Journal of Computer Science

Original Research Paper

Model Classification for Predicting the Post-Translational Modification (PTM) Glycosylation in Sequence O Using an Extreme Gradient Boosting Algorithm

[1,2]Damayanti, [3]Sutyarso, [4]Akmal Junaidi and [4]Favorisen Rosyking Lumbanraja

[1]Faculty of Mathematics and Natural Science, Universitas Lampung, Lampung, Indonesia
[2]Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Lampung, Indonesia
[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia
[4]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia

Article history
Received: 06-10-2023
Revised: 28-12-2023
Accepted: 29-02-2024

Corresponding Author:
Damayanti
Faculty of Mathematics and Natural Science, Universitas Lampung, Lampung, Indonesia
Email: damayanti@teknokrat.ac.id

Abstract: Post Translational Modification (PTM) is an important mechanism involved in regulating protein function. Post-translational modification refers to the addition of covalent and enzymatic modifications of proteins in protein biosynthesis, which has an important role in modifying protein function and regulating gene expression. One of the post-translational modifications is glycosylation. Glycosylation is the addition of a sugar group to a protein structure. One type of glycosylation is glycosylation, which occurs in sequence O. Glycosylation has been linked to several illnesses, including diabetes, cancer, and the flu. Therefore, it is important to anticipate the occurrence of glycosylation by carrying out predicted glycosylated or non-glycosylated data. Glycosylation prediction has been widely done using manual laboratory techniques, which results in the prediction process being long and expensive for lab equipment. To overcome this, computerized data is needed that can predict glycosylation more quickly. The data used is glycosylation data on sequence O obtained from the UniProt website, which can be openly accessed. This study aimed to improve the accuracy of post-translational modification glycosylation in sequence O prediction using the method of extreme gradient boosting as a framework for gradient enhancement that tends to be faster. This accuracy is increased by conducting feature extraction experiments with the following types: AAIndex, hydrophobicity, sable, composition, CTD, and PseAAC. Feature selection uses the MRMR approach. Evaluation using k-fold cross-validation. The results of this study indicate the prediction performance of post-translational modification glycosylation in sequence O with an accuracy value of 100%. The study's findings indicate that the XGBoost algorithm performs better than other research that has been conducted.

Keywords: Glycosylation, XGBoost, Machine Learning, Sequence

Introduction

Post-translational modification (PTM), often known as post-translational modification, is an important mechanism involved in regulating protein function. The term 'post-translational modification' describes enzymatic and covalent additions made to proteins during or following protein production. These modifications are crucial for controlling gene expression and altering the function of proteins (Minguez *et al*., 2012). Post-translational modification refers to the addition of covalent and enzymatic modifications of proteins during protein biosynthesis, which has an important role in modifying protein function and regulating gene expression (Tak *et al*., 2019). Phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, and lipidation are all examples of post-translational modification (Caragea *et al*., 2007; Yang and Han, 2017). When

compared to other post-translational modifications, glycosylation has the most variability since it involves the addition of sugar groups to the protein structure. When an enzyme can create a glycan, which links a nucleotide sugar to one of the amino acids, in this case, asparagine, glycosylation takes place. The most challenging step in protein modification is glycosylation (Naik *et al*., 2018). Glycosylation can also be tied to various diseases, including diabetes, cancer, and influenza (Taherzadeh *et al.*, 2019; Weerapana and Imperiali, 2006). Glycan alterations can result in illness (Everest-Dass *et al.*, 2018). Alzheimer's disease can also be brought on by glycosylation. Alzheimer's disease is a brain illness that causes memory loss in its victims (Regan *et al.*, 2019).

Therefore, predicting glycosylation is very important to determine whether a sequence is glycosylated or not. Glycosylation prediction has traditionally been performed using manual laboratory techniques, which makes the prediction process lengthy and expensive due to the required lab equipment. While numerous techniques, such as genetic modification, genetic engineering, and other biological studies, have been utilized for glycosylation prediction, these methods still necessitate repeated testing, thereby prolonging the process (Li *et al.*, 2019). With its advancement, computational technology is required to more swiftly forecast glycosylation (Lumbanraja *et al.*, 2018). Therefore, to solve these issues, a computational data model using machine learning techniques to predict glycosylation must be created (Chien *et al.*, 2020). In the study of protein function, computational approaches are crucial for understanding post-translational modification (Bateman *et al.*, 2014).

Machine learning is a branch of artificial intelligence that involves the development of algorithms capable of completing tasks, often similar to those performed by humans. Through the use of computers, machine learning enables the prediction of data based on patterns and trends identified within datasets. The field has undergone significant advancements and development, employing various methodologies to enhance the capabilities of machine learning algorithms (Vieira *et al*., 2020) Despite the little data, machine learning techniques are thought to be able to make predictions with a better level of accuracy (Lumbanraja *et al.*, 2019).

An identical problem has been previously addressed in several earlier experiments involving glycosylation prediction. These experiments also utilized machine learning techniques for glycosylation prediction, but the accuracy results still require improvement. The previous study demonstrated a glycosylation projection accuracy of 77-86%. (Alkuhlani *et al*., 2023). Additionally, the study also rectified O-glycosylation, giving a 90.7% accuracy rate (Li *et al.*, 2015). The purpose of this study was to increase glycosylation prediction accuracy by using the extreme gradient boosting algorithm. Gradient boosting is an algorithm capable of identifying the best solutions to a wide range of issues, particularly in the areas of regression, classification, and ranking. The fundamental idea of this algorithm is to continuously modify the learning parameter.

XGBoost is a framework for gradient enhancement introduced by Friedman in 2001 that tends to be more efficient, scalable, and faster (Chen and Guestrin, 2016; Chen and He, 2024; Zhang and Zhan, 2017). The XGBoost package includes solutions for linear models and tree-learning algorithms (Chien *et al.*, 2020). XGBoost has reliable features such as speed in performance and a customizer that supports objective and evaluation functions, resulting in better performance across various datasets (Chen *et al*., 2015). Currently, XGBoost is the most popular algorithm for addressing machine learning challenges. XGBoost provides a more structured approach to creating regression tree structures, yielding improved results and simplifying the model. Essentially, the XGBoost technique represents an algorithmic evolution from gradient tree-boosting ensembles. The XGBoost method was chosen due to its additional features that enhance computational efficiency. XGBoost can effectively handle a wide range of regression and classification scenarios. The computation involves assembling a collection of trees, known as XGBoost.

To achieve superior performance, we conducted experiments involving several feature extraction methods. Specifically, glycosylation prediction was performed using five types of feature extraction: Amino Acid Index (AAIndex), hydrophobicity, Solvent Accessibility (Sable), Composition Transition and Distribution (CTD), and Pseudo-Amino Acid Composition (PseAAC). The purpose of feature extraction is to convert string data into numerical data suitable for computer processing.

The novelty and contribution of our proposed study lie in the addition of SABLE feature extraction and hydrophobicity, which were not previously explored in order to enhance glycosylation predictive accuracy. Furthermore, efforts to enhance glycosylation-O prediction accuracy will involve selecting features using the Minimum Redundancy Maximum Relevance (MRMR) approach. This research is crucial as it identifies glycosylated proteins as key subjects for research in diagnosing diseases resulting from the glycosylation process. Nearly all proteins in human and other mammalian cells undergo glycosylation.

Materials and Methods

The data for this research consists of amino acid sequences for various types of O-glycosylation, which are sequences of serine obtained from the UniProt website (https://www.UniProt.org/). The European Molecular

Biology Organization collaborated to create the protein sequence and annotation database UniProt (Bateman *et al*., 2014). Data in the form of human proteins were extracted from these databases. We investigated the amino acids of the various types of O-glycosylation based on accurate data. The research stages consist of data collection, data preprocessing, feature extraction, feature selection, modeling and evaluation, and results in Fig. 1.

This section presents the research method. The research stages consist of data collection, data preprocessing, feature extraction, feature selection, modeling and evaluation, and results in Fig. 1.

Data Collection

Various types of O-glycosylation, which consist of sequences of serine, were openly retrieved from the UniProt website (https://www.UniProt.org/). The European Molecular Biology Organization collaborated to create the protein sequence and annotation database UniProt (Bateman *et al*., 2014). Data in the form of human proteins were extracted from these databases. We investigated the amino acid compositions of the various types of O-glycosylation based on accurate data. We used R-Studio software for our analysis. The UniProt Database is a website that provides protein sequences (Bateman *et al*., 2015). The following are the algorithms used to collect the data (Bateman *et al*., 2015). The following are the algorithms used to collect the data.

Algorithm 1: Collecting Data

```
library(protr)
NegIndependent_O<-
read.csv("D:/GLIKOSILASI/DATA
/NegIndependent_O.txt", header = FALSE, sep="\t")
windows=21;
jumlah_seq=1
seq_iter=array()
sink('New_NegIndependent_O.fasta')
for(j in 1:nrow(NegIndependent_O)){
   prots<-getUniProt(NegIndependent_O[j,2])
   start=NegIndependent_O[j,3]-((windows-1)/2)
   end=NegIndependent_O[j,3]+((windows-1)/2)
    sq=substr(prots[[1]],start,end)

    if(nchar(sq)==windows){

#cat(paste('>',NegIndependent_O[j,2],NegIndependent_
O[j,3],'\n'))
          cat(paste(sq,'\n'))
          seq_iter[jumlah_seq]=sq
          jumlah_seq=jumlah_seq+1
 }
}

sink()
```

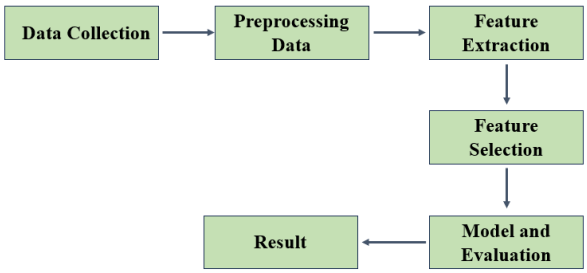

Fig. 1: The research stages

The collected data is presented in the form of a collection grouped based on benchmark data and independent data, each of which contains negative and positive values. Negative data represents instances that are not glycosylated, while positive data represents instances that are glycosylated. The data used is presented in Table 1.

Preprocessing Data

The initial data amounted to 1925, consisting of benchmark data and independent data, which include negative and positive data, respectively, as shown in Table 1. The displayed data represent 21 residues of protein glycosylation, derived from sequences of both 10 lengths of the right sequence and 10 lengths of the left sequence. Algorithm 1 illustrates this information. Following data collection, the next step involves data preprocessing. One of the phases in data preprocessing is data cleaning The first thing that needs to be done is cleaning the data, which is sometimes referred to as data preparation. This suggests that further screening of the unprocessed data is needed. Next, eliminate or erase any inaccurate, superfluous, or incomplete data. Skipredundant is the technique used. Example of the data collected in Fig. 2. Next, we eliminate redundant data to obtain optimal data. The tools we use to eliminate redundant data are skip redundant no greater than 30% (Li *et al*., 2015; Chien *et al*., 2020). This was a crucial step in removing sequence redundancy and preventing overestimations of machine learning-based classifiers' performance (Li *et al*., 2015). This procedure proved crucial for removing sequence redundancy and preventing overestimations of machine learning-based classifiers' performance. The cleaned data totaled 204 instances, ready to be processed for modeling. The amount of cleaned data can be seen in Table 2. The dataset contains both negative and positive labels, where the negative label indicates that the dataset is not glycosylated, while the positive label indicates that the dataset is glycosylated. The O-sequence data for length 21 is presented in Table 3.

Table 1: Dataset of glycosylation sequence O on the website UniProt

Types of glycosylation	Amount of data	Actual
NegBenchMark_O	1018.000	Negative
NegIndependent_O	258.000	Negative
PosBenchMark_O	520.000	Positive
PosIndependent_O	129.000	Positive
Total	1.925	

The post-translational modification glycosylation in sequence O dataset available on the UniProt website is 1.925 sequences

Table 2: Clean data

Types of glycosylation	Amount of data	Actual
NegBenchMark_O	53	Negatif
NegIndependent_O	37	Negatif
PosBenchMark_O	76	Positif
PosIndependent_O	38	Positif
Total	204	

Table 3: O-sequence data

ID protein	Position	Types of glycosylation	Sequence	Label
Q8TDJ6	588	NegBench Mark_O	LLHQEGMSVGS PHGSQPHSRS	Not glycosylated
Q8TDJ6	2711	NegBench Mark_O	IGEEYDRESKSS DDVDYRGST	Not glycosylated
E7EV10	52	NegBench Mark_O	VVCFYRRRDIS NTLIMLADKH	Not glycosylated
E7EV10	552	NegBench Mark_O	KKPNVIRSTP SLQTPTTKRML	Not glycosylated
P13647	62	NegBench Mark_O	AGACGVGGYG SRSLYNLGGSK	Not glycosylated
P13647	232	NegBench Mark_O	LLFRTSLKFRN THLGKKGSEI	Not glycosylated
E9PJ03	98	NegBench Mark_O	QISGVKKLM HSSSLNNTSISR	Not glycosylated
E9PJ03	100	NegBench Mark_O	SGVKKLMH SSSLNNTSISRFG	Not glycosylated
E5RJ61	21	NegBench Mark_O	PGSVSPSR DSSVPGSPSSIV	Not glycosylated
E9PJU3	81	NegBench Mark_O	QQFLPQFPED SAEQQNELILA	Not glycosylated
A3KN83	588	PosBench Mark_O	TIVMTKTPP VTTNRQTITLTK	Glycosylated
E7ENI0	2711	PosBench Mark_O	GALQQKIPG VSTPQTLAGTQK	Glycosylated
O14639	552	PosBench Mark_O	VRDRMIHRST SQGSINSPVYS	Glycosylated

Feature Extraction

The collected data requires further extraction by performing hydrophobicity, Amino Acid Index (AAIndex), solvent accessibility (Sable), Composition Transition and Distribution (CTD), and Pseudo-Amino Acid Composition (PseAAC). Feature extraction aims to convert string data into numeric data. This stage involves feature extraction so that the resulting data can be used (Khaire and Dhanalakshmi, 2022). Feature extraction is very important in the machine learning process. Feature extraction aims to produce higher accuracy. The feature extraction stage is one of the most critical steps in machine learning research.

This feature extraction aims to increase the accuracy of predicting Post-Translational Modification (PTM) to sequence O. There are five types of feature extraction used in this study.

Amino Acid Index (AAindex)

The package used for feature extraction uses the Amino Acid Index (AAindex), namely the BioSeqClass package with the feature index () function. This feature consists of 21 amino acid sequences, which are then stored in the example directory in a document storage format with the format. pep.

Hydrophobicity

The package used for feature extraction using hydrophobicity is the BioSeqClass package with the featureHydro (). The feature consists of a 21-residue amino acid sequence, which is then saved in the example directory in the document with the format. pep. The physical characteristic of a chemical that makes it resistant to water masses is called hydrophobicity (Verlicchi *et al*., 2013).

Composition, Transition and Distribution (CTD)

The package used for feature extraction uses Composition, Transition, and Distribution (CTD), namely the BioSeqClass package with the featured () function. This feature consists of 21 amino acid sequences; then, it is stored in the example directory with a document storage format with format. Text.

Pseudo Amino Acid Composition (PseAAC)

In this feature extraction, the package used is the BioSeqClass package with the featurePseudoAAComp() function. This feature consists of 24 amino acid sequences, which are then stored in the example directory in a document storage format with format. Pep.

Solvent Accessibility (Sable)

Sable is a website used for generic structure prediction that identifies the folds of a given sequence that are most compatible. The name of the protein sequence and the amino acid sequence can be entered on the website Sable Protein https://sable.cchmc.org/ to acquire results. The feature extraction result is then delivered via email.

The five extractions are combined into one using the bind () function. All positive and negative data extractions are combined and then saved in CSV format. Then the next step is the labeling of each class. The label consists of 0 and 1, meaning class 0 is negative and class 1 is positive. Data extraction using Sable is shown in Fig. 2.

Figure 2 shows the sequence's length on the first line and the feature extraction output using Sable is shown on the second line. Solvent Accessibility (SABLE) transforms 21 characters. Character sequences from the feature extraction procedure into numeric values. Numerical data that can be processed by XGBoost modeling.

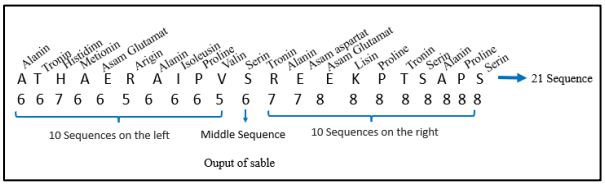

Fig. 2: Output of sable

Table 4: Feature extraction method contributes

Feature extraction type	Total feature contribution	Percentage (%)
Sable	11	44
AAIndex	1	4
CTD	3	12
Hydrophobicity		
	1	4
PseAAC	9	36
	Total	100

Based on Table 4, it is known that the feature contributing the most to the increase in post-translation modification for predicting O-glycosylation is the extraction of the SABLE feature, which accounts for 44%. Meanwhile, the smallest contribution is shown by the AAIndex extraction and the hydrophobicity characteristics, each contributing 4%.

Feature Selection

The data collected is in the form of benchmark and independent data, which are then combined into a dataset. To achieve optimal accuracy, this study employs feature selection using the Minimum Redundancy Maximum Relevance (mRMR) technique. mRMR is a technique that filters features based on two variables: relevance and redundancy. The primary objective of the mRMR technique is to reduce redundancy between features while retaining the most relevant features related to the target variable for forecasting or classification. The feature selection stage aims to select features with the highest relevance to the target variable and minimize redundancy. The application of MRMR feature selection to O-glycosylation was investigated in this research (Chien *et al*., 2020; Alkuhlani *et al*., 2022). he MRMR feature selection process consists of two stages: selecting features with the highest correlation level to generate the most relevant features, and then refining the selection from the first stage to minimize redundancy between the selected features. This algorithm is believed to enhance model accuracy by reducing data. The way that mRMR functions is as follows:

1. Minimal redundancy: Minimal redundancy: This stage determines how each feature relates to the target or class you wish to predict using relevance calculation. Metrics such as mutual information, correlation, and other statistically significant variables are typically used to measure it

The next stage involves reducing redundancy within the features themselves, after determining which features exhibit a strong association with the objective. One approach to achieve this is by selecting features with high correlation among several others and removing features that are highly interconnected.

2. Maximum significance: sThis stage prioritizes the selection of attributes that are most relevant to the intended audience. Features that exhibit a strong correlation with the target variable are given priority in this strategy

This process considers the relationship of each feature to the target features to ensure the diversity of information represented by those features.

This study employs the mRMR approach for feature selection. The objective of the feature selection stage is to enhance accuracy. The mRMR feature selection process requires the use of the mRMR library, specifically the mRMR.classic() function. The feature selection in mRMR involves selecting 25 features. The document will be saved in .csv format to select the desired variables. The use of mRMR was chosen because this approach has been shown to improve accuracy.

Model and Evaluation

This stage involves modeling and evaluation to predict post-translational modifications in O-glycosylation. During this phase, glycosylation prediction modeling is performed using the extreme gradient boosting method. Gradient boosting has evolved into the XGBoost technique. The machine learning model utilized in the XGBoost approach is built using gradient decision trees and is expected to enhance performance. Gradient boosting, as a technique, is capable of identifying optimal solutions to various problems, particularly those involving regression and classification. The fundamental concept of this algorithm is to minimize the loss function by adjusting parameters during iterative learning. To mitigate model complexity and prevent overfitting, XGBoost constructs a regression tree structure using a more regularized model, resulting in improved outcomes (Ma *et al*., 2020). At present, the most widely used algorithm for solving machine learning problems is XGBoost. With XGBoost, regression tree structures can be created in a more structured manner, which can simplify the model and yield better results.

The following equation represents the training loss and regularization terms that constitute the objective in Eq. 1:

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

The function shows the training loss, while Ω is the regularization term function. Then θ is the parameter used (Zhang and Zhan, 2017). The function of defining training loss can be seen in Eq. 2:

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

The yi function shows the valid actual value, while the $\hat{y}i$ function shows the predicted value. Then, the function n is the number of iterations used to get a better agreement, which can be seen in Eq. 3:

$$L(\theta) = -[yi \log(\hat{y}i) + (1 - yi) \log(1 - \hat{y}i)] \quad (3)$$

where, yi is the actual value considered accurate and $\hat{y}i$ is the predicted result of the model, while *n* is the number of iterations of the input value for the related model.

Next, we conducted an evaluation using cross-validation. This cross-validation is employed to assess the accuracy of the model. The cross-validation strategy utilizes the k-fold cross-validation method, where the data utilized for the model creation process is referred to as training data, and the data employed for model validation is referred to as testing data. The dataset was randomly split, and k experiments were performed to evaluate the cross-validation performance. Accuracy data is averaged after experiments using the k^{th} partition data, which is utilized as both training and testing data. In this study, cross-validation was performed up to five times to achieve the desired prediction results. Figure 3 illustrates the k-fold cross-validation simulation.

The results of the prediction performance of post-translational modifications in O-glycosylation are represented in a matrix, which is presented in terms of Accuracy (ACC), Sensitivity (Sn), Specificity (Sp), and Matthews Correlation Coefficient (MCC). In the confusion matrix, the number of accurately predicted glycosylation sites is known as True Positive (TP). The number of glycosylation sites incorrectly predicted as positive is False Positive (FP). The number of correctly predicted non-O-glycosylation sites is True Negative (TN). The number of non-glycosylated sites incorrectly predicted as negative is known as False Negative (FN).

The following are the equations used for *Accuracy*, *Sn*, *Sp*, and MCC:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

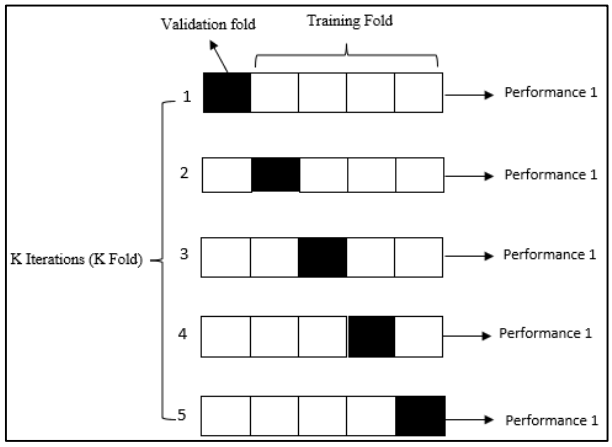

Fig. 3: K-fold cross-validation simulation

$$Sn = \frac{TP}{TP+FN} \quad (5)$$

$$Sp = \frac{TN}{TN+FP} \quad (6)$$

$$MCC = \frac{TP\ X\ TN - FP\ X\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

Results and Discussion

The research findings will be explained in detail below, based on the investigation.

Performance Evaluation

This research has succeeded in carrying out five types of feature extraction. The five feature extraction techniques have contributed to improving the accuracy of glycosylation prediction. How each feature extraction method contributes to improving the accuracy of glycosylation prediction can be seen in Table 4. The SABLE feature contributes the most to increasing the accuracy of post-translational prediction of glycosylation modifications in sequence O. The increase in accuracy of glycosylation prediction is also influenced by feature selection using the mRMR technique. This technique takes optimal features from the most relevant features with a low level of redundancy.

The predictive performance of post-translational alterations in O-glycosylation is depicted by a matrix, which is presented in terms of accuracy, sensitivity, specificity, and Matthews correlation coefficient. The performance of each benchmark data, training data, and independent data is shown in Table 3. Consequently, the accuracy of glycosylation prediction on sequence O using the XGBoost algorithm tends to increase. This is demonstrated by a benchmark accuracy of 99.27% and an independent accuracy of 100%, as indicated in Table 5: Post Translational Modification (PTM) glycosylation results.

Table 5: Results of the glycosylation on sequence O test

Glycosylation sequence O	Accuracy	Sensitivity	Specificity	Matthews correlation coefficient
Benchmark	99,27	98,13	100	98,64
Independent	100	100	100	100

Table 6: Comparative research

Year	Author	Method	Accuracy (%)	Year
2015	Li *et al*. (2015)	Random forest	95	2015
2020	Chien *et al*. (2020)	XGBoost	94,60	2020
2023	Alkuhlani *et al*. (2023)	XGBoost	77,86	2023
2023	Damayanti *et al*. (2023)	XGBoost	100	2023

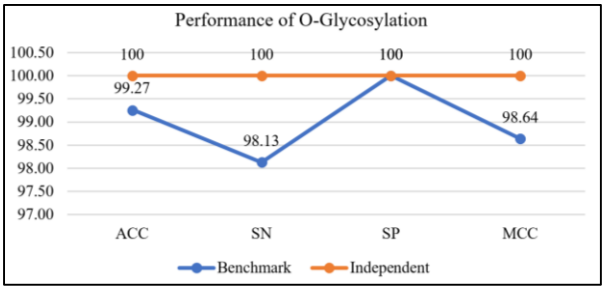

Fig. 4: Performance of Post Translational Modification (PTM) to sequence O

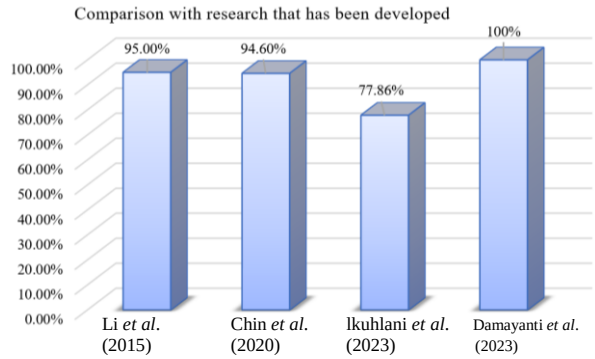

Fig. 5: Comparison with research that has been developed

Table 5 shows the accuracy, sensitivity, specificity, and Matthew correlation coefficient values for the research conducted. The results of the study can also be seen in Fig. 4. The results showed that the accuracy rate of the XGBoost algorithm is a recommendation that can be used for the prediction of Post Translational Modification (PTM) glycosylation on sequence O. Prediction using the XGBoost algorithm was 100%. Testing using the XGBoost model shows that the accuracy prediction for benchmark data tends to be higher. Namely, 99.27% compared to the independent 100%, thus outperforming previous studies. In the survey, benchmark data accuracy was 95%, while independent data was 95% (Pitti *et al*., 2019). Then, for research by Chien *et al*., (2020), the study's results showed an accuracy rate for Post Translational Modification (PTM) glycosylation on sequence O independent data of 94.6%. Furthermore, this study also outperformed research (Alkuhlani *et al*., 2023). This study discusses independent data on o-glycosylation with an accuracy of 77.86%.

Comparative Research

This study shows that the accuracy results tend to be higher compared with research that has been developed previously. In comparison to previous research, there was a 5% increase. The increase in the accuracy of predictions of post-translational modification of glycosylation in sequence O was influenced by several experiments that we carried out. First, we cleaned the data using skip redundant from 1925 data to 204 so that the processed data tends to be small. This can increase the accuracy value because the data used is the most optimal data. First, we cleaned the data using skip redundant from 1925 data to 204 so that the processed data tends to be small. This can increase the accuracy value because the data used is the most optimal data. Second, we carried out five types of feature extraction experiments, each of which contributed to increasing the accuracy value of O-glycosylation predictions. Third, we carried out feature selection using mRMR to obtain the most correlated features and those with the lowest level of redundancy. Fourth, we used the XGBoost algorithm modeling to improve O-glycosylation predictions so that, when compared with previous research, the accuracy of our research tends to increase. A comparison of the results of this research with several previous studies can be seen in Table 6.

A comparison of the results of the research that has been carried out with previous research can be seen in Fig. 5 tends to have a higher accuracy value, namely 100%. with an increase in accuracy of 5%. The findings of this study demonstrate that the performance of the XGBoost algorithm offers a greater level of accuracy than glycosylation prediction using a variety of previous methods. Such as for research by Chien *et al*., (2020) the study's results showed an accuracy rate for Post Translational Modification (PTM) glycosylation on sequence O independent data of 94.6%. Furthermore, this study also outperformed research (Alkuhlani *et al*., 2023).

The XGBoost algorithm is a recommendation that can be used to predict Post Translational Modification (PTM) glycosylation on O sequence glycosylation. Things that increase the accuracy of glycosylation predictions include: First, the data processed is small, namely 204 out of 1925, so it can affect the quality of the data. Second, reliability in handling XGBoost features can overcome the problem of high-impact or irrelevant features. This algorithm can identify and extract the most important features from the data. Comparing the research findings to earlier studies, the XGBoost algorithm performs better in this study. Thus, XGBoost can prioritize these features and ignore features that are less influential. This helps improve accuracy by focusing on the most informative features. Thus, the XGBoost algorithm can be recommended for future research by adding larger amounts of data with predictions of other post-translational modifications.

Conclusion

Prediction of Post Translational Modification (PTM) Glycosylation on Sequence O using the XGBoost algorithm can be used to increase accuracy. The results of this research show that the accuracy value is 100%, which is shown in Fig. 5. The evaluation of the model uses k-fold cross-validation with a parameter of k = 5. The increase in accuracy that has been achieved is supported by five types of feature extraction techniques, namely: AAindex, hydrophobicity, CTD, SABLE, and PseAAC. Each type of feature extraction contributed to increasing the accuracy of O glycosylation predictions. SABLE contributed the most, namely 44%, AAIndex contributed 4%, CTD contributed 12%, hydrophobicity contributed 4% and PseAAC contributed 36%. Furthermore, this very high increase in accuracy is also influenced by feature selection using the MRMR approach technique. This technique aims to reduce the dimensions of the dataset by selecting the most relevant and least redundant features. Prediction of O-glycosylation using the Extreme Gradient Boosting algorithm tends to be faster and saves experimental costs. This research succeeded in improving compared to previously developed research. The implications of this research can be used further in the clinical field for drug development. Glycosylation prediction by applying feature extraction techniques and feature selection and modeling using XGBoost can predict glycosylation quickly. This research has limitations with datasets, so it can be developed in further research by adding the amount of data with a sequence length of 51.

Acknowledgment

The author would like to express his deepest gratitude to the Universitas Teknokrat Indonesia for continuous support for the successful completion of this study and thank the reviewers for their valuable comments and suggestions that contributed to the improvement of this study.

Funding Information

The research paper focusing on glyskosylation in sequence O was from Universitas Teknokrat Indonesia.

Author's Contributions

Damayanti: Contributed to the research, design, literature review, implementation of the proposed algorithm, analysis of results, comparison with other existing algorithms, and written of the manuscript.

Sutyarso: Contributed to conceptualization, edited, and reviewed.

Akmal Junaidi: Contributed to conceptualization, analysis of results, and manuscript written, edited, and reviewed.

Favorisen Rosyking Lumbanraja: Contribution of conceptualization, design, results analysis, a manuscript written.

Ethics

The authors declare that there are no ethical issues that may arise after the publication of this manuscript.

References

Alkuhlani, A., Gad, W., Roushdy, M., & Salem, A. B. M. (2022). Pustackngly: Positive-unlabeled and stacking learning for n-linked glycosylation site prediction. *IEEE Access*, *10*, 12702-12713. https://doi.org/10.1109/ACCESS.2022.3146395

Alkuhlani, A., Gad, W., Roushdy, M. I., & Salem, A. B. M. (2023). Prediction of O-Glycosylation Site Using Pre-Trained Language Model and Machine Learning. *International Journal of Intelligent Computing and Information Sciences*, *23*(1), 41-52. https://doi.org/10.21608/IJICIS.2023.160986.1218

Bateman, A. *et al*. (2015). UniProt: A hub for protein information. Nucleic Acids Research, 43(D1), pp. D204-D212. https://doi.org/10.1093/nar/gku989

Bateman, A. *et al*. (2014). UniProt: The universal protein knowledgebase. Nucleic Acids Research. *Oxford University Press*, *45*(D1), pp. D158-D169. https://doi.org/10.1093/nar/gkw1099

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., & Honavar, V. (2007). Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics*, *8*, 1-13. https://doi.org/10.1186/1471-2105-8-438

Chen, T. & He, T. (2024). XGboost: Extreme Gradient Boosting. pp. 1-3. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: Extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1-4. https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf

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. https://doi.org/10.1109/ACCESS.2020.3022629

Everest-Dass, A. V. Moh, S. X. E. Ashwood, C. Shathili, M. M. A. & Packer, H. N. (2018). Human disease glycomics: Technology advances enabling protein glycosylation analysis - part 2 Taylor and Francis. https://doi.org/10.1080/14789450.2018.1448710

Khaire, U. M., & Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, *34*(4), 1060-1073. https://doi.org/10.1016/j.jksuci.2019.06.012

Li, F., Li, C., Wang, M., Webb, G. I., Zhang, Y., Whisstock, J. C., & 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. https://doi.org/10.1093/bioinformatics/btu852

Li, F., Zhang, Y., Purcell, A. W., Webb, G. I., Chou, K. C., Lithgow, T., ... & Song, J. (2019). Positive-unlabelled learning of glycosylation sites in the human proteome. *BMC Bioinformatics*, *20*, 1-17. https://doi.org/10.1186/s12859-019-2700-1

Lumbanraja, F. R., Nguyen, N. G., Phan, D., Faisal, M. R., Abipihi, B., Purnama, B., ... & Satou, K. (2018). Improved protein phosphorylation site prediction by a new combination of feature set and feature selection. *Journal of Biomedical Science and Engineering*, *11*(6), 144-157. http://repository.lppm.unila.ac.id/id/eprint/9593

Lumbanraja, F. R., Mahesworo, B., Cenggoro, T. W., Budiarto, A., & Pardamean, B. (2019). An evaluation of deep neural network performance on limited protein phosphorylation site prediction data. *Procedia Computer Science*, *157*, 25-30. https://doi.org/10.1016/j.procs.2019.08.137

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. *Computers in Biology and Medicine*, *121*, 103761. https://doi.org/10.1016/j.compbiomed.2020.103761

Minguez, P., Letunic, I., Parca, L., & Bork, P. (2012). PTMcode: A database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic Acids Research*, *41*(D1), D306-D311. https://doi.org/10.1093/nar/gks1230

Naik, H. M., Majewska, N. I., & Betenbaugh, M. J. (2018). Impact of nucleotide sugar metabolism on protein N-glycosylation in Chinese Hamster Ovary (CHO) cell culture. *Current Opinion in Chemical Engineering*, *22*, 167-176. https://doi.org/10.1016/j.coche.2018.10.002

Pitti, T., Chen, C. T., Lin, H. N., Choong, W. K., Hsu, W. L., & Sung, T. Y. (2019). N-GlyDE: A two-stage N-linked glycosylation site prediction incorporating gapped dipeptides and pattern-based encoding. *Scientific Reports*, *9*(1), 15975. https://doi.org/10.1038/s41598-019-52341-z

Regan, P., McClean, P. L., Smyth, T., & Doherty, M. (2019). Early stage glycosylation biomarkers in Alzheimer's disease. *Medicines*, *6*(3), 92. https://doi.org/10.3390/medicines6030092

Tak, I. R. Ali, F. Dar, J. S. Magray, A. R. Ganai, B. A. & Chishti, M. Z. (2019). Posttranslational Modifications of Proteins and Their Role in Biological Processes and Associated Diseases. *Protein Modificomics*, (pp.1-35) https://doi.org/10.1016/B978-0-12-811913-6.00001-1

Taherzadeh, G., Dehzangi, A., Golchin, M., Zhou, Y., & Campbell, M. P. (2019). SPRINT-Gly: Predicting N-and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties. *Bioinformatics*, *35*(20), 4140-4146. https://doi.org/10.1093/bioinformatics/btz215

Verlicchi, P., Zambello, E., & Al Aukidy, M. (2013). Removal of pharmaceuticals by conventional wastewater treatment plants. *In Comprehensive Analytical Chemistry*, (Vol. *62*, pp. 231-286). Elsevier. https://www.sciencedirect.com/science/article/abs/pii/B9780444626578000082

Vieira, S., Lopez Pinaya, W. H., & Mechelli, A. (2020). Introduction to machine learning in Machine Learning. *Methods and Applications to Brain Disorders*. https://doi.org/10.1016/B978-0-12-815739-8.00001-8

Weerapana, E., & Imperiali, B. (2006). Asparagine-linked protein glycosylation: From eukaryotic to prokaryotic systems. *Glycobiology*, *16*(6), 91R-101R. https://doi.org/10.1093/glycob/cwj099

Yang, X., & Han, H. (2017). Factors analysis of protein O-glycosylation site prediction. *Computational Biology and Chemistry*, *71*, 258-263. https://doi.org/10.1016/j.compbiolchem.2017.09.005

Zhang, L., & Zhan, C. (2017, May). Machine learning in rock facies classification: An application of XGBoost. *In International Geophysical Conference, Qingdao, China, 17-20 April 2017*, (pp. 1371-1374). Society of Exploration Geophysicists and Chinese Petroleum Society. https://doi.org/10.1190/IGC2017-351