

Performance evaluation of feature extraction to improve the classification of PTM in C-glycosylation using XGBoost

Damayanti¹, Favorisen Rosyking Lumbanraja², Akmal Junaidi², Sutyarso³, Gregorius Nugroho Susanto³, Nirwana Hendrastuty¹

¹Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Bandar Lampung, Indonesia

²Department of Computer Science, Faculty of Mathematics and Natural Science, University of Lampung, Bandar Lampung, Indonesia

³Department of Biology, Faculty of Mathematics and Natural Sciences, University of Lampung, Bandar Lampung, Indonesia

Article Info

Article history:

Received Mar 15, 2024

Revised Oct 29, 2024

Accepted Nov 19, 2024

Keywords:

Feature selection

Glycosylation

Machine learning

Post-translational modification

Prediction

Protein

Sequence

ABSTRACT

Protein function is regulated by an important mechanism known as post-translational modification (PTM). Covalent and enzymatic protein modifications are added during protein biosynthesis, and such alterations significantly influence the regulation of gene activity and the functionality of proteins. Glycosylation, one type of PTM, involves adding sugar groups to a protein's structure. Numerous illnesses, such as diabetes, cancer, and the flu, have been linked to glycosylation. Therefore, it is critical to predict the presence of glycosylation, whether it occurs or not. Currently, predicting glycosylation sites is still done manually using biological methods, which require repeated experiments and a significant amount of time. To address these challenges, it is essential to rapidly develop computational data models using machine learning methods. In this study, the extreme gradient boosting (XGBoost) method is implemented, and C-glycosylation data is obtained from the publicly accessible UniProt website. The objective is to enhance the accuracy of C-glycosylation prediction using the XGBoost method. Feature extraction is performed using amino acid index (AAindex), composition, transition, and distribution (CTD), solvent AccessiBiLitiEs (SABLE), hydrophobicity, and pseudo amino acid composition (PseAAC) to improve accuracy. The minimum redundancy maximum relevance (MRMR) method is applied for feature selection. The findings of the study demonstrate that the PTM C-glycosylation prediction achieved 100%.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Damayanti

Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia

Bandar Lampung, Lampung, Indonesia

Email: damayanti@teknokrat.ac.id

1. INTRODUCTION

Post-translational modification (PTM) is a vital process that impacts the regulation of protein activity [1]. During protein biosynthesis, PTM involves covalent and enzymatic alterations, which are fundamental for the regulation of gene expression and the adjustment of protein functions. Examples of PTMs include phosphoryl [2], [3]. Glycosylation is one of the PTMs that occur in eukaryotic cells and is characterized by the addition of carbohydrate moiety to proteins [4]. This modification affects a number of biological processes, including protein folding, intercellular communication, protein metabolism, and immune responses [5].

Glycosylation is one of the post-translational protein modifications in eukaryotic cells that influence various biological processes, including protein folding, cell-cell interactions, and immune responses [5], [6].

Protein glycosylation is a major PTM event that is poorly understood in eukaryotic cells, which contributes to diverse functions that range from protein folding and communication between cells, to immune regulation [7]. Glycosylation is the covalent addition of carbohydrates to proteins post-translationally carbohydrates are composed of glycans, sugars or saccharides with complex linear or branched structures arranged by covalently bound monosaccharide molecules [8]. Although this is a general categorization, there are 4 primary types of glycosylation: N-glycosylation (where the sugar connects through an amide nitrogen), O-glycosylation (the linking oxygen), C-glycosylation, and glycosylphosphatidylinositol (GPI) anchor [5], [9], [10]. N-linked glycosylation is a sugar bound to an asparagine residue, while O-glycosylation is a sugar bound to a serine residue [11]. C-glycosylation bound to tryptophan residues. glycosylation contributes to the prevention of various diseases, such as bone, nerve, and other diseases [12].

These observations state that there are changes in sub-structural glycans from various diseases, including Alzheimer's [13], cancer [14], neurological diseases in the context of glycol, diabetes glycan features, and antibody glycan structural features. Glycosylation changes observed in various diseases are important in understanding disease progression and advances in treatment [12], [15]. The problem of predicting glycosylation sites is still being conducted manually by implementing biological methods. These methods still require experiments with repeated performance, thus taking a considerable amount of time. This is important to address those challenges such that data-driven models are needed by integrating machine learning methods in order to predict the glycosylation sites efficiently [16], [17]. This research focuses on discussing in C-glycosylation. Machine learning is an approach in artificial intelligence (AI) that is widely used to imitate human behavior to solve problems automatically [18]-[22]. With the use of machine learning computing, it is hoped that it can increase the accuracy of glycosylation predictions. The machine learning algorithm used is extreme gradient boosting (XGBoost). What we aim to achieve in this research is an increase in the accuracy of glycosylation predictions when compared to previous research.

Improvement to analyze C-glycosylation data and conducts several feature extraction experiments and feature selection to obtain optimal data ready for processing by machine learning. Several related studies previously developed include research on PTM using sequence data obtained from the UniProt website with a sequence length of 15 sequences. Feature extraction used amino acid index (AAindex), physicochemical properties of proteins position-specific, scoring matrices (PSSMs), and residue conservation score. Feature selection uses minimum redundancy maximum relevance (MRMR), then modeling and evaluation use the random forest algorithm. The results of this research show an accuracy rate of 95% [22]. Then, previously, we discussed research on glycosylation prediction using deep learning. The length of the sequence developed is 21. The results of this research show an accuracy rate of 83.20% [23]. Then research discussing C-glycosylation uses a sequence length of 31. Feature extraction uses a support vector machine (SVM). Modeling using XGBoost. The research results show that the accuracy value of C-glycosylation prediction is 77.68% [24]. The research then discusses glycosylation prediction using sequence datasets with a length of 21. They applied feature extraction techniques including binary, AAindex, amino acid composition (AAC), parallel correlation pseudo amino acid composition (PC-PseAAC), series correlation pseudo amino acid composition (SC-PseAAC), motif, relative surface accessibility/absolute surface accessibility (RSA/ASA), secondary structure (SS), and signal. Their research findings revealed an accuracy rate of 94.68% [25].

Despite numerous studies conducted, there is still potential to improve the accuracy of predicting post-translational glycosylation modifications in C-glycosylation. We propose feature extraction using AAindex, hydrophobicity, solvent AccessiBiLitiEs (SABLE), composition, transition, and distribution (CTD), and PseAAC. The extraction of hydrophobicity and SABLE features represents a novel emphasis that has not been utilized in previous glycosylation prediction research. The extraction of hydrophobicity and SABLE features is a novelty of this study. The MRMR feature selection technique and XGBoost modeling approach are also employed in this study. Subsequently, the predictive performance of C-glycosylation will be compared with previously developed research [26]. This research plays important role for drug development in the area of clinical. Most of the proteins in human and other mammalian are undergoing the process of glycosylation; nonetheless, any deviations lead to numerous diseases such as cancer, Alzheimer's disease as well as many infection pathobiological diseases among others [27].

2. METHOD

This section describes the steps taken in conducting research. Starting from the stage of collecting data, preprocessing data, feature extraction (using AAindex, hydrophobicity, SABLE, CTD, and PseAAC), feature selection (using MRMR), and classification (using XGBoost), performance measurement. The research stages can be seen in Figure 1.

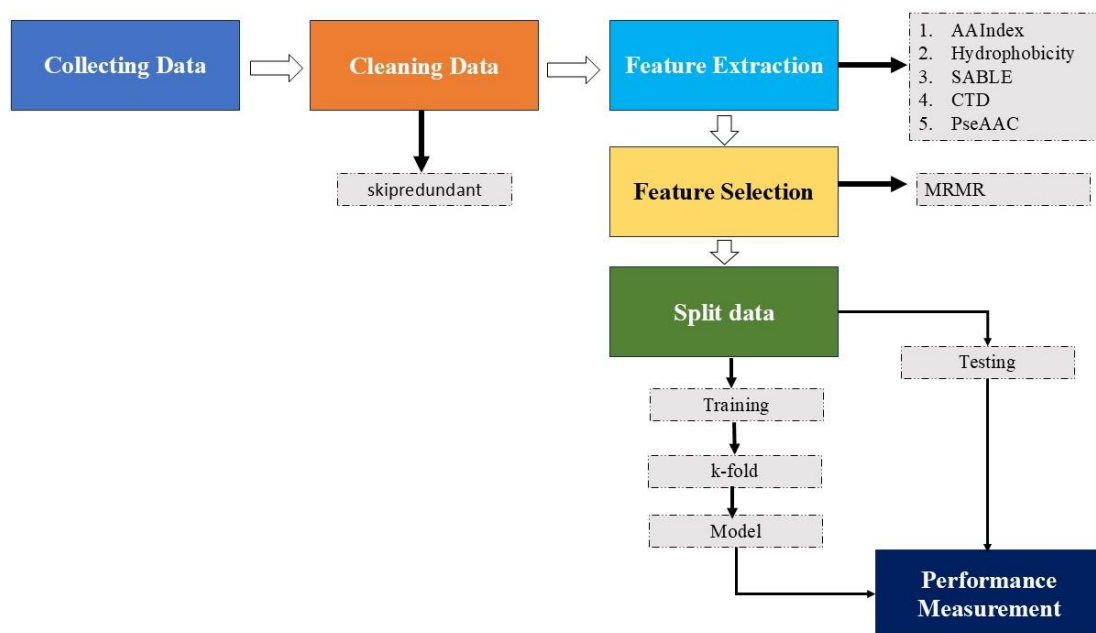


Figure 1. Research stage classification of PTM in C-glycosylation

2.1. Collecting data

The data analysis stage is identifying the data used in the form of sequence data, which is processed online in the Uniprot database (<https://www.uniprot.org/>) [28]. This stage is data preprocessing. Sequence data is taken with a length of 21 residues: 10 residues from the right and 10 residues from the left [25]. The C-glycosylation is in the form of text data in the form of a sequence [28], [29]. The following is an example of an amino acid sequence from a C-glycosylation site where positive data appears, namely "C S P S S C L M T E W G E W D E C S A T C". Data is collected both from benchmark data and independent. Each dataset includes both negative and positive. This data is a class or target that shows glycosylated or non-glycosylated data.

2.2. Cleaning data

The protein sequence order, which has been successfully collected, is then reprocessed to generate optimal data by performing data cleaning using the skip redundant tools. The purpose of cleaning data is to ensure that the data used for further analysis or modeling is of high quality and reliable. This includes various actions such as removing missing, handling outlier values, changing inappropriate data formats, and so on. cleaning data is important because poor data quality can lead to inaccurate.

2.3. Feature extraction

Feature extraction aims to obtain data or characteristics from a class [30]. Feature extraction aims to improve accuracy and performance in protein glycosylation prediction [31]. This stage converts text data into numerical data so that it can be processed by the learning machine. There were five types of feature extraction, namely:

2.3.1. Amino acid index

The AAindex based feature extraction package and the Aaindex() function incorporated into BioSeqClass. The feature contains 21 amino acid sequences. Figure 2 Aacode snippet for an Aaindex feature extraction.

```

if (interactive()) {
  file = file.path(path.package("BioSeqClass"),
"example", "New_PosIndependent_C.pep")
  seq = as.matrix(read.csv(file, header = F)) [,1]
  AI_pos = featureAAindex(seq, "ANDN920101") [,-22]
}

```

Figure 2. Program code for AAIndex feature extraction

2.3.2. Hydrophobicity

For hydrophobicity, we used BioSeqClass package to extract the feature by Hydro() function. Example of the program code used for extraction hydrophobicity is shown in Figure 3.

```
if (interactive()) {
  file_pos2 = file.path(path.package("BioSeqClass"),
    "example", "New_PosIndependent_C.pep")
  seq1_pos = as.matrix(read.csv(file_pos2, header =
    F))[,1]
  H1_pos = featureHydro(seq1_pos, "kpm")[,-22]
}
```

Figure 3. Program code for hydrophobicity feature extraction

2.3.3. Solvent accessiBiLitiEs

SABLE is a tool that is developed to determine the suitable folding for a sequence with arbitrary structure. For results, the users have to provide the protein name and its amino acid sequence through the SABLE proteins web-site <https://SABLE.cchmc.org/>. An example of feature extraction is seen in Figure 4.

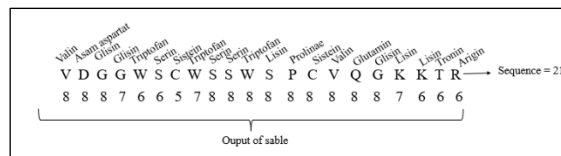


Figure 4. Output of SABLE

2.3.4. Composition, transition, and distribution

This research also utilizes the feature extraction of CTD [21], [32], [33]. The program code snippet used to obtain CTD feature extraction results can be seen in Figure 5.

```
if (interactive()) {
  file_neg1 = file.path(path.package("BioSeqClass"),
    "example", "id_NegIndependent_C.pep")
  tmp_neg = readAAStringSet(file_neg1)
  proteinSeq_neg = as.character(tmp_neg)
  #CTD1_neg = featureCTD(proteinSeq_neg, class =
  elements("aminoacid"))
  CTD2_neg = featureCTD(proteinSeq_neg, class =
  aaClass("aaV"))[, -22:-26]
  CTD2_neg[is.na(CTD2_neg)] <- 0
}
```

Figure 5. Program code for CTD feature extraction

2.3.5. Pseudo amino acid composition

For this feature, the BioSeqClass. The relevant code for PseAAC feature extraction is shown in Figure 6. The feature extraction stage functions to identify each feature. Each feature has different dimensions in each feature extraction. Following are the dimensions of each feature in Table 1.

```
if(interactive()){
  file_pos3 = file.path(path.package("BioSeqClass"),
    "example", "New_PosIndependent_C.pep")
  seq2_pos = as.matrix(read.csv(file_pos3,header=F, sep
    = ""))
  PAC4_pos = featurePseudoAACComp(seq2_pos,4)
}
```

Figure 6. Program code for PseAAC

Description	Dimensions	Percentage (%)
AAndex	21	19
Hindrophobicity	21	19
SABLE	21	19
CTD	21	19
PseAAC	24	24
Total	109	

2.4. Feature selection

Feature selection is very important in building a better classification so that the resulting data can be used [34]. Feature selection is used to reducing over fitting, reducing the number of features, or eliminate the irrelevant features that has lower prediction accuracy and achieve the better solutions. The feature selection method used here is MRMR [25]. A portion of the program code that was employed to realize the mRMR results is shown in Figure 7.

```
library(mRMR)
Independen_C<-read.csv("D:/DISERTASI
S3/GLIKOSILASI/IndependenC/Independen_C.csv"
, header = TRUE, sep=",")
Independen_C<-
mRMR.data(data=data.frame(Independen_C[,3:11
1, drop=FALSE]))
MRMR_Test<-mRMR.classic("mRMR.Filter",
data=Independen_C, target_indices=109,
feature_count = 25)
solutions(MRMR_Test)
```

Figure 7. Program code for feature selection using MRMR

MRMR feature selection is the stage of selecting features with the highest target correlation with the class or output of the prediction and the lowest redundancy correlation [35]. At this stage, the simplest is determined, namely, finding which features are most relevant and exclude data redundancy. The features used are those features decided by an optimal procedure.

2.5. Classification

This stage is the stage of machine learning modeling using 0 algorithms for classification [26]. Post-translation modification in C-glycosylation. XGBoost is a method used to solve supervised learning problems. XGBoost consists of training data (xi) which can predict target data (yi). XGBoost performance can be seen in the following equation: the objective function consists of training losses and regularization terms [36], which can be seen in (1):

$$Obj(\theta) = \mathcal{L}(\theta) + \Omega(\theta) \quad (1)$$

the L function shows the training data, while Ω is the parameter used [9]. The function to define training can be seen in (2):

$$\mathcal{L}(\theta) = \sum_{i=1}^n t(y_i, \hat{y}_i) \quad (2)$$

There are various types of methods available in the evaluation procedure out of which k-fold cross validation technique is successfully implemented and in this process, data used in developing the model is called training data while the data used for validating the model is called testing data [37]. Evaluation uses k-fold cross-validation five times. The steps of the k-fold simulation are shown in the Figure 8.

2.6. Performance measurement

Evaluate the model using a confusion matrix. The model was evaluated using several indicators, including accuracy (ACC), sensitivity (SN), specificity (SP), and the Matthew correlation coefficient (MCC) [38], [39] as shown in (3)-(7):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (6)$$

True positive (TP) is the number of glycosylation sites got correctly identified. False positive (FP) is the number of glycosylation sites classified as positive for glycosylation sites in the any specific condition. True negative (TN) indicates the number of non glycans where the program had successfully predicted them not to be glycosylated. False negative (FN) represents the number of non-glycosylated position classified as glycosylated.

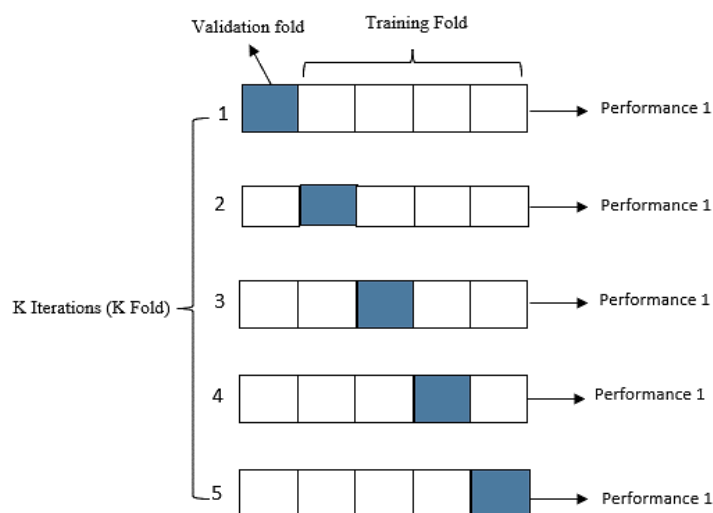


Figure 8. The k-fold simulation

3. RESULTS AND DISCUSSION

Feature extraction means the pinpointing of component attributes of the input raw data of the whole data source. The principal purpose is to reduce the size of the data, remove unwanted and isolate down important or representative variables for subsequent analysis. In this research utilized five feature extraction methods, namely: AAindex, SABLE, hydrophobicity, CTD, and PseAAC. Various feature extraction techniques used in the present study have shown to give improved results for the C-glycosylation prediction. This study shows how each feature extraction feature works to enhance the prediction accuracy as stated in Table 2.

Table 2. The contribution of each feature

Feature extraction type	Total feature contribution	Percentage (%)
SABLE	3	6
AAIndex	12	24
CTD	9	18
Hydrophobicity	12	24
PseAAC	14	28
Total	50	100

The next step is features selection, which involves choosing a subset of features that are most relevant or significant for analysis or modeling purposes. This process aims at reducing the data dimensionality, the enhancement of model accuracy, reduction of the problem of overfitting as well as increased understanding of how various features are related to the target variable. According to Table 2, the MRMR technique results in the selection of 50 features. Each extracted feature contributes to an improvement in the accuracy of glycosylation PTM prediction. Among the five feature extraction techniques utilized, PseAAC demonstrates the highest contribution, accounting for 28% of the total. This dominance of PseAAC feature extraction indicates its greater impact compared to other extraction methods. These features serve as a numerical representation of amino acid sequences on proteins, which can be utilized as features to train prediction models. The contribution of PseAAC feature extraction is deemed greater than that of other feature extraction techniques due to its suitability and sensitivity to glycosylation-related data. PseAAC is

commonly employed to identify distances or relationships between amino acids in sequences, suggesting its significant impact on glycosylation prediction.

The cross-validation method used in the study is the K-fold cross-validation process where the sample data is split into a training and a testing set. K- fold cross validation procedure involves partitioning of the dataset into K random subset which will be in affect K times used for training and testing. The study is opted for the five-fold cross-validation of the data set such that the data was split into equal five parts. In each loop, one part was used for validation and the remaining four segments were used in the training of the model. Five such splits were enacted and in each of the splits the corresponding subset was used for testing only. The XGBoost algorithm demonstrates superior performance, achieving higher accuracy compared to earlier studies. Our investigation reveals the widespread glycosylation of the carboxylate amino acid sequence. This improvement is attributed to the utilization of various feature extraction techniques, including SABLE, AAindex, CTD, hydrophobicity and PseAAC, along with the MRMR feature selection method. Further details and findings are provided in Table 3.

Table 3. The findings of the C-glycosylation prediction

Glycosylation data	ACC (%)	SN (%)	SP (%)	MCC (%)
C_glycosylation is not using MRMR	95	86.67	100	87.20
C_glycosylation using MRMR	100	100	100	100

Based on Table 3, the results of the study using the selection of MRMR features are 100%; without MRMR, they are 95%. This suggests that the MRMR technique is used in data analysis to determine which subset of features among the available feature sets is most relevant. The results of the current studies point that the carboxylate of amino acid sequence is almost always glycosylated. The test results using cross-validation with five-fold repetition can be seen in Figure 9.

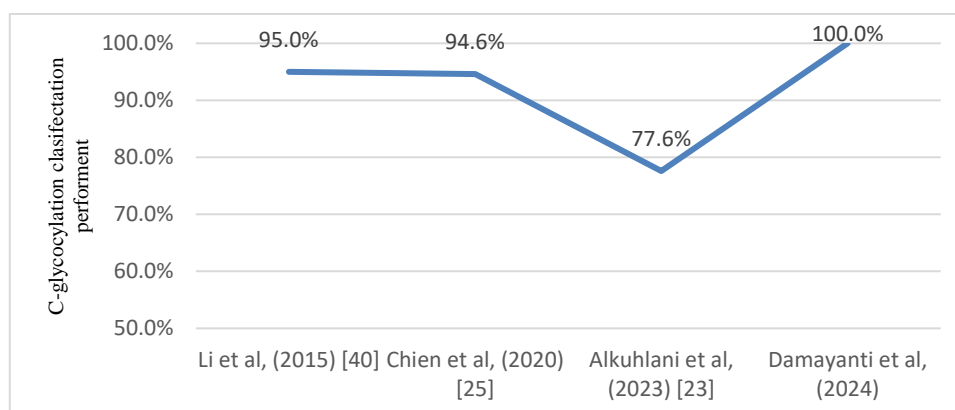


Figure 9. Comparison of performance analyses of previously developed approaches

Based on Figure 9, the performance of this study tends to be better than previous studies. This study achieved an accuracy of 100% when compared to the previous study, which was only 95.00%. The increase in performance resulting from the use of the approach carried out in this study reached 5% [40]. The results of the test also show that quantity of the feature affects the accuracy value.

4. CONCLUSION

Glycosylation prediction in C-glycosylation using the XGBoost algorithm consists of benchmark and independent data. Glycosylation prediction begins with feature extraction, which aims to convert the extracted string-type dataset into numeric-type features. They are five types of feature extraction techniques: SABLE, AAindex, CTD, hydrophobicity, PseAAC. After that, feature selection is done based on the MRMR method. Modeling using the XGBoost algorithm with k-fold testing 5 (five) times. Each iteration value is multiplied five times, and then the average value of the iteration is obtained. Next, the modeling was tested using cross-validation. The results of independent C-glycosylation data testing achieved an accuracy value of 100%.

ACKNOWLEDGEMENTS

The authors is grateful to Universitas Teknokrat Indonesia for their continuous sponsorship and for the grant obtained from Universitas Teknokrat Indonesia, research grant no 009/UTI/LPPMI/E.1.1/VIII/2023. The author also wishes to thank the reviewers for their feedback as their comments have been very useful in improving this work.




REFERENCES

- [1] F. Esmaili, M. Pourmirzaei, S. Ramazi, S. Shojailangari, and E. Yavari, "A review of machine learning and algorithmic methods for protein phosphorylation sites prediction," *Genomics, Proteomics, and Bioinformatics*, vol. 21, no. 6, pp. 1266–1285, 2023, doi: 10.1016/j.gpb.2023.03.007.
- [2] Q. Zhong *et al.*, "Protein posttranslational modifications in health and diseases: Functions, regulatory mechanisms, and therapeutic implications," *MedComm*, vol. 4, no. 3, pp. 1–112, 2023, doi: 10.1002/mco2.261.
- [3] W. R. Qiu, B. Q. Sun, X. Xiao, Z. C. Xu, and K. C. Chou, "iPTM-mLys: Identifying multiple lysine PTM sites and their different types," *Bioinformatics*, vol. 32, no. 20, pp. 3116–3123, 2016, doi: 10.1093/bioinformatics/btw380.
- [4] X. Yang and H. Han, "Factors analysis of protein O-glycosylation site prediction," *Computational Biology and Chemistry*, vol. 71, pp. 258–263, 2017, doi: 10.1016/j.compbiolchem.2017.09.005.
- [5] T. Pitti, C. T. Chen, H. N. Lin, W. K. Choong, W. L. Hsu, and T. Y. Sung, "N-GlyDE: a two-stage N-linked glycosylation site prediction incorporating gapped dipeptides and pattern-based encoding," *Scientific Reports*, vol. 9, 2019, doi: 10.1038/s41598-019-52341-z.
- [6] D. Wang *et al.*, "MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization," *Nucleic Acids Research*, vol. 48, no. W1, pp. W140–W146, 2021, doi: 10.1093/NAR/GKAA275.
- [7] Y. Zhang and L. Sun, "Sweetening the deal: glycosylation and its clinical applications," *Journal of Biomedical Sciences*, vol. 9, no. 3, pp. 1–7, 2020, doi: 10.36648/2254-609x.9.3.9.
- [8] Y. Mazola, G. China, and A. Musacchio, "Integrating bioinformatics tools to handle glycosylation," *PLoS Computational Biology*, vol. 7, no. 12, pp. 1–8, 2011, doi: 10.1371/journal.pcbi.1002285.
- [9] L. Zhang and C. Zhan, "Machine learning in rock facies classification: an application of XGBoost," *International Geophysical Conference*, Qingdao, China, 2017, pp. 1371–1374, doi: 10.1190/igc2017-351.
- [10] A. Kumar, V. Narayanan, and A. Sekhar, "Characterizing post-translational modifications and their effects on protein conformation using NMR spectroscopy," *Biochemistry*, vol. 59, no. 1, pp. 57–73, 2020, doi: 10.1021/acs.biochem.9b00827.
- [11] G. Taherzadeh, A. Dehzangi, M. Golchin, Y. Zhou, and M. P. Campbell, "SPRINT-Gly: Predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties," *Bioinformatics*, vol. 35, no. 20, pp. 4140–4146, 2019, doi: 10.1093/bioinformatics/btz215.
- [12] A. V. Everest-Dass, E. S. X. Moh, C. Ashwood, A. M. M. Shathili, and N. H. Packer, "Human disease glycomics: technology advances enabling protein glycosylation analysis—part 2," *Expert Review of Proteomics*, 2018, doi: 10.1080/14789450.2018.1448710.
- [13] W. Li, H. L. Li, J. Z. Wang, R. Liu, and X. Wang, "Abnormal protein post-translational modifications induces aggregation and abnormal deposition of protein, mediating neurodegenerative diseases," *Cell & Bioscience*, vol. 14, no. 1, pp. 1–14, 2024, doi: 10.1186/s13578-023-01189-y.
- [14] H. Javadikasgari, E. G. Soltesz, and A. M. Gillinov, "Surgery for Atrial Fibrillation," *Atlas of Cardiac Surgical Techniques*, pp. 479–488, 2018, doi: 10.1016/B978-0-323-46294-5.00028-5.
- [15] D. J. Vigerust, "Protein glycosylation in infectious disease pathobiology and treatment," *Open Life Sciences*, vol. 6, no. 5, pp. 802–816, 2011, doi: 10.2478/s11535-011-0050-8.
- [16] X. Hou, Y. Wang, D. Bu, Y. Wang, and S. Sun, "EMNGly: predicting N-linked glycosylation sites using the language models for feature extraction," *Bioinformatics*, vol. 39, no. 11, pp. 1–8, 2023, doi: 10.1093/bioinformatics/btad650.
- [17] T. T. D. Nguyen, N. Q. K. Le, T. A. Tran, D. M. Pham, and Y. Y. Ou, "Incorporating a transfer learning technique with amino acid embeddings to efficiently predict N-linked glycosylation sites in ion channels," *Computers in Biology and Medicine*, vol. 130, 2021, doi: 10.1016/j.compbiomed.2021.104212.
- [18] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Introduction to machine learning," in *Machine Learning: Methods and Applications to Brain Disorders*, 2019, pp. 1–20, doi: 10.1016/B978-0-12-815739-8.00001-8.
- [19] A. Ławryniewicz and V. Tresp, "Introducing machine learning," *Perspectives on Ontology Learning*, AKA Heidelberg/IOS Press vol. 18, pp. 35–50, 2014.
- [20] R. Sharma, S. S. Kamble, A. Gunasekaran, V. Kumar, and A. Kumar, "A systematic literature review on machine learning applications for sustainable agriculture supply chain performance," *Computers & Operations Research*, vol. 119, 2020, doi: 10.1016/j.cor.2020.104926.
- [21] Y. C. A. P. Reddy, P. Viswanath, and B. E. Reddy, "Semi-supervised learning: a brief review," *International Journal of Engineering & Technology*, vol. 7, no. 1.8, pp. 81–85, 2018, doi: 10.14419/ijet.v7i1.8.9977.
- [22] N. Fujii, T. Takata, N. Fujii, K. Aki, and H. Sakaue, "D-Amino acids in protein: the mirror of life as a molecular index of aging," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1866, no. 7, pp. 840–847, 2018, doi: 10.1016/j.bbapap.2018.03.001.
- [23] Y. Deng, Y. Fu, H. Zhang, X. Liu, and Z. Liu, "Protein post-translational modification site prediction using deep learning," *Procedia Computer Science*, vol. 198, no. 2021, pp. 480–485, 2021, doi: 10.1016/j.procs.2021.12.273.
- [24] A. Alkhlani, W. Gad, and M. Roushdy, "Prediction of o-glycosylation site using pre-trained language model and machine learning," *International Journal of Intelligent Computing and Information Sciences (IJICIS)*, vol. 23, no. 1, pp. 41–52, 2023, doi: 10.21608/ijicis.2023.160986.1218.
- [25] C.-H. Chien, C.-C. Chang, S.-H. Lin, C.-W. Chen, Z.-H. Chang, and Y.-W. Chu, "N-GlycoGo: predicting protein N-glycosylation sites on imbalanced data sets by using heterogeneous and comprehensive strategy," *IEEE Access*, 2020, doi: 10.1109/access.2020.3022629.
- [26] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," *Computers in Biology and Medicine*, vol. 121, 2020, doi: 10.1016/j.compbiomed.2020.103761.
- [27] P. Regan, P. L. McClean, T. Smyth, and M. Doherty, "Early stage glycosylation biomarkers in Alzheimer's disease," *Medicines*, vol. 6, no. 3, p. 92, 2019, doi: 10.3390/medicines6030092.




- [28] A. Bateman *et al.*, “UniProt: A hub for protein information,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015, doi: 10.1093/nar/gku989.
- [29] A. Bateman *et al.*, “UniProt: The universal protein knowledgebase,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D158–D169, 2017, doi: 10.1093/nar/gkw1099.
- [30] A. G. Sorkhi, J. Pirgazi, and V. Ghasemi, “A hybrid feature extraction scheme for efficient malonylation site prediction,” *Scientific Reports*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-08555-9.
- [31] L. Guo, D. Rivero, J. Dorado, C. R. Munteanu, and A. Pazos, “Automatic feature extraction using genetic programming: an application to epileptic EEG classification,” *Expert Systems with Applications*, vol. 38, no. 8, pp. 10425–10436, 2011, doi: 10.1016/j.eswa.2011.02.118.
- [32] R. Hou, J. Wu, L. Xu, Q. Zou, and Y. J. Wu, “Computational prediction of protein arginine methylation based on composition–transition–distribution features,” *ACS Omega*, vol. 5, no. 42, pp. 27470–27479, 2020, doi: 10.1021/acsomega.0c03972.
- [33] F. Indriani, K. R. Mahmudah, B. Purnama, and K. Satou, “ProtTrans-Glutar: incorporating features from pre-trained transformer-based models for predicting glutarylation sites,” *Frontiers in Genetics*, vol. 13, no. May, pp. 1–11, 2022, doi: 10.3389/fgene.2022.885929.
- [34] U. M. Khaire and R. Dhanalakshmi, “Stability of feature selection algorithm: A review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, 2022, doi: 10.1016/j.jksuci.2019.06.012.
- [35] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, “MRMR: An R package for parallelized mRMR ensemble feature selection,” *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, 2013, doi: 10.1093/bioinformatics/btt383.
- [36] T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [37] D. Berrar, “Cross-validation,” *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 2018. .
- [38] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, “Confusion-matrix-based kernel logistic regression for imbalanced data classification,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 9, pp. 1806–1819, 2017, doi: 10.1109/TKDE.2017.2682249.
- [39] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, pp. 1–13, 2020, doi: 10.1186/s12864-019-6413-7.
- [40] F. Li *et al.*, “GlycoMine: A machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome,” *Bioinformatics*, vol. 31, no. 9, pp. 1411–1419, 2015, doi: 10.1093/bioinformatics/btu852.

BIOGRAPHIES OF AUTHORS






Damayanti    received the magister computer of the Institut Teknologi Sepuluh Nopember in 2016. Currently, she is the Associate Professor at Department of Information Systems, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia. Her research interests include bioinformatics, computer science, information systems, and machine learning. She can be contacted at email: damayanti@teknokrat.ac.id.






Favorisen Rosyking Lumbanraja    is an Associate Professor at the Department of Computer Science, Faculty of Mathematics and Natural Science, University of Lampung, Indonesia. an M.Si. in Institut Pertanian Bogor, Indonesia, in 2011 and completed his Ph.D. at Kanazawa University, Japanese, in 2017. His research interests are bioinformatic, expert systems, machine learning, and computer science. He can be contacted at email: favorisen.lumbanraja@fmipa.unila.ac.id.






Akmal Junaidi    is an Associate Professor at the Department of Computer Science, Faculty of Mathematics and Natural Science, University of Lampung, Indonesia. an M.Sc. in Universiteit Twente, Dutch, in 2003 and completed his Dr.rer.nat. at Technische Universität Dortmund, German, in 2016. His research interests are expert systems, machine learning, computer science, and image processing. He can be contacted at email: akmal.junaidi@fmipa.unila.ac.id.






Sutyarso    is currently Professor at the Department of Biology, Faculty of Mathematics and Natural Sciences, University of Lampung, Indonesia. Currently, he is a lecturer at the University of Lampung, Indonesia. His research interests are in biology and medicinal plants. He can be contacted at email: sutyarso.1957@fmipa.unila.ac.id.



Gregorius Nugroho Susanto    is currently a Professor at the Department of Biology, Faculty of Mathematics and Natural Sciences, University of Lampung, Indonesia. an M.Sc. in Mississippi State University, USA in 1994 and completed his Ph.D. at Universite Montpellier II, France in 2001. His research interests are in biology, especially in aquatic animal physiology. He can be contacted at email: gregorius.nugroho@fmipa.unila.ac.id.



Nirwana Hendrastuty    received the magister computer of the Universitas Gadjah Mada in 2021. Currently, she is the Associate Professor at Department of Information Systems, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia. Her research interests include computer science, information systems, and machine learning. She can be contacted at email: nirwanahendrastuty@teknokrat.ac.id.