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

12 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

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
Biography 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")
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=strsplit(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()
Table 1: Dataset of glycosylation sequence O on the website UniProt

<table>
<thead>
<tr>
<th>Types of glycosylation</th>
<th>Amount of data</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>NegBenchMark_O</td>
<td>1018.000</td>
<td>Negative</td>
</tr>
<tr>
<td>NegIndependent_O</td>
<td>258.000</td>
<td>Negative</td>
</tr>
<tr>
<td>PosBenchMark_O</td>
<td>520.000</td>
<td>Positive</td>
</tr>
<tr>
<td>PosIndependent_O</td>
<td>129.000</td>
<td>Positive</td>
</tr>
<tr>
<td>Total</td>
<td>1,925</td>
<td></td>
</tr>
</tbody>
</table>

The post-translational modification glycosylation in sequence O dataset available on the UniProt website is 1,925 sequences.

Table 2: Clean data

<table>
<thead>
<tr>
<th>Types of glycosylation</th>
<th>Amount of data</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>NegBenchMark_O</td>
<td>53</td>
<td>Negative</td>
</tr>
<tr>
<td>NegIndependent_O</td>
<td>37</td>
<td>Negative</td>
</tr>
<tr>
<td>PosBenchMark_O</td>
<td>76</td>
<td>Positif</td>
</tr>
<tr>
<td>PosIndependent_O</td>
<td>38</td>
<td>Positif</td>
</tr>
<tr>
<td>Total</td>
<td>204</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: O-sequence data

<table>
<thead>
<tr>
<th>ID</th>
<th>protein</th>
<th>Position</th>
<th>Types of glycosylation</th>
<th>Sequence</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q8TJD6</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>LLHQEHEMSVGS</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>Q8TJD6</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>PHQGQPHRS</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>E7EV10</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>VVCPFRRRSS</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>E7EV10</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>NTLMLADKHK</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>P13647</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>KQPNVIRSTP</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>P13647</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>SLYQFTTKRML</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>E9P03</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>LGFRTSLKFRN</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>E9P03</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>GSVKKLMM</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>E5R61</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>DSSVPGPSIV</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>E9P1U3</td>
<td>NegBench</td>
<td>Mark_O</td>
<td>QQFLPQPFPD</td>
<td>Not glycosylated</td>
<td></td>
</tr>
<tr>
<td>A3KN93</td>
<td>PosBench</td>
<td>Mark_O</td>
<td>TIVMTKTPP</td>
<td>Glycosylated</td>
<td></td>
</tr>
<tr>
<td>E7EN80</td>
<td>PosBench</td>
<td>Mark_O</td>
<td>VTTNRQTIITLTK</td>
<td>Glycosylated</td>
<td></td>
</tr>
<tr>
<td>O14639</td>
<td>PosBench</td>
<td>Mark_O</td>
<td>VSTPOTLACTQK</td>
<td>Glycosylated</td>
<td></td>
</tr>
</tbody>
</table>

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 feature () function. This feature consists of 21 amino acid sequences; then, it is stored in the example directory in 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 featurePseAAC() 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.
XGBoost employs feature selection by constructing a regression tree structure using a minimal redundancy maximum relevance (mRMR) technique. This approach involves selecting features with high correlation and minimal redundancy to generate the most relevant features related to the target variable. The primary objective 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 involved selecting 25 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.

Table 4: Feature extraction method contributes

<table>
<thead>
<tr>
<th>Feature extraction type</th>
<th>Total feature contribution</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sable</td>
<td>11</td>
<td>44</td>
</tr>
<tr>
<td>AAIIndex</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>CTD</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>PseAAC</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>Total</td>
<td>44</td>
<td>100</td>
</tr>
</tbody>
</table>

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 AAIIndex 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 MRRM feature selection to O-glycosylation was investigated in this research (Chien et al., 2020; Alkuhlani et al., 2022). The MRRM 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: This 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.

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:

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

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

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

The $y_i$ 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:

$$L(\theta) = -[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$  \hspace{1cm} (3)

where, $y_i$ 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 kth 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:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$  \hspace{1cm} (4)

$$Sn = \frac{TP}{TP+FN}$$  \hspace{1cm} (5)

$$Sp = \frac{TN}{TN+FP}$$  \hspace{1cm} (6)

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$  \hspace{1cm} (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

<table>
<thead>
<tr>
<th>Glycosylation sequence O</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Matthews correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td>99.27</td>
<td>98.13</td>
<td>100</td>
<td>98.64</td>
</tr>
<tr>
<td>Independent</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 6: Comparative research

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Method</th>
<th>Accuracy (%)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Li et al. (2015)</td>
<td>Random forest</td>
<td>95</td>
<td>2015</td>
</tr>
<tr>
<td>2020</td>
<td>Chien et al. (2020)</td>
<td>XGBoost</td>
<td>94.60</td>
<td>2020</td>
</tr>
<tr>
<td>2023</td>
<td>Alkuhlani et al. (2023)</td>
<td>XGBoost</td>
<td>77.86</td>
<td>2023</td>
</tr>
<tr>
<td>2023</td>
<td>Damayanti et al. (2023)</td>
<td>XGBoost</td>
<td>100</td>
<td>2023</td>
</tr>
</tbody>
</table>

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 future 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 glykosylation 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


