{,}71 \text{ x } 100\%$$

$$= 71\%$$

3. *Specificity* (SP)

   Specificity adalah pengujian dengan hasil probabilitas kebenaran memprediksi negatif dibandingkan dengan data negatif keseluruhan (Bekkar *et al.*, 2013). Perhitungan nilai *specificity* dapat menggunakan Persamaan (8).

$$Precision = \frac{TN}{TN+FP} \quad (8)$$

Contoh:

$$Precision = \frac{20}{20+23} = 0{,}74$$

$$= 0{,}46 \text{ x } 100\%$$

$$= 46\%$$

4. *Matthews Correlation Coefficient* (MCC)

   Matthews Correlation Coefficient (MCC) pertama kali dikembangkan oleh B.W. Matthews untuk menilai akurasi prediksi struktur protein. Kemudian menjadi ukuran kinerja yang banyak digunakan dalam penelitian biomedis (Boughorbel *et al.,* 2017). MCC diperlukan untuk menghitung perbandingan data *aktual* dengan data prediksi (Chicco dan Jurman, 2020). Perhitungan untuk *Matthews Correlation Coefficient* (MCC) dapat menggunakan Persamaan (9).

$$MCC = \frac{TPxTN-FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FN)(TN+FN)}} \quad (9)$$

Contoh:

$$MCC = \frac{(23x9)-(8x20)}{\sqrt{(23+8)(23+9)(20+8)(20+9)}}$$

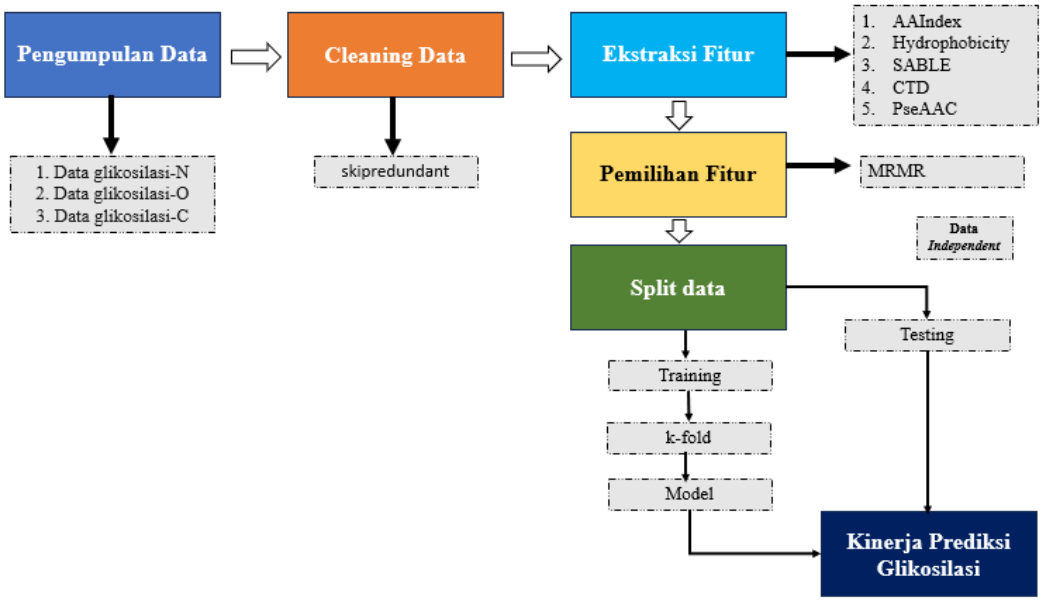
$$= \frac{207-160}{31+32+28+29} = \frac{47}{120} = 0{,}39$$

III. METODOLOGI PENELITIAN

Pelaksanaan penelitian diperlukan tahapan yang sistematis agar penelitian dapat berjalan dengan baik. Pada bab ini menjelaskan bagaimana tahapan dan kerangka penelitian yang dilakukan.

3.1 Tahapan Penelitian

Alur kerja dalam penyelesaian penelitian ini memiliki beberapa tahapan kegiatan yang dilakukan. Tahapan penelitian dapat dilihat pada Gambar 10.



Gambar 10. Tahapan Penelitian Kasifikasi Glikosilasi

Penjelasan tahapan penelitian adalah sebagai berikut:

1. **Pengumpulan Data**

Tahap pengumpulan data yaitu mengidentifikasi data-data yang digunakan berupa data *sequence* yang diperolah secara online pada *database* Uniprot https://www.uniprot.org/. UniProt adalah situs web yang menyediakan urutan protein (Bateman *et al.*, 2015). Tahap pengumpulan data pengambilan data pada

web UniProt dilakukan secara otomatis menggunakan algoritma pemrograman dengan *tools RStudio*. *Sequence* diambil sepanjang 21 *sequence* protein. Panjang *sequence* 10 sebelah kirI dan 10 *sequence* sebelah kanan. Informasi dataset awal yang disajikan terdiri dari ID protein, dan posisi protein terlihat pada Tabel 3.

Tabel 3. Dataset Awal Protein

No.	ID Protein	Posisi Protein
1	Q7Z443	1187
2	Q7Z443	372
3	Q7Z443	897
4	Q7Z443	1152
5	Q7Z443	924
6	P17097	424
7	P17097	18
8	P17097	131
9	P17097	64
10	P17097	645
11	H3BTB2	104
12	H3BTB2	24
13	H3BTB2	68
14	H3BTB2	169
15	H3BTB2	97
16	Q9C0A6	834
17	Q9C0A6	206

Kemudian ID protein dari data tersebut diinputkan pada web UniProt https://www.uniprot.org/ menggunakan pemrograman untuk mendapatkan data urutan *sequence* protein glikosilasi secara otomatis. Berikut potongan kode program untuk mengambil data pada web UniProt dapat dilihat pada Kode Program 1.

```
library(protr)
NegBenchMark_N<-read.csv("D:/DISERTASI S3/GLIKOSILASI/DATA
BERSIH/NegBenchMark_N.txt", header = FALSE, sep="\t")
windows=21;

jumlah_seq=1
seq_iter=array()

sink('Posisi_NegBenchMark_N.fasta')
for(j in 1:nrow(NegBenchMark_N)){
```

```
  prots<-getUniProt(NegBenchMark_N[j,2])
  start=NegBenchMark_N[j,3]-((windows-1)/2)
  end=NegBenchMark_N[j,3]+((windows-1)/2)
  sq=substr(prots[[1]],start,end)

  if(nchar(sq)==windows){

cat(paste('>',NegBenchMark_N[j,2],NegBenchMark_N[j,3],'\n'))
    cat(paste(sq,'\n'))
    seq_iter[jumlah_seq]=sq
    jumlah_seq=jumlah_seq+1
  }
}
sink()
```

Kode Program 1. Kode program untuk mengambil data pada web Uniprot.

Data yang terkumpul sebanyak 3378 yang terdiri dari sequens protein glikosilasi data kelas negatif dan positif. Data kelas negatif sebanyak 2245 sedangkan data kelas positif sebanyak 1133. Pengumpulan data dari *website* dengan mengambil urutan *sequence* protein glikosilasi dengan panjang urutan 21. Data awal yang digunakan pada penelitian ini dapat dilihat pada Tabel 4.

Tabel 4. Data Awal Penelitian

Jenis Glikosilasi	Jumlah Data	Kelas	Jumlah
NegBenchMark_N	667	Negatif	2245
NegBenchMark_C	108	Negatif	
NegBenchMark_O	1018	Negatif	
NegIndependent_N	166	Negatif	
NegIndependent_C	28	Negatif	
NegIndependent_O	258	Negatif	
PosBenchMark_N	333	Positif	1133
PosBenchMark_C	55	Positif	
PosBenchMark_O	520	Positif	
PosIndependent_N	83	Positif	
PosIndependent_C	13	Positif	
PosIndependent_O	129	Positif	
Total	3378		

2. Praproses Data

Tahap praproses data merupakan tahapan dimana dataset yang diambil dari web UniProt kemudian diproses menjadi informasi berupa urutan *sequence* protein

sepanjang 21 *sequence* protein. Data terdiri dari data glikosilasi dengan kelas negatif dan positif. Data *sequence* dapat dilihat pada Tabel 5.

Tabel 5. Data Protein Sequence

ID Protein	Position	Types of Glycosylation	Sequence	Class
Q86U86	900	NegBenchMark_N	ALSYTTKHLHNDVEKERKEKL	Not glycosylated
Q86U86	1262	NegBenchMark_N	NDILLCESRYNESDKQMKKFK	Not glycosylated
Q8WVF1	195	NegBenchMark_N	MFLKDKVQNNNGRFVLPVSGP	Not glycosylated
Q8WVF1	12	NegBenchMark_N	SVRTLPLLFLNLGGEMLYILD	Not glycosylated
F8WDF8	14	NegBenchMark_N	SIPAFLYFLDNLIVFYVLSYL	Not glycosylated
F8WDF8	109	NegBenchMark_N	LLFRTSLKFRNTHLGKKGSEI	Not glycosylated
Q03938	467	NegBenchMark_N	ECGKAFKRSSNLTTHKISHTE	Not glycosylated
Q03938	159	NegBenchMark_N	TYVKVSHIFSNSNRHKIRDTG	Not glycosylated
H0YHD9	97	NegBenchMark_N	PPGGIPQVTVNKSLLAPLNVE	Not glycosylated
H0YHD9	105	NegBenchMark_N	TVNKSLLAPLNVEMDPEIQRV	Not glycosylated
O14672	278	PosBenchMark_N	ISFMVKRIRINTTADEKDPTN	Glycosylated
O14672	439	PosBenchMark_N	NNKFSLCSIRNISQVLEKKRN	Glycosylated
O60449	1103	PosBenchMark_N	SEVKSRQTLQNASETVKYLNN	Glycosylated
O60449	865	PosBenchMark_N	GLKAIKNKIANISGDGQKWWI	Glycosylated
O60449	934	PosBenchMark_N	TKLPFICEKYNVSSLEKYSPD	Glycosylated
O75636	189	PosBenchMark_N	LRVELEDFNGNRTFAHYATFR	Glycosylated
O95857	137	PosBenchMark_N	NCCGFRSVNPNDTCLASCVKS	Glycosylated
O95858	189	PosBenchMark_N	CGVPYTCCIRNTTEVVNTMCG	Glycosylated
O95866	32	PosBenchMark_N	SLDGRPGDRVNLSCGGVSHPI	Glycosylated
P00450	138	PosBenchMark_N	KEHEGAIYPDNTTDFQRADDK	Glycosylated
P00450	358	PosBenchMark_N	LQAFFQVQECNKSSSKDNIRG	Glycosylated
P00450	397	PosBenchMark_N	PSGIDIFTKENLTAPGSDSAV	Glycosylated
Dst….				

Praproses data terdiri dari beberapa tahapan berikut ini:

a. *Cleaning Data*

Urutan *sequence* protein yang telah berhasil dikumpulkan kemudian diproses kembali untuk menghasilkan data optimal dengan melakukan *cleaning data* menggunakan tools *skipredundant* dengan tingkat redundansi 30%. Hal ini dilakukan agar menghindari *overfitting*. *Overfitting* merupakan salah satu proses dalam pembelajaran mesin yang tidak diinginkan yang terjadi ketika model pembelajaran mesin memberikan prediksi akurat untuk data pelatihan tetapi tidak untuk data baru https://www.bioinformatics.nl/cgi-bin/emboss/skipredundant. Berikut pada Tabel 6 dataset optimal yang akan diproses pada tahap selanjutnya.

Tabel 6. Dataset Optimal

Jenis Glikosilasi	Jumlah Data	Kelas	Jumlah
NegBenchMark_N	545	Negatif	1879
NegBenchMark_C	87	Negatif	
NegBenchMark_O	871	Negatif	
NegIndependent_N	136	Negatif	
NegIndependent_C	26	Negatif	
NegIndependent_O	214	Negatif	
PosBenchMark_N	328	Positif	1088
PosBenchMark_C	55	Positif	
PosBenchMark_O	491	Positif	
PosIndependent_N	82	Positif	
PosIndependent_C	12	Positif	
PosIndependent_O	120	Positif	
Total	2967		

Berdasarkan Tabel 6 adalah dataset optimal yang siap untuk diproses ke tahap selanjutnya berjumlah 2967 yang terdiri dari data negatif berjumlah 1879 dan data positif berjumlah 1088. Selanjutnya menghilangkan data redundansi atau data duplikat dan menghindari data *overfitting*. Pengolahan data duplikat menggunakan *tools skipredundant* dengan parameter tingkat redundansi sebesar 30% Berikut dataset optimal yang siap diproses ke tahap ekstraksi fitur dapat lihat pada Tabel 7.

Tabel 7. Dataset Optimal Siap Diproses

Jenis Glikosilasi	Jumlah Data	Kelas	Jumlah
NegBenchMark_N	52	Negatif	286
NegBenchMark_O	53	Negatif	
NegBenchMark_C	87	Negatif	
NegIndependent_N	31	Negatif	
NegIndependent_O	37	Negatif	
NegIndependent_C	26	Negatif	
PosBenchMark_N	51	Positif	260
PosBenchMark_O	76	Positif	
PosBenchMark_C	55	Positif	
PosIndependent_N	28	Positif	
PosIndependent_O	38	Positif	
PosIndependent_C	12	Positif	
Total	546		

Data pada Tabel 7 merupakan data yang akan diproses pada tahap ekstraksi fitur menggunakan AAIndex, *Hydrophobicity*, SABLE, CTD, dan PseAAC.

b. Ekstraksi Fitur

Data optimal yang telah dikumpulkan selanjutnya diproses dengan melakukan ekstraksi fitur. Ekstraksi fitur bertujuan untuk mendapat data atau ciri dari suatu kelas. Ekstraksi fitur bertujuan untuk meningkatkan kinerja akurasi dalam prediksi protein Glikosilasi (Guo *et al.,* 2011). Ada lima ekstraksi fitur yang digunakan yaitu: *Amino Acid Index* (AAIndex), *Hydrophobicity*, *Solvent AccessiBiLitiEs* (SABLE), *Composition, Transition, and Distribution* (CTD), *Pseudo Amino Acid Composition* (PseAAC).

Tahap ekstraksi fitur berfungsi untuk mengidentifikasi dari masing-masing setiap fitur. Setiap fitur memiliki dimensi berbeda-beda pada setiap ekstraksi fitur. Eksperimen dari beberapa ekstraksi fitur akan berkontribusi dalam peningkatan akurasi prediksi glikosilasi. Berikut kontribusi dari setiap ekstraksi fitur pada Tabel 8.

Tabel 8. Kontribusi Ekstraksi Fitur

Deskripsi	Dimensi	Prosentase
AAIndex	21	19%
Hindrophobicity	21	19%
SABLE	21	19%
CTD	21	19%
PseAAC	24	24%
Total	109	

c. Seleksi Fitur

Prediksi glikosilasi sering kali melibatkan analisis data molekuler untuk memahami dan memprediksi glikosilasi pada protein atau molekul lainnya. Pemilihan fitur dapat membantu meningkatkan kinerja model prediktif dengan memilih fitur-fitur yang paling informatif atau relevan dan mengurangi kompleksitas model dan risiko *overfitting* di mana model pembelajaran mesin terlalu baik menyesuaikan diri

dengan data pelatihan tetapi gagal generalisasi dengan baik ke data baru. Pemilihan fitur adalah proses pemilihan subset dari fitur yang relevan dari kumpulan data. Dalam prediksi glikosilasi fitur-fitur dapat mencakup berbagai karakteristik atau atribut yang terkait dengan proses glikosilasi, seperti sifat-sifat kimia atau struktur biologis dari molekul-molekul yang terlibat.

Pemilihan fitur sangat penting dilakukan dalam membangun klasifikasi yang lebih baik agar data yang dihasilkan dapat digunakan (Khaire dan Dhanalakshmi 2022). Ada beberapa alasan mengapa tahapan seleksi fitur diperlukan dalam *machine learning* diantaranya sebagai berikut:

1. Meningkatkan Kinerja Model

   Dengan menghilangkan fitur yang kurang relevan atau tidak informatif maka dapat meningkatkan kinerja model. Fitur-fitur yang tidak memberikan kontribusi signifikan terhadap prediksi dapat menyebabkan kompleksitas model yang tidak perlu.
2. Mengurangi *Overfitting*

   Model *machine learning* cenderung *overfitting* ketika memiliki terlalu banyak fitur dibandingkan dengan jumlah sampel pelatihan. Pemilihan fitur dapat membantu mengurangi *overfitting* dengan meminimalkan kompleksitas model.
3. Mempercepat Pelatihan

   Dengan mengurangi jumlah fitur maka waktu yang diperlukan untuk melatih model dapat berkurang terutama bekerja dengan dataset yang besar.
4. Mengurangi Dimensi Data

   Dengan mengurangi jumlah fitur dapat membantu mengatasi masalah dimensi tinggi sehingga dapat meningkatkan efisiensi.

Pemilihan fitur dilakukan untuk mengurangi jumlah fitur dan menghilangkan redundansi untuk meningkat peningkatan performa prediksi serta memperoleh hasil yang komprehensif. Seleksi fitur yang digunakan yaitu *Minimum Redundancy Maximum Relevance* (MRMR) (Ding dan Peng, 2003). Seleksi fitur MRMR

merupakan tahapan pemilihan fitur dengan target korelasi tertinggi dengan kelas atau output dari prediksi dan korelasi redundansi yang terendah (De Jay *et al.*, 2013).

MRMR memilih fitur-fitur yang memiliki hubungan langsung dengan fitur target. MRMR membantu mengatasi masalah dimensi tinggi pada dataset dengan memilih fitur yang lebih informatif dan relevan. Dengan mengurangi jumlah fitur yang digunakan, MRMR dapat meningkatkan efisiensi komputasi dan mengurangi waktu yang dibutuhkan dalam pemodelan. MRMR adalah metode seleksi fitur yang bertujuan untuk memilih fitur yang memiliki korelasi tinggi terhadap suatu kelas (*output*) dan korelasi rendah terhadap data yang sama (Radovic *et al.*, 2017).

Proses seleksi MRMR menggunakan *library* MRMR. Memilih fitur MRMR menggunakan fungsi *MRMR.classic().* Pemilihan fungsi MRMR terdiri dari 25 fitur, 50 fitur, dan 75 fitur.

3. Pembagian Data

Pembagian data adalah proses membagi dataset menjadi data dengan tujuan untuk melatih, memvalidasi, dan menguji model *machine learning*. Pembagian ini dilakukan untuk memastikan bahwa model yang dikembangkan mampu menggeneralisasi dengan baik ke data yang belum pernah dilihat sebelumnya. Berikut yang digunakan dalam pembagian data:

1. Data Pelatihan (*Training Data*):

   Merupakan data yang digunakan selama proses pelatihan model. Sebagian besar data digunakan untuk pelatihan. Data Traning pada penelitian ini 80 % dari total data.

2. Data Validasi (*Validation Data*)

   Merupakan data yang digunakan untuk mengevaluasi kinerja model selama pelatihan. Evalusai model menggunakan *K-Fold Cross-Validation*. Evaluasi model pada data validasi membantu menghindari *overfitting* dan meningkatkan kemampuan model untuk menggeneralisasi. *K-Fold Cross-*

Validation merupakan sebuah teknik yang digunakan untuk mengevaluasi kinerja model pada data training tersebut. Data training digunakan dalam setiap iterasi *K-Fold Cross-Validation* untuk melatih model dan mengukur kinerjanya. Pada penelitian ini menggunakan *Cross-Validation* sebanyak 5 kali.

3. Data Uji (*Testing Data*)

   Merupakan data yang tidak pernah terlibat selama pelatihan atau validasi. Data testing digunakan untuk menguji kinerja model setelah proses pelatihan selesai. Data testing pada penelitian ini sebesar 20%. Tahap pembagian data penting dilakukan untuk mengevaluasi kinerja model pada data yang tidak pernah dilihat selama proses pelatihan atau validasi. Hal ini memberikan gambaran yang lebih objektif tentang kemampuan model dalam menggeneralisasi ke data baru.

4. Pemodelan dan Evaluasi

Tahap ini melakukan pemodelan dengan menggunakan pendekatan *Extreme Gradient Boosting* (XGBoost) (Chen dan Guestrin, 2016). Pemodelan menghasilkan nilai dari masing masing data *benchmark* dan *independent*. Kemudian model di Uji dengan menggunakan *Cross-Validation 5-fold* menghasilkan nilai *accuracy*, *sensitivity*, *specificity*, dan MCC.

3.2 Alat

Adapun alat yang digunakan pada penelitian ini adalah sebagai berikut:

1. *Software*

Software atau perangkat lunak merupakan alat yang dibutuhkan dalam melakukan penelitian. *Software* yang digunakan adalah sebagai berikut:

a. Sistem Operasi: Windows 10 pro 64 bit
b. RStudio versi 3.6.1 dan RStudio versi 3.6.4
c. *Package* caret versi 6.0-84,
d. *Package* BioSeqClass versi 1.40.0
e. *Package* protr versi 1.6-2

f. *Package* MRMRe 2.1.0
g. *Package* Xgboost versi 4.6-14.
h. *Python* menggunakan *Google Collab*
i. *Skipredundant*
j. *Web UniProt*
k. *Microsoft Excel* 2016
l. *Microsoft Word* 2016

2. ***Hardware***

Penelitian ini membutuhkan perangkat keras atau *hardware* untuk mendukung dalam penyelesaian penelitian ini. Adapun spesifikasi *hardware* yang digunakan adalah sebagai berikut:

a. *Laptop processors*: Intel® *Core*(TM) i7
b. 5600U CPU @ 2.60GHz.
c. Memory *RAM* 8 GB
d. SK *hynic* SC210 mSTA 128GB

IV. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah diuraikan, maka dapat disimpulkan sebagai berikut:

1. Ekstraksi fitur yang diterapkan dalam peningkatan prediksi glikosilasi terdapat lima jenis yaitu AAIndex, *Hydrophobicity*, CTD, SABLE, dan PseAAC. Masing-masing fitur memiliki kontribusi untuk peningkatan prediksi glikosilasi. Fitur AAIndex paling berkontribusi besar pada peningkatan prediksi glikosilasi-N yaitu sebesar 24%. Sedangkan fitur SABLE berkontribusi besar pada peningkatan prediksi glikosilasi-O yaitu sebesar 44%. Sedangkan fitur *Hydrophobicity* berkontribusi besar yaitu 27%, dan fitur PseAAC berkontribusi 27% pada peningkatan akurasi prediksi glikosilasi-C.
2. Seleksi fitur *Minimum Redundancy Maximum Relevance* (MRMR) berhasil berkontribusi dalam peningkatan prediksi glikosilasi. Fitur terpilih menggunakan teknik MRMR yaitu 25 fitur, 50 fitur, dan 75 fitur.
3. Performa predikasi *Post-translational modification* glikosilasi N, glikosilasi-O, dan glikosilasi-C mencapai kinerja tinggi yaitu masing-masing nilai akurasi sebesar 100%. Peningkatan akurasi dipengaruhi jumlah data yang sedikit setelah melalui tahapan pembersihan data.
4. Perbandingan kinerja prediksi glikosilasi-N, glikosilasi-O, dan glikosilasi-C dalam penelitian ini menunjukkan peningkatan dari penelitian sebelumnya. Peningkatan tersebut mencapai 5% dibandingkan dengan penelitian yang dilakukan oleh *Li et al*. pada tahun 2025, 5,4% dibandingkan dengan penelitian yang dilakukan oleh Chien *et al*. pada tahun 2020, dan 4,89% dibandingkan dengan penelitian yang dilakukan oleh Akuhlani *et al*. pada tahun 2022.

5.2. Saran

Penelitian ini telah menghasilkan kinerja yang baik namun dapat kembangkan lebih lanjut dengan mempertimbangan hal-hal sebagai berikut:

1. Dataset yang digunakan dapat mengambil panjang 51 urutan asam amino, 20 ke kanan dan 20 ke kiri sehingga diperoleh informasi yang lebih banyak lagi.
2. Penelitian lanjutan dapat melakukan prediksi jenis PTM yang lain dengan menggunakan algoritma seperti *adaboost algorithm.*
3. Penelitian tentang glikosilasi penting dilakukan lebih lanjut untuk memperluas pemahaman tentang klinis yang berkaitan dengan penyakit dan pengembangan obat.

DAFTAR PUSTAKA

Acar T, Arayici PP, Karahan M, Akdeste Z. 2015. Applications of Molecular Genetics in Personalized Medicine Post-Translational Modificatıons of Proteins. USA, 19 hlm. https://www.researchgate.net/publication/326835654

Akram M, Asif HM, Uzair M, Akhtar N, Madni A, Ali Shah SM, Hasan ZU, Ullah A. 2011. Amino Acids: A Review Article. *Journal of Medicinal Plants Research*. 5(17):3997-400

Alkuhlani A, Gad W, Roushdy M. 2023. International Journal of Intelligent Prediction Of O-Glycosylation Site Using Pre-Trained. 23(1):41–52.doi:10.21608/ijicis.2023.160986.1218.

Alkuhlani A, Gad W, Roushdy M, Salem ABM. 2022. PUStackNGly: Positive-Unlabeled and Stacking Learning for N-Linked Glycosylation Site Prediction. *IEEE Access*. 10:12702–12713.doi:10.1109/ACCESS.2022.3146395.

Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, Antunes R, Arganiska J, Bely B, Bingley M, *et al.* 2015. UniProt: A Hub for Protein Information. Nucleic AcidsRes. 43 (D1): D204–D212.doi:10.1093/nar/gku989.

Bekkar M, Djemaa HK, Alitouche TA. 2013. Evaluation Measures for Models Assessment Over Imbalanced Data Sets. *J. Inf. Eng. Appl*. 3(10):27–38.

Berrar D. 2018. Cross-validation. *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.* 1–3:542–545.doi:10.1016/B978-0-12-809633-8.20349-X.

Boughorbel S, Jarray F, El-Anbari M. 2017. Optimal Classifier For Imbalanced Data Using Matthews Correlation Coefficient Metric. *PLoS One*. 12(6):1–17.doi:10.1371/journal.pone.0177678.

Brostrom MA, Brostrom CO. 2007. Protein Synthesis. *Encyclopedia of Stress*. Elsevier Inc. 258-265.doi: 10.1016/B978-012373947-6.00315-9

Reddy YCAP, Viswanath P, Eswara Reddy B. 2018. Semi-supervised Learning: a Brief Review. *Int. J. Eng. Technol.* 7(1.8):81.doi:10.14419/ijet.v7i1.8.9977.

Caragea C, Sinapov J, Silvescu A, Dobbs D, Honavar V. 2007. Glycosylation site prediction using ensembles of Support Vector Machine Classifiers. *BMC Bioinformatics*. 8:438.doi:10.1186/1471-2105-8-438.

Cawley GC, Talbot NLC. 2003. Efficient Leave-One-Out Cross-Validation Of Kernel Fisher Discriminant Classifiers. *Pattern Recognit.* 36(11):2585–2592.doi:10.1016/S0031-3203(03)00136-5.

Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 13-17-Augu:785–794.doi:10.1145/2939672.2939785.

Chen T, He T. 2014. Xgboost: Extreme Gradient Boosting. *R Lect.*(2016):1–84.

Chen T, He T, Benesty M, Khotilovich V, Tang Y. 2018. Xgboost: Customized Extreme Gradient Boosting. :1–4.

Chicco D, Jurman G. 2020. The Advantages Of The Matthews Correlation Coefficient (Mcc) Over F1 Score And Accuracy In Binary Classification Evaluation. *BMC Genomics*. 21(1):1–13.doi:10.1186/s12864-019-6413-7.

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.doi:10.1109/access.2020.3022629.

Ding C, Peng H. 2003. Minimum Redundancy Feature Selection From Microarray Gene Expression Data. *Proc. 2003 IEEE Bioinforma. Conf. CSB 2003*. 3(2):523–528.doi:10.1109/CSB.2003.1227396.

Kumar, K 2018. (PDF) Data Analysis using R and Python. *ResearchGate*.(June), https://www.researchgate.net/publication/342231305

Dridi S, Supervised Learning - A Systematic Literature Review, 2015. *A Comp. Anal. Linear Regres. Support Vector Regres.*, pp. 1–8.

El-Sayed DMK. 2020. Amino acids and protein chemistry part 1 https://www.researchgate.net/publication/345985616,

Everest-Dass A V., Moh ESX, Ashwood C, Shathili AMM, Packer NH. 2018. Human Disease Glycomics: Technology Advances Enabling Protein Glycosylation Analysis–Part 2. *Expert Rev. Proteomics*. 15(4).

doi:10.1080/14789450.2018.1448710.Fig_3_PTM_transport.

Friedman J, Tibshirani R, Hastie T. 2000. Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by The Authors). *Ann. Stat.* 28(2):337–407.doi:10.1214/aos/1016120463.

Noriko F, Takata T, Norihiko F , Aki K, Sakaue H. 2018. D-Amino Acids In Protein: The Mirror Of Life As A Molecular Index of Aging. *Biochim. Biophys. Acta-Proteins Proteomics*. 1866 (7):840–847. doi:10.1016/j.bbapap.2018.03.001.

Govindan G, Nair AS. 2011. Composition, Transition and Distribution (CTD) - A Dynamic Feature for Predictions Based on Hierarchical Structure of Cellular Sorting. *Proc. - 2011 Annu. IEEE India Conf. Eng. Sustain. Solut. INDICON-2011*. 38(8):10425-10436.doi:10.1109/INDCON.2011.6139332.

Guo L, Rivero D, Dorado J, Munteanu CR, Pazos A. 2011. Automatic Feature Extraction Using Genetic Programming: An Application To Epileptic Eeg Classification. *Expert Syst. Appl.* 38(8):10425–10436. doi:10.1016/j.eswa.2011.02.118.

Guruprasad L. 2019. Protein Structure. *Resonance*. doi:10.1007/s12045-019-0783-7.

Hamidah I, Subkhi N, Ratnasari A. 2020. Validasi Media Pembelajaran Alat Peraga Sintesis Protein Berbahan Baku Limbah Plastik. *Rep. Biol. Educ. - J. UMMI*. 1(2):42–51.

Huang YA, You ZH, Chen X, Chan K, Luo X. 2016. Sequence-Based Prediction Of Proteinprotein Interactions Using Weighted Sparse Representation Model Combined With Global Encoding. *BMC Bioinformatics*. 17(1):1–11. doi:10.1186/s12859-016-1035-4.

Jay ND, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B. 2013. MRMRe: An R Package for Parallelized MRMR Ensemble Feature Selection. *Bioinformatics*. 29(18):2365–2368. doi:10.1093/bioinformatics/btt383.

Jo I, Lee S, Oh S. 2019. Improved Measures of Redundancy and Relevance MRMR Feature Selection. *Computers*. 8(2):1–14. doi:10.3390/computers8020042.

Kadakeri S, Arul MR, Bordett R, Duraisamy N, Naik H, Rudraiah S. 2020. *Protein synthesis and characterization*. Elsevier Ltd.

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAIndex: Amino Acid Index Database, Progress Report 2008. *Nucleic Acids Res.* 36(SUPPL. 1):202–205. doi:10.1093/nar/gkm998.

Khaire UM, Dhanalakshmi R. 2022. Stability of Feature Selection Algorithm: A review. *J. King Saud Univ. - Comput. Inf. Sci.* 34(4):1060–1073.doi:10.1016/j.jksuci.2019.06.012.

Kimball SR, Jefferson LS. 2006. New Functions for Amino Acids: Effects on Gene Transcription and Translation. *American Journal of Clinical Nutrition*. 3(2).doi: 10.1093/ajcn/83.2.500s

Kyte J, Doolittle RF. 1982. A Simple Method for DIsplaying The Hydropathic Character of a Protein. *J. Mol. Biol.* 157(1):105–132.doi:10.1016/0022-2836(82)90515-0.

Li F, Li C, Wang M, Webb GI, Zhang Y, Whisstock JC, Song J. 2015. GlycoMine: A Machine Learning-based Approach for Predicting N-, C-and O-linked Glycosylation in The Human Proteome. *Bioinformatics*. 31(9):1411–1419. doi:10.1093/bioinformatics/btu852.

Ma B, Meng F, Yan G, Yan H, Chai B, Song F. 2020. Diagnostic Classification Of Cancers Using Extreme Gradient Boosting Algorithm And Multi-Omics Data. *Comput. Biol. Med.* doi:10.1016/j.compbiomed.2020.103761.

Macek B, Forchhammer K, Hardouin J, Weber-Ban E, Grangeasse C, Mijakovic I. 2019. Protein Post-Translational Modifications In Bacteria. *Nat. Rev. Microbiol.* 17(11):651–664. doi:10.1038/s41579-019-0243-0.

Mann M, Jensen ON. 2003. Proteomic Analysis of Post-Translational modification. 21(March):255–261.doi: 10.1038/nbt0303-255

Mazola Y, Chinea G, Musacchio A. 2011. Integrating bioinformatics tools to handle glycosylation. *PLoS Comput. Biol.* 7(12):1–8. doi:10.1371/journal.pcbi.1002285.

Menditto A, Patriarca M, Magnusson B. 2007. Understanding The Meaning of Accuracy, Trueness And Precision. *Accredit. Qual. Assur.* 12(1):45–47.

doi:10.1007/s00769-006-0191-z.

Minguez P, Letunic I, Parca L, Bork P. 2013. PTMcode: A Database Of Known And Predicted Functional Associations Between Post-Translational Modifications In Proteins. *Nucleic Acids Research* 41(D1)doi:10.1093/nar/gks1230.

Misran A, HaniffJaafar A. 2018. Protein. *Postharvest Physiology and Biochemistry of Fruits and Vegetables*. doi.org/10.1016/B978-0-12-813278-4.00015-4

Ohsaki M, Wang P, Matsuda K, Katagiri S, Watanabe H, Ralescu A. 2017. Confusion-Matrix-Based Kernel Logistic Regression for Imbalanced Data Classification. *IEEE Trans. Knowl. Data Eng.* 29(9):1806–1819.doi:10.1109/TKDE.2017.2682249.

Ohtsubo K, Marth JD. 2006. Glycosylation in Cellular Mechanisms of Health and Disease. *Cell*. doi:10.1016/j.cell.2006.08.019.

Pitti T, Chen CT, Lin HN, Choong WK, Hsu WL, Sung TY. 2019. N-GlyDE: A Two-Stage N-Linked Glycosylation Site Prediction Incorporating Gapped Dipeptides And Pattern-Based Encoding. *Sci. Rep.* doi:10.1038/s41598-019-52341-z.

Powers DMW. 2020. Evaluation: From Precision, Recall And F-Measure To Roc, Informedness, Markedness And Correlation. (May). doi:10.9735/2229-3981.

Prabakaran S, Lippens G, Steen H, Gunawardena J. 2012. Post-translational modification: Nature's Escape From Genetic Imprisonment And The Basis For Dynamic Information Encoding. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 4(6):565–583. doi:10.1002/wsbm.1185.

Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. 2016. iPTM-mLys: Identifying Multiple Lysine PTM Sites and Their Different Types. *Bioinformatics*. 32(20):3116-3123.doi:10.1093/bioinformatics/btw380.

Radovic M, Ghalwash M, Filipovic N, Obradovic Z. 2017. Minimum Redundancy Maximum Relevance Feature Selection Approach For Temporal Gene Expression Data. *BMC Bioinformatics*. 18(1):1–14.doi:10.1186/s12859-016-1423-9.

Regan P, McClean PL, Smyth T, Doherty M. 2019. Early Stage Glycosylation

Biomarkers in Alzheimer's Disease. *Medicines*. 6(3):92. doi:10.3390/medicines6030092.

Šafránková J, Annual Conference of Doctoral Students (19 2010.06.01-04 Prague), WDS'10 (19 2010.06.01-04 Prague), Week of Doctoral Students 2010 (19 2010.06.01-04 Prague). 2010. 19th Annual Conference of Doctoral Students, WDS'10 "Week of Doctoral Students 2010", Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic, June 1, 2010 to June 4, 2010 : [proceedings of contributed papers]. Pt. 1 Mathematics and. 1999 (December):31–36.

Sumardjo D. 2006. Pengantar Kimia: Buku Panduan Kuliah Mahasiswa Kedokteran. *Positron*.

Taherzadeh G, Dehzangi A, Golchin M, Zhou Y, Campbell MP. 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. doi:10.1093/bioinformatics/btz215.

Tanaka M, Okutomi M. 2014. A Novel Inference of a Restricted Boltzmann Machine. *Proceedings - International Conference on Pattern Recognition*. IEEE Access. doi:10.1109/ICPR.2014.271. https://ieeexplore.ieee.org/document/6976981

Vieira S, Pinaya LWH, Mechelli A. 2019. Introduction to Machine Learning. *Machine Learning: Methods and Applications to Brain Disorders*. Elsevier. 1-20.doi: 10.1016/B978-0-12-815739-8.00001-8

Vigerust DJ. 2011. Protein Glycosylation In Infectious Disease Pathobiology And Treatment. *Cent. Eur. J. Biol.* 6(5):802–816. doi:10.2478/s11535-011-0050-8.

Lyashenko AV. 2023. No Title Cross-Validation in Machine Learning: How to Do It Right. [diunduh 2023 Jun 19]. Tersedia pada: https://neptune.ai/blog/cross-validation-in-machine-learning-how-to-do-it-right

Wagner M, Adamczak R, Porollo A, Meller J. 2005. Linear Regression Models For Solvent Accessibility Prediction In Proteins. *J. Comput. Biol.* 12(3):355–369. doi:10.1089/cmb.2005.12.355.

Weerapana E, Imperiali B. 2006. Asparagine-linked Protein Glycosylation: From Eukaryotic to Prokaryotic Systems. *Glycobiology*. PMID. 16510493**.**doi:10.1093/glycob/cwj099.

Yang Q. 2020. Encyclopedia of Big Data. *Encycl.BigData*.(October).doi:10.1007/978-3-319-32001-4.

Yang X, Han H. 2017. Factors analysis of protein O-glycosylation site prediction. *Comput. Biol. Chem.* 71:258263.doi:10.1016/j.compbiolchem.2017.09.005.

Zare M, Mohabatkar H, Faramarzi FK, Beigi MM, Behbahani M. 2015. Using Chou's Pseudo Amino Acid Composition and Machine Learning Method to Predict the Antiviral Peptides. *Open Bioinforma. J.* 9(1):13–19. doi:10.2174/1875036201509010013.

Zhang L, Luo S, Zhang B. 2016. Glycan Analysis of Therapeutic Glycoproteins. *MAbs*. 8(2):205–215. doi:10.1080/19420862.2015.1117719.

Zhang L, Zhan C. 2017. Machine Learning in Rock Facies Classification: An Application of XGBoost. :1371–1374. doi:10.1190/igc2017-351.

Zhang Y, Sun L. 2020. Sweetening the Deal: Glycosylation and its Clinical Applications. *J. Biomed. Sci.* 9(3):1–7. doi:10.36648/2254-609x.9.3.9.

LAMPIRAN

Hasil Penelitian Klasifikasi *Post Translational Modification* (PTM) Glikosilasi-O

Lampiran Bukti Submit pada jurnal ***Journal of Computer Science***