Integra: Journal of Integrated Mathematics and Computer Science
Journal Homepage: https://integrajimcs.com/index.php/integra/index
integra
Journal of Integrated Mathematics and Computer Science

Research Paper

Application of Random Forest Method Classification for Glycosylation in Lysine Protein Sequences

Silfia Fitriyana[1], Admi Syarif[1], Favorisen Rossyking Lumbanraja[1*], Mohammad Reza Faisal[2]

[1] *Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Lampung, Lampung, 35145, Indonesia*
[2] *Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat, Banjarmasin, 70123, Indonesia*
[*] **Corresponding author**: facvorisen.lumbanraja@fmipa.unila.ac.id

Keywords
Glycosylation, Lysine Protein, Extraction Features, Classification, Random Forest

Abstract
Grouping glycosylated lysine proteins into groups according to the type of glycosylation seen in the lysine protein sequence is known as glycosylation in the lysine protein sequence. In this work, the sensitivity, specificity, accuracy, and Matthew's correlation coefficient (MCC) of the random forest approach for classifying the glycosylation of lysine protein sequences were examined. With 214 positive and 406 negative data, the lysine protein dataset derived from benchmark data contains 620 total proteins with a protein length of 15 sequences. 90% of the dataset is used for training, while 10% is used for testing. Using the R package BioSeqClass version 1.44.0, feature extraction employed protein descriptors, specifically AA Index, CTD, and PseAAC, with a total of 60 features. The Random Forest classification algorithm was used to reprocess the results with *Mtry* values of 4, 8, and 16. The number of trees (*ntree*) was randomly set to 250, 500, 750, and 1000. The best results were achieved with a dataset split of 90% training data and 10% test data, using *Mtry* of 42 and 1000 trees, resulting in 89.97% sensitivity, 92.79% specificity, 80.76% MCC, and 90.42% accuracy. These results demonstrate that the combination of feature extraction and the Random Forest algorithm is effective in classifying lysine proteins.

Received: 16 April 2024, Accepted: 1 June 2024
https://doi.org/10.26554/integra.20241218

1. INTRODUCTION

Post-translational modifications (PTMs) involve the covalent alteration of amino acids within a protein's primary sequence, leading to a greater diversity of protein forms. These modifications play a critical role in regulating various biological processes, including protein localization within cells, protein stability, and the control of enzymatic functions. So far, over 90,000 distinct PTMs have been identified through large-scale biochemical and biophysical analyses, offering valuable insights into their significance in health and disease and highlighting proteomics as a powerful tool [1]. Depending on physiological needs and metabolic conditions, proteins may undergo various modifications such as phosphorylation, glycosylation, acetylation, ubiquitination, sulfation [2].

Glycosylation is one type of post-translational modification, a reaction that occurs between proteins and glucose at high concentrations, this reaction is also called the Maillard reaction [3]. The glycosylation reaction or Maillard reaction is a reaction between the amine group of the protein and the aldehyde group of glucose that can form reactive products, which can further modify the protein. This reaction is characterized by the occurrence of nonenzymatic browning between reducing sugars and reactive free amino acids from proteins [4].

Glycosylation will give rise to non-enzymatic reactions that generally occur post translational modification of proteins induced spontaneously by condensation of glucose derivatives and intermediate metabolic compounds with free amine groups on lysine or arginine residues. The first step of the Maillard reaction is the formation of schiff base inclusion complexes into proteins. This initial product of glycation undergoes reversible

rearrangement to form the Amadori inclusion complex product. Schiff bases and Amadori products then undergo rearrangement, oxidation, and or drying through chemical pathways to produce inclusion complexes that irreversibly become proteins, namely Advanced Glycation End Products (AGEs) [3, 5].

Based on these causal factors, a classifier is needed that is expected to help the medical world. This can be done with classification methods in data mining that match the information in helping to achieve the results of the classifier [6].

This study classifies data that is systematically organized into a group to find out an individual is in a particular group, in this study the classifier method used is random forest. A popular classification technique that blends the concepts of bagging classification trees and randomizing subset features, based on tree decisions and using aggregation ideas is Random Forest [7, 8, 9, 10]. The scheme is to build an ensemble predictor with a set of decision trees grown on data subspaces chosen at random. Despite their popularity and usefulness, random forests' statistical characteristics have not been well studied, and little is known about the algorithm's underlying mathematical strength. Particularly on high-dimensional data where the tree structure explicitly represents interactions among characteristics, random forest classifiers have been demonstrated to generate trees with minimal bias and low correlation between individual trees, resulting in effective classifiers. Some researches regarding the implementation of random forest in health/medicine include [11, 12, 13, 14] and many more.

In this study we will discuss about random forest method classification for glycosylation in lysine protein sequences.

2. METHODS

Protein and Amino Acids The primary macromolecule that organisms require is protein. The preferred function of protein is to synthesize new proteins according to the needs of the body. Protein contains 20 different kinds of amino acids. Each of the 20 different kinds of amino acids found in proteins has unique chemical characteristics [15].

The human body needs proteins, and the primary building blocks of proteins are amino acids, which are used by the body for metabolism. Amino acids, which are the building blocks of proteins and contain both carboxylic and amino groups, are important regulators of many gene expression-related activities. Amino acids are classified as either essential or non-essential. The body can synthesize non-essential amino acids, while essential amino acids cannot and must be acquired through diets high in protein. There are ten types of amino acids. Table 1 shows the 10 amino acids.

2.1 Glycosylation

As already stated in [4], Glycosylation reactions or Maillard reactions are reactions between the amine groups of proteins and the aldehyde groups of glucose that can form reactive products, which can further modify proteins. Figure 1 shows the main types of human glycosylation.

Table 1. Various Kinds of Amino Acids [16]

Essensial Amino Acids	Non-Essensial Amino Acids
Lysine	Cysteine
Methionine	Tyrosine
Valine	Serine
Tryptophan	Alanine
Isoleucine	Asparagines
Histidine	Aspartic Acid
Phenylalanine	Glutamic Acid
Threonine	Glycine
Leucine	Hydroxylysine
Arginine	Proline

2.2 Preprocessing

The preprocessing stage is data cleaning to eliminate repetitive data or the same data, using two procedures to remove sequences that have similarities greater than 10%, namely if a pair of proteins reaches a sequence identity percentage greater than the shortest sequence is discarded and if a pair of proteins has a sequence identity percentage that lies outside the shortest sequence range is discarded. Once the sequences have been removed the list contains only data that has no similarity. Data using the benchmark has a positive class dataset and negative dataset getting the number of protein sequences positive data 214 sequences and negative 406. Table 2 displays the data after preprocessing.

Table 2. Data After Preprocessing

Data set	Number of Protein Sequences
Positive	214
Negative	406

2.3 Feature Extraction

Feature Extraction is an important process in classifying an object to find the mapping of the original features into new features to increase the accuracy of the results of classifying an object [17]. This research uses the protein descriptor method to divide the problem into a set of subproblems to propose a more efficient algorithmic solution. The feature extraction used is protein descriptor using BioSeqClass package, and this study uses three kinds of feature extraction as follow.

2.3.1 CTD

A trait called Composition, Transition, and Distribution (CTD) is used to forecast where proteins should be targeted. There are 21 factors in each CTD, including hydrophobicity, normalized van der Waals volume, polarity, secondary structure, solvent accessibility, and chain length, which includes the first 25, 50, 75, and 100% of a protein sequence [18].

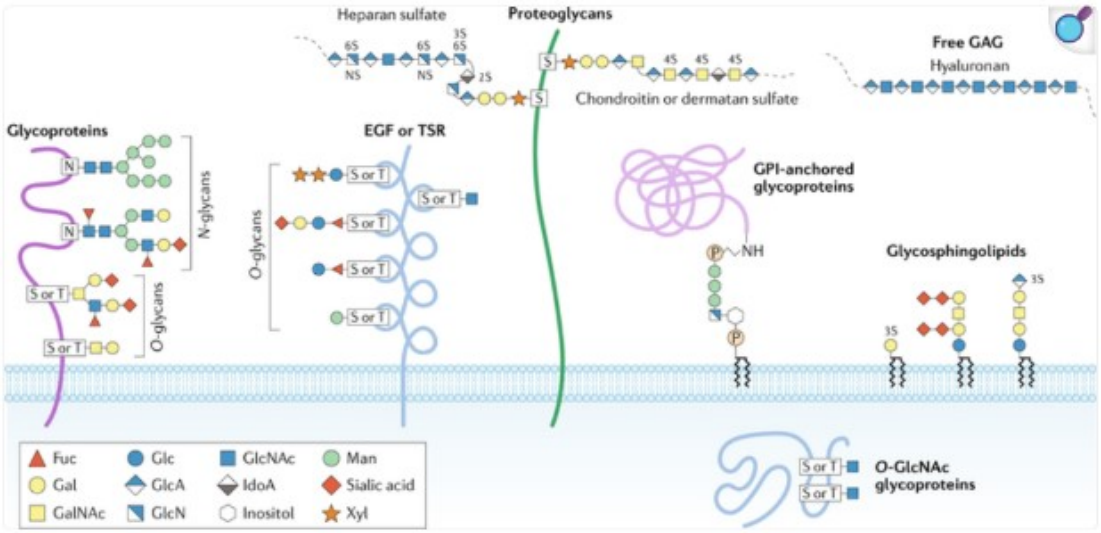

Figure 1. Main Types of Human Glycosylation [19]

2.3.2 AA Index

Numerous physicochemical and biological characteristics of amino acids and amino acid pairs are represented by the AA index, a database of numerical indexes. The three components of AAindex are as follows: AAindex1 for amino acid indexes with 20 numerical values. With its sequence length, the AA index has 15 variables [17].

Table 3. The Total of All Variables

Feature Extraction	Variable
CTD	21
AAIndex	15
PseAAC	24
The number of variables	60

2.3.3 Pseudo AminoAcid Composition

The technique known as pseudo amino acid composition retains sequence information while serving as a tool for creating biological sequences using vectors. Of the 24 variables in PseAAC, the first 20 features reveal the makeup of 20 amino acids [16].

The variables in this study amounted to 60 variables, namely CTD consisting of 21 variables, AAindex 15 variables, Pseudo Amino Acid Composite 24 variables. Feature extraction CTD, PseAAC, and AA index are stored in the form of CSV files. Table 3 displays the total of all variables.

2.4 Random Forest

Random forest is a classification and regression-based strategy with a decision tree aggregation step. This approach was selected due to its ability to handle very large volumes of training data, reduce errors, provide good classification accuracy, and work well with incomplete data [20].

Important features of the random forest implementation include using bootstrap sampling to construct prediction trees, having each decision tree use random predictors, and then integrating the results of all the decision trees using a majority vote for classification. In order to generate a forest with k trees, random feature extraction is used, where m explanatory variables are chosen at random where $m << p$. The optimal parser is then chosen based on (m) explanatory variables, and the process is repeated k times. After modelling the training data with the random forest package in R programming, the steps involved in developing a classification model using the random forest algorithm are completed.

Strength is the average or expected strength measure of a single tree's accuracy. *Mtry* is the number of distinct predictors that should be tried at each node, and nodesize is the terminal node's minimum size. Regression tree count is denoted by ntree. In order to obtain the best ntree, the tree should be strengthened and have a small correlation in order to be created till the mistake is minor. For the random forest approach to produce the best results, m predictor variables must be chosen at random and k trees must be created.

The use of the right m will result in OOB error depending on the correlation between trees and the strength of each tree in the random forest where increasing correlation can increase OOB error while increasing trees can decrease OOB error. OOB error is calculated from the proportion of misclassification of random forest results from all original data. The research conducted a random selection of ntree values, namely the default in the random forest in R programming 500 ntree and also the comparison values of 250, 750, and 1000.

The research testing step employs Cross Validation, a statistical evaluation method that compares learning algorithms by separating the data into two segments: one for learning or training the model and the other for validating the model. This cross

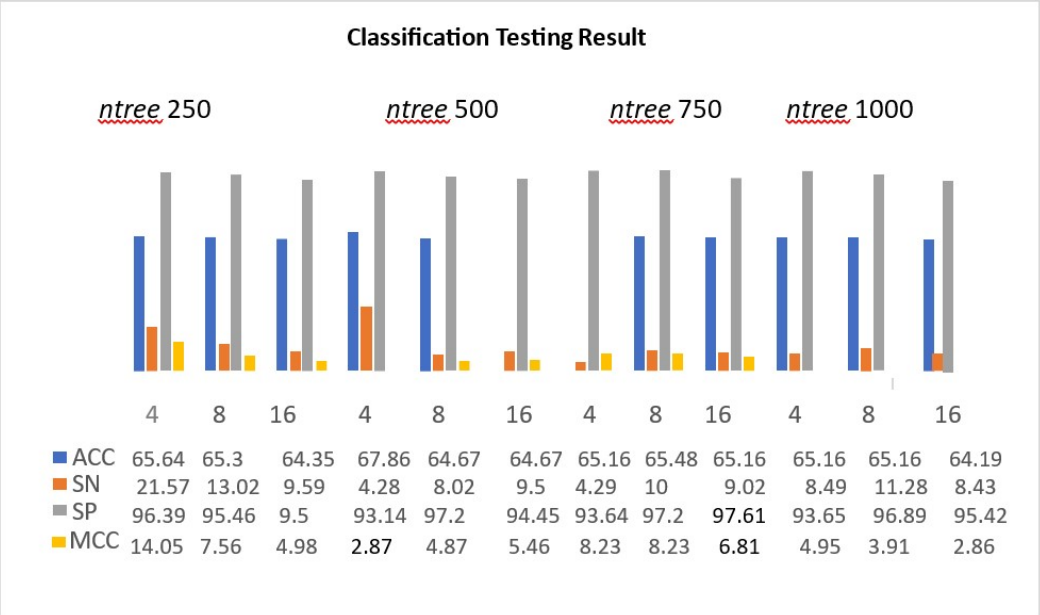

Figure 2. Lysine Protein Glycosylation Classification Testing Results

validation is used for statistical methods to get prediction results [21]. Assessment of prediction results is done using matrix evaluation, which is to calculate the performance of the classification model using the confusion matrix [22]. The parameters in the matrix can be seen in Equations 1, 2, 3 and 4, namely Accuracy, Sensitivity, Specificity, and Matthew Correlation Coefficient.

$$\text{ACC} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\text{SN} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{SP} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{4}$$

where:
TP (True Positive): The number of positive data accurately classified by the algorithm.
TN (True Negative): The number of negative data accurately classified by the algorithm.
FP (False Positive): the amount of positive data incorrectly classified by the algorithm.
FN (False Negative): the amount of negative data incorrectly classified by the algorithm.

3. RESULTS AND DISCUSSION

Glycosylation classification on the lysine protein sequence was carried out using the random forest method with 3 experiments, namely different feature extraction with a total of 60 variables and using parameters *Mtry* 4, 8, and 16 as well as ntree 250, 500, 750, and 1000. From the test results of glycosylation classification, lysine protein gets the highest accuracy at *Mtry* 4 and ntree 500 which is 67.8% while the smallest accuracy is at *ntree* 1000 and *Mtry* 16 at 64.1%.

Table 4. The Overall Out of Bag Error Results

	Ntree			
Mtry	250	500	750	1000
4	35.03%	34.85%	34.69%	34.74%
8	35.03%	34.83%	34.71%	34.62%
16	34.73%	34.71%	34.82%	34.40%

Testing using random forest there is a level of misclassification or out of bag error (OOB) which gets the results of each *Mtry* and ntree that for *Mtry* 4 the out of bag error continues to decrease and the lowest result is when ntree 750 with an out of bag error value of 34.69%, *Mtry* 8 shows out of bag error with the lowest result when ntree 1000 with an out of bag error value of 34.62%. While *Mtry* 16 shows the out of bag error with the lowest result at the time of ntree 1000 with an out of bag error value of 34.40%. Table 4 shows the out of bag error.

3.1 Comparative analysis with previous study

Previous research using the Support Vector Machine (SVM) by [6] method obtained greater accuracy than experiments using the random forest method, the random forest method has less good prediction performance results than SVM, namely getting the results of 67.86% accuracy, 4.21% sensitivity, 97.20% specificity, and 2.87% MCC. So that the random forest method has not been able to produce the best use in the classification of protein lysine glycosylation in the data. Table 5 shows the comparison results.

Table 5. Classification Performance Comparison

Method	ACC	SN	SP	MCC
SVM [6]	68.91	58.74	73.99	0.38
RF	67.86	4.21	97.20	2.87

4. CONCLUSIONS

From the results of glycosylation classification using random forest, it can be concluded that the data used (glycosylated protein data that is still in string format) was first converted using three extraction features consisting of AA index, composition, transition, and distribution (CTD), and pseudo amino acid composition (PseAAC). Glycosylation was classified using random forest with parameters *Mtry* 4, 8, and 16, and ntree 250, 500, 750, and 1000. The highest results were obtained using *ntree* 4 and *Mtry* 500, with an accuracy of 67.86%, sensitivity of 4.21%, specificity of 97.20%, and MCC of 2.87%. When comparing the results obtained by [6], which used the Support Vector Machine (SVM) method, it was found that SVM had better final values. Therefore, it can be said that in this case, random forest still has relatively low final results when influenced by feature extraction and the parameters used.

5. ACKNOWLEDGEMENT

The author would like to thank Software Engineering Laboratory, Universitas Lampung, and Department of Computer Science, Universitas Lambung Mangkurat for the support given.

REFERENCES

[1] M. Audagnotto and M. Dal Peraro. Protein post-translational modifications: In silico prediction tools and molecular modeling. *Computational and Structural Biotechnology Journal*, 15:307–319, 2017.

[2] D. Pascovici, J. X. Wu, M. J. McKay, C. Joseph, Z. Noor, K. Kamath, Y. Wu, S. Ranganathan, V. Gupta, and M. Mirzaei. Clinically relevant post-translational modification analyses—maturing workflows and bioinformatics tools. *International Journal of Molecular Sciences*, 20(1):16, 2019.

[3] N. Apriani, E. Suhartono, I. Z. Akbar, and U. L. Mangkurat. Korelasi kadar glukosa darah dengan kadar advanced oxidation protein products (aopp) tulang pada tikus putih model hiperglikemia. *JKM*, 11(1):48–55, 2011.

[4] E. Suhartono and B. Setiawan. Modifikasi protein akibat pembebanan glukosa dengan model reaksi glikosilasi nonenzimatik *In Vitro*. *Jurnal Ilmiah*, 08:40–47, 2008.

[5] M. He, X. Zhou, and X. Wang. Glycosylation: Mechanisms, biological functions and clinical implications. *Signal Transduction and Targeted Therapy*, 1:194, 2024.

[6] Y. Xu, L. Li, J. Ding, L.-Y. Wu, G. Mai, and F. Zhou. Glypseaac: Identifying protein lysine glycation through sequences. *Gene*, 602:1–7, 2017.

[7] S. R. Künzel, T. F. Saarinen, E. W. Liu, and J. S. Sekhon. Linear aggregation in tree-based estimators. *Journal of Computational and Graphical Statistics*, 31(3):917–934, 2022.

[8] C. D. Sutton. Classification and regression trees, bagging, and boosting. In C. R. Rao, E. J. Wegman, and J. L. Solka, editors, *Handbook of Statistics*, volume 24, pages 303–329. Elsevier, 2005.

[9] G. Biau and E. Scornet. A random forest guided tour. *Test*, 2016.

[10] C. Kern, T. Klausch, and F. Kreuter. Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1):73–93, 2019.

[11] F. Mbonyinshuti, J. Nkurunziza, J. Niyobuhungiro, and E. Kayitare. Application of random forest model to predict the demand of essential medicines for non-communicable diseases management in public health facilities. *Pan African Medical Journal*, 42:89, 2022.

[12] M. L. Wallace, L. Mentch, B. J. Wheeler, et al. Use and misuse of random forest variable importance metrics in medicine: Demonstrations through incident stroke prediction. *BMC Medical Research Methodology*, 23:144, 2023.

[13] W. Hong, Y. Lu, X. Zhou, S. Jin, J. Pan, Q. Lin, S. Yang, Z. Basharat, M. Zippi, and H. Goyal. Usefulness of random forest algorithm in predicting severe acute pancreatitis. *Frontiers in Cellular and Infection Microbiology*, 12:893294, 2022.

[14] P. Liu, Y. Liu, H. Liu, L. Xiong, C. Mei, and L. Yuan. A random forest algorithm for assessing risk factors associated with chronic kidney disease: Observational study. *Asian Pacific Island Nursing Journal*, 8, 2024.

[15] B. Alberts, A. Johnson, J. Lewis, et al. *Molecular Biology of the Cell*. Garland Science, New York, 4th edition, 2002.

[16] M. Akram, H. M. Asif, M. Uzair, N. Akhtar, A. Madni, S. M. Ali Shah, Z. U. Hasan, and A. Ullah. Amino acids: A review article. *Journal of Medicinal Plants Research*, 5(17):3997–4000, 2011.

[17] 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*, 38(8):10425–10436, 2011.

[18] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. Aaindex: Amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(SUPPL. 1):202–205, 2008.

[19] C. Reily, T. J. Stewart, M. B. Renfrow, and J. Novak. Glycosylation in health and disease. *Nature Reviews Nephrology*,

15:346–366, 2019.

[20] A. Primajaya and B. N. Sari. Random forest algorithm for prediction of precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining*, 1(1):27, 2018.

[21] S. Ohannessian. Historical background. In *Language in Zambia*, pages 271–291. 2017.

[22] M. Bekkar, H. K. Djemaa, and T. A. Alitouche. Evaluation measures for models assessment over imbalanced data sets. *Journal of Information Engineering and Applications*, 3(10):27–38, 2013.