Journal of Computer Science

Original Research Paper

Optimizing N-linked Glycosylation Site Prediction in Human Proteins with Ensemble Stacking and Cross-Validation

Mubina Malik and Jaimin Undavia

Department of Computer Science and Applications, CMPICA, CHARUSAT, Charotar University of Science and Technology (CHARUSAT), CHARUSAT Campus, Changa, India

Article history
Received: 21-07-2024
Revised: 30-08-2024
Accepted: 11-09-2024

Corresponding Author:
Mubina Malik
Department of Computer Science and Applications, CMPICA, CHARUSAT, Charotar University of Science and Technology (CHARUSAT), CHARUSAT Campus, Changa, India
Email: mubinamalik.mca@charusat.ac.in

Abstract: The most frequent post-translational modification of proteins in all territories is glycosylation which impacts many biological activities. The most significant and critical of these modifications is N-linked glycosylation which is associated with various human diseases including diabetes cancer Inflammation Alzheimers and atherosclerosis. This article illustrates recent advances in knowledge of biology that are eventually targeting the computer science sector. Moreover-identification of N-linked glycosylation helps to understand the biological system of humans and the mechanism of glycosylation. Machine learning techniques became very important for the N-linked glycosylation prediction from human protein because the experimental process is time-consuming and costly. This article proposes an ensemble machine learning approach for N-linked glycosylation prediction integrating updated and experimentally verified databases (UniProtKB dbPTM and nGlycositeAtlas) with an optimal window size of 21. MMSeq2 clustering with a threshold of 0.3 was employed to eliminate duplicate and similar protein sequences for improved dataset preparation. A total of 9040 features were extracted using various descriptors including sequence structural and physicochemical features. ANOVA F-score CHI2 and Mutual Information were used as ensemble feature selection techniques the combination of all these results generated 182 desirable features for the final model training. The model was then trained using cross-validation methods and ensemble stacking using four base classifiers: SVM LR XGBoost and RF. The prediction result demonstrates that ensemble stacking techniques with cross-validation give a more reliable and promising result than the individual base classifiers. Moreover, ensemble Stacking with cross-validation performs better than the individual classifier with an Accuracy of 99.99% Precision of 99.98% Recall of 100% AUC of 99.94% MCC of 99.96%, and F-score 99.99%.

Keywords: Machine Learning, Ensemble Stacking, XGBoost, Random Forest, SVM Cross Validation, Protein N-Linked Glycosylation

Introduction

Protein Glycosylation is one of the most important protein Post-Translational Modifications (PTM) in Eukarya Bacteria and Archaea (Moremen *et al*., 2012). The process of adding a sugar molecule to a protein-lipid or other organic molecule both inside and outside of the cell is known as protein glycosylation. Glycans are the carbohydrates that are connected to lipids and proteins during this process specifically to a specific residue that forms a glycosidic bond. The most complicated post-translational modification is glycosylation which is due to the greater number of enzyme steps required. Recent developments in artificial intelligence eliminate the limitations of experiment-based glycosylation detection. N-linked glycosylation O-linked glycosylation C-linked glycosylation S-linked glycosylation phoglycosylation and glypiation are some of the numerous forms of glycosylation (Ząbczyńska and Pochec, 2015). The most significant kind of all is N-linked glycosylation. Both the ER and the Golgi complex engage in N-linked glycosylation. The process of an oligosaccharide (glycan) being attached to the amide nitrogen of an asparagine (Asn) residue in a protein is known as N-linked glycosylation in biochemistry. N-linked glycosylation often takes place in the sequence N-X-S/T (N-asparagine S-serine T-threonine)

Science Publications

while it may also occur in N-X-C (C-cysteine) where X can be any amino acid other than proline (Gavel and Heijne, 1990). Furthermore, the wide variety of glycans linked to proteins and limiting the examination of certain glycosylation functions make it difficult to comprehend. It is difficult to experimentally characterize N-linked glycosides in glycoproteins since doing so is costly time-consuming and technically difficult. Our main objective is to Predict the N-linked glycosylation site using machine-learning techniques for accurate site prediction from human protein sequences considering the limitations of available techniques.

A number of publications were examined in order to identify the best approach as well as the dataset methodology and constraints of the current models for the prediction of human protein N-linked glycosylation. There are two approaches for predicting protein N-linked glycosylation (1) Protein sequences-based approach and (2) Protein structure-based approach. Additionally, there are two categories for protein sequence-based approaches: Residue level and sequence level. Additionally, it demonstrates that the optimal method for achieving high accuracy is one that is based on N-linked glycosylation sequences (Malik and Undavia, 2022). According to Birgit and Frank Eisenhaber, "glycosylation prediction is still not acceptable and sequence-based approach has low prediction rate because the number of glycosyltransferases is not investigated and indeed" (Eisenhaber and Eisenhaber, 2010). Manikandan Muthu Sechul Chun and others have emphasized the bioinformatics resources that are already accessible and concluded that there is a significant gap between the tools that are currently available and real-world applications. Even though many different glycosylation prediction methods have been created only 1% of them have been employed to study glycosylation in tumors (Muthu *et al*., 2020). Most of the machine learning and deep learning prediction models have assessed their performance at every N in protein sequences without the confirmation of N-X-S/T sequon according to several authors. Moreover, additional factors like disordered regions and physicochemical properties can be leveraged to provide more precise results (Chien *et al*., 2020; Pakhrin *et al*., 2021).

Literature Review

We have reviewed articles based on the protein N-linked glycosylation prediction of human protein to identify the research gap. Following are a few sequence-based feature prediction approaches for N-linked glycosylation: NetNGlyc uses Artificial Neural Network (ANN) (Gupta and Brunak, 2001) GPP uses RF (Hamby and Hirst 2008) EnsembleGly uses ensemble SVM (Caragea *et al*., 2007) GlocoPP uses RF (Chauhan *et al*., 2012) GlycoEP uses SVM (Chauhan *et al*., 2013) NGlycoGo uses XGBoost (Chien *et al*., 2020) GlycoMine uses RF (Li *et al*., 2015) and SprintGly uses DNN and SVM to construct large dataset (Taherzadeh *et al*., 2019). The two structure-based approach models that have been developed are GlycoMine^{struct}, which uses RF (Li *et al*., 2016), and NGlycPred, which uses RF (Chuang *et al*., 2012). However, some of the hybrid approaches using both sequence and structural features are N-GlycDE uses SVM (Pitti *et al*., 2019) DeepNGlycPred uses deep neural network (Pakhrin *et al*., 2021) PUStackNGly uses ensemble stacking (Alkuhlani *et al*., 2022) and LMNGlycPred uses Deep Learning Approach (Pakhrin *et al*., 2023) With the exception of a few machine learning approaches such as NetNGlyc (Gupta and Brunak, 2001) N-GlycDE (Pitti *et al*., 2019) PUStackNGly (Alkuhlani *et al*., 2022) and a few deep learning approaches such as DeepNGlycPred (Pakhrin *et al*., 2021) LMNGlycPred (Pakhrin *et al*., 2023) nearly every model listed in the aforementioned paragraphs evaluated their performance using residue N without confirming N-X-[S/T] motif to identify N-linked glycosylation. As a result, performance was overstated and generated high accuracy. In order to achieve comparable performance, the N-X-[S/T] consensus sequence needs to be taken into account for analysis. This literature review highlights the drawbacks of previous models of prediction despite their comparatively high accuracy. Overlooking the high accuracy of available models following motivating factors were found to have better accuracy in addressing the limitations.

- The presence of N-X-S/T sequon was not considered
- The datasets used in the previous model are relatively small
- Experimentally confirmed protein sequence data were not used
- Incomplete amino acid information for feature encoding was taken in an experimental study
- Feature selection techniques used without proper comparative study
- Inappropriate window size was used
- Disordered regions and physicochemical properties were ignored in many previous studies.

N-X-[S/T] consensus sequence should be considered to include in the analysis for attaining similar performances. However, since one-third to half of the consensus sequence is buried deep inside the protein and is inaccessible to the glycosylation enzyme its existence does not prove N-linked glycosylation (Schulz, 2012; Nita-Lazar *et al*., 2005). Therefore, a performance evaluation with only consensus sequences may result in false positives. Along with the sequence features a few predicted structural features are also used to enhance the accuracy of the prediction model. These include disordered residue Secondary Structure (SS) to check helix-strand-coil in a sequence and Accessible Surface Area (ASA) to check the accessibility of N residue. While utilizing structural features GlycoMine^{struct} (Li *et al*., 2016) SprintGly (Taherzadeh *et al*., 2019) evaluate the performance without verifying the N-X-[S/T] sequence. Recent developments in DeepNGlycPred (Pakhrin *et al*., 2021) with its focus on deep learning and the SVM-based approach in N-GlycDE (Pitti *et al*., 2019) have made key steps toward predicting N-linked glycosylation from protein sequences based on the N-X-[S/T] consensus sequence.

Proposed Model

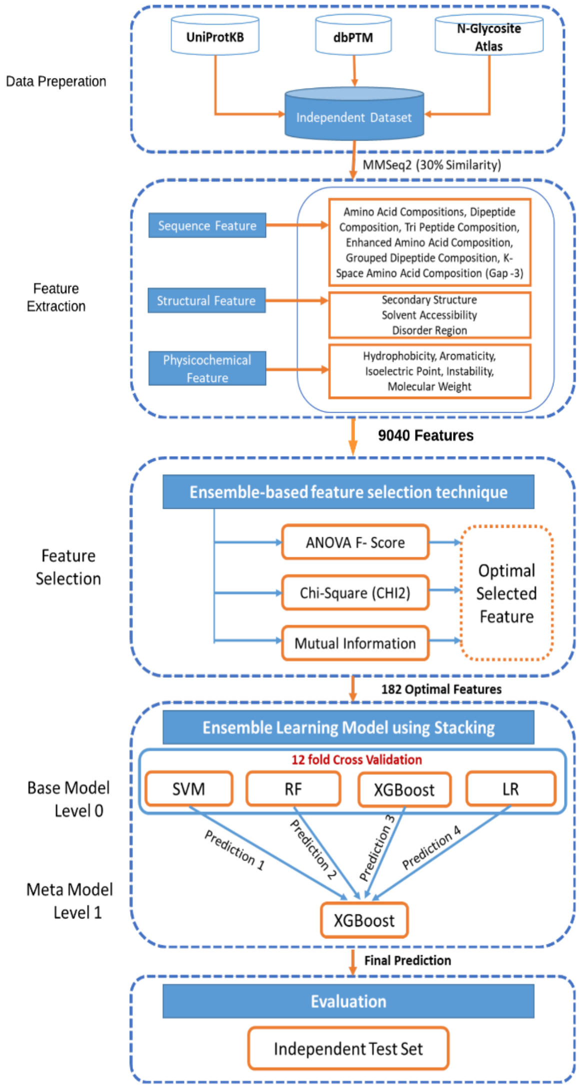

Fig. 1: Proposed model

The proposed approach illustrated in Fig. (1) uses a comprehensive approach to predict N-linked glycosylation sites in human proteins which involves five steps (1) Data preparation (2) Feature extraction (3) Feature selection (4) Ensemble Learning Model using Stacking and (5) Evaluation.

Step (1) data preparations: The model's data is combined from N-Glycosite Atlas dbPTM and UniProtKB into a single independent dataset. Sequence similarity is decreased using MMSeq2 ensuring a high-quality non-redundant dataset. Step (2) feature extraction: Sequence structural and physicochemical features are extracted. Step (3) Feature selection: To determine which features from the original set are the most informative an ensemble-based feature selection technique is applied using ANOVA F-score Chi-Square (CHI2) and mutual information. Step (4) Ensemble learning model using Stacking: The model combines many classifiers at two levels using a stacking ensemble learning technique. Support Vector Machine (SVM) Random Forest (RF) XG Boost and Logistic Regression (LR) are the base models (Level 1). The predictions of these models are combined using a meta-model (Level 2) in this case XG Boost classifier that was selected based on its performance and robustness. Step (6) evaluation: Evaluation of the model is performed on an independent dataset to ensure generalizability. The detailed description of each step is demonstrated in the below section.

Materials and Methods

Data Preparation and Preprocessing

Data is the main obstacle to protein analysis so we selected the dataset to include reliable and experimentally verified protein sequence data for N-linked glycosylation with the confirmed consensus sequence N-X-S/T. In this study, we have selected three datasets with a different version: UniProtKB (ver. 2022) (Caragea *et al*., 2007) dbPTM (ver. 2019) (Pakhrin *et al*., 2021), and n-GlycositeAtlas (ver. 2016) (Chauhan *et al*., 2012). These data are filtered to match the objective of the problem statement. For the UniProtKB Dataset (ver. 2022) we used the filter including reviewed and human data PTM as glycosylation with keyword N-linked to filter newly added and updated sequence information from UniProtKB this dataset consists of 8995 experimentally confirmed sequences from 3330 unique proteins. The dbPTM Dataset contains 481 distinct glycosides from 222 distinct proteins. The n-GlycositeAtlas Dataset includes 9260 unique proteins and 24383 proteins. For more research and forecasting these datasets were integrated. To obtain reliable and confirmed data these three datasets were combined before being preprocessed. It is necessary to select a proper window size for the protein sequence as the whole protein sequence cannot be processed at a time to train the model. We have studied a few research articles and important features such as protein structure hydrophobicity for N-linked glycosylation prediction and have selected the window size 21 for further model implementation (Malik and Undavia, 2022) (Chauhan *et al*., 2013) (Li *et al*., 2015). We have further studied and divided the proteins that had modifications at the N-terminal the C-terminal and everywhere in between to ensure the accuracy of N-linked glycosylation prediction at various locations. After the combining dataset, we got 33858 N-linked glycosylation sites with a total of 11255 unique proteins. The statistics of the data that we have selected for the model are described in Table (1).

Classification and clustering are two main tasks in machine learning research. Sequence-clustering algorithms that can analyze massive volumes of sequencing data are becoming more and more necessary as Next-Generation Sequencing (NGS) technology advances. It is required to cluster the protein sequence and eliminate duplicate or similar identity protein sequences according to a threshold to increase the efficiency of sequence analysis and reduce sequence redundancy from the chosen datasets. A number of tools are commonly known for this purpose including MMSeqs2-Fast and sensitive clustering for large datasets (Hauser *et al*., 2016; Steinegger and Söding, 2017) USEARCH-Versatile clustering and searching (Edgar, 2010) CD-HIT Sequence similarity-based clustering (Fu *et al*., 2012; Li and Godzik, 2006) KClust Kmer similarity-based clustering. We have chosen MMSeqs2 for clustering and the search performed rapidly and sensitively enough to find sequence matches down to 30% residue-wise sequence identity and fulfilled by MMseqs2. According to Steinegger and Söding 2017 in aligning sequences, MMseqs2 is more sensitive and best in accuracy compared to USEARCH and CD-HIT. Their study showed that MMseqs2 was able to find some kinds of homologous sequences not detected by other programs. Thus, it has enabled more comprehensive studies to be performed (Steinegger and Söding, 2017). Therefore, the MMseqs2 is a protein sequence analysis tool economically viable since it is open-source free software. On the other hand, alternatives such as USEARCH would be an issue for some researchers since they are propriety programs requiring licenses. Table (2). Describe the details of the various feature comparisons for the clustering algorithms and the results clearly show that MMSeqs2 is outperformed as compared to the rest of the algorithms.

For dataset preparation, we used MMSeq2 techniques to remove the redundancy with a 0.3 threshold and as a result, we got 30225 protein sequence clusters as positive samples to train and test the model. Also, negative samples were collected from the nGlycDE (Pitti *et al*., 2019). We have considered 3964 negative protein sequences for human protein which have N at the middle position in protein sequences from the N-GlycDE dataset. The alignment sequence logo and the position-specific scoring matrix depict the positive and negative datasets shown in Figs. (2-3) respectively.

From the aligned sequence logos, it is clearly identified that the data that is considered for the prediction confirms the consensus sequence N-X-S/T as positive samples and N at position 11 in the negative samples. Moreover, previous work indicates that N-X-S/T is the simplest substrate for Oligo Saccharyl Transferase (OST) to transfer N-linked glycans; nevertheless, it has been demonstrated that the improved sequons F-X-N-X-T and F-X-X-N-X-T which include the adjacent aromatic amino acid phenylalanine (F) ensure successful N-glycosylation (Chen *et al*., 2013; Culyba *et al*., 2011). After confirming the accuracy of N-glycosylation sites dataset is split into train and test using machine learning techniques with 80% (24180 positive 3171 negative glycosites) and 20% (6045 positive 793 negative glycosites) respectively with positive and negative data which can be used for the prediction model.

Table 1: Protein N-linked glycosylation dataset insights

Dataset	Total Number of n-linked glycosite	C-terminal (End) Modified	Between Modified	N-terminal (Begin) Modified
dbPTM	481	5	476	0
n-GlycositeAtlas	24382	258	24124	0
UniProt	8995	76	8378	541
Total		339	32978	541

Table 2: Feature comparison of clustering techniques for protein sequences

Feature	MMSeqs2	USEARCH	CD-HIT	KClust
Speed	Very High	High	High	High
Accuracy	High	High	High	Moderate
Memory usage	Moderate	Moderate	Low	Low
Scalability	Yes	Yes	Yes	Yes
Handles large datasets	Yes	Yes	Yes	Yes
Supports parallel processing	Yes	Yes	Yes	Yes
Easy to use	Yes	Yes	Yes	No
Open-source	Yes	No	Yes	Yes

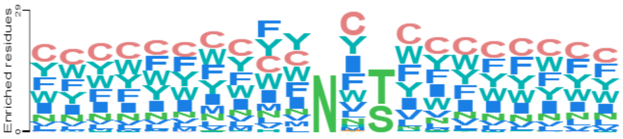

Fig. 2: Sequence logo of positive protein sequences

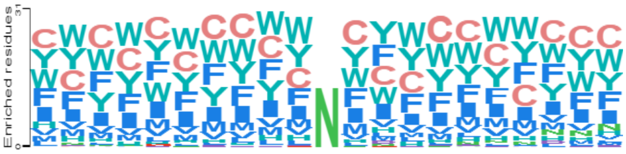

Fig. 3: Sequence logo of negative protein sequences

Feature Extraction

Dimensionality reduction is the process of reducing the number of variables or features in review. Dimensionality reduction can be divided into two subcategories which are Feature Selection and Feature Extraction. Feature extraction and selection are critical steps in developing accurate and interpretable prediction models. The 20 letters that make up a protein sequence are called amino acids. To predict n-linked glycosylation using machine learning or deep learning the amino acid sequence must be converted to binary representation. Protein sequence data can be encoded by three encoding methods which convert protein sequences into numeric features (Malik and Undavia, 2022). Feature extraction is the process of extracting a set of valuable features from raw data that may be used for prediction. Using a variety of sequence-based features that we extract from protein sequences; our method captures the properties of N-linked glycosylation sites. Based on the techniques listed in Table (3) protein features are extracted. Protein sequence-based features structural features and physicochemical properties are all extracted using Bio-python and iLearnPlus (Chen *et al*., 2021; Cock *et al*., 2009).

For N-linked glycosylation, three types of features are important to apply machine learning techniques for precise prediction. (1) Sequence-based feature (2) Structural features and (3) Physicochemical properties. As N-linked glycosylation identification is based on the Amino Acid Property's Secondary structure folding and various physicochemical properties We have encoded a total of 9040 features which include 8846 sequence-based features 127 structural features and 67 Physicochemical features mentioned in Table (3). Sequence-based features include the various forms of amino acid compositions of each possible amino acid and its groups such as A N S NS TP EAF etc. Structural features include secondary structures such as alpha helix beta-sheet and coil at each position. Solvent Accessibility represents specific amino acids at each position that is accessible to the surface or not which include Exposed Buried and Intermediate state.

Table 3: Protein sequence encoding techniques

Binary encoding method	One-hot encoding method (20-bit), PAM Metrics (6-bit), PAM Metrics (5-bit)
Substitution Metrix	Position independent – PAM, BLOSUM.
	Position dependent – PSSM, PSI-BLAST
Physicochemical property encoding	VHSE Scale

Feature Selection

To determine which features are most important to use in machine learning algorithms the feature selection approach is considered. Feature selection techniques are used to reduce the number of input variables by eliminating unnecessary or redundant features and limiting the set of features down to those that are most significant to the machine learning model. Techniques for feature selection not only eliminate unnecessary features but also optimize the model reduce variance and shorten training times. When using feature selection techniques, it is crucial to consider the nature of the problem statement for each input variable in the dataset. Both continuous (floating point and integer) and categorical (Boolean ordinal and nominal) variables are commonly employed as inputs. We obtained both types of variables as features after feature encoding. Physicochemical qualities molecular weight and amino acid compositions are examples of continuous numerical data whereas solvent accessibility and secondary structure are examples of categorical data.

Feature Selection-Collectively or Individually: Feature selection techniques can be applied either to all features together or separately to individual feature types depending on the context and objectives. Our main objective is to develop a reliable and efficient prediction model for N-linked glycosylation in human proteins using protein sequence structure and physicochemical properties. The relationship between protein sequence structure and physicochemical properties is deeply interconnected forms the foundation of protein biology and impacts the N-linked glycosylation site (Ramírez and Locher, 2023). The way a protein folds into its 3D structure is driven by interactions between the amino acids in the sequence such as hydrogen bonding hydrophobic interactions and disulfide bonds. Slight alterations in the sequence can lead to misfolding and potentially result in alteration of the glycosylation process. The physicochemical properties of a protein are derived from its amino acid sequence and influence how the protein behaves in different environments.

For N-linked glycosylation where complex interactions between sequence structure and physicochemical properties determine the glycosylation sites applying feature selection to all features together is particularly effective. This method is superior to applying feature selection to each type separately because it captures the complex interactions between sequence structure and physicochemical properties providing a more comprehensive and accurate prediction model. (Pakhrin *et al*., 2023; Ramírez and Locher, 2023).

Techniques for Feature Selection: The optimal feature selection methods are chosen according to the input and output variables. Targeting the problem statement and encoded feature we have selected ensemble feature selection techniques using ANOVA

F-score (Suresh and Naidu, 2022; Hasan and Hasan, 2020) CHI2 (Liu and Setiono, 1995; Zhang *et al*., 2020) and Mutual Information (Jorge and Pablo, 2014) for the feature selection. Table (4) Describes the individual feature selection technique with input and output variables. Mutual Information is applicable to both continuous and categorical data while ANOVA F-Score is best suited for continuous data and CHI2 for categorical data. The ratio of within-group variability to between-group variability is compared using the ANOVA f-score and is expressed in Eq. (1):

$$f = \frac{\sum_{i=0}^{k} n_i(\bar{x}_i - \bar{x})^2/(k-1)}{\sum_{i=0}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2/(N-k)} \quad (1)$$

where, k is the number of groups n_i is the total of observations in the i^{th} group x_i is the mean of the i^{th} group $\bar{x}$ is the overall mean of the observed group and N is the total number of observations.

Based on the variations between observed and expected frequencies for every cell in the contingency table the chi-square statistic (χ^2) is computed. It is stated in Eq. (2):

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

where, O_i is the observed frequency and E_i is the expected frequency i.

Mutual information is frequently used in feature selection to assess how closely a feature relates to the target variable in classification tasks. The mutual information $I(X;\ Y)$ between a feature X and the target variable Y is defined in Eq. (3):

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(xy)\ log\left(\frac{p(xy)}{p(x)p(y)}\right) \quad (3)$$

Table 4: Feature selection techniques

Method	Input VARIABLE	Output variable	Purpose
ANOVA F-Score	Numerical	Categorical	Statistically significant differences in means between two or more groups.
CHI2	Categorical	Categorical	Assess the association between two categorical variables.
Mutual Information	Numerical or Categorical	Categorical	Measures the relevance or association between a feature and the target variable

Table 5: Feature selection outcome of ensemble method using ANOVA F-score, CHI2, mutual information

Description Group	Descriptor	No. of Feature	No. of Selected Feature
Sequence based features	Amino Acid Composition	20	11
	Dipeptide Composition	400	21
	Tripeptide Composition	8000	64
	CKSAAP (Gap 3)	400	11
	Enhanced Amino Acid Composition	21	10
	Grouped Amino Acid Composition	5	4
Structural feature	Secondary Structure	63	32
	Solvent Accessibility	63	14
	Disorder Region	1	1
	Hydrophobicity	63	12
Physicochemical properties	Molecular Weight	4	2
	Aromaticity		
	Isoelectric Point		
	Instability Index		
3	14	9040	182

Ensemble Model Approach Classifiers

The use of ensemble techniques in machine learning approaches has gained popularity recently as seen in multiple approaches. Wang *et al.* have implemented an ensemble two-stage model for cancer survival prediction (Wang *et al*., 2019) Suraj Gattani *et. al.* used a two-stage ensemble approach for the prediction of protein-carbohydrate binding (Gattani *et al*., 2019) Xiao *et al*. proposes an ensemble learning method to improve traffic incident detection using SVM and KNN (Xiao, 2019; Alkuhlani *et al*., 2022) applied ensemble method to predict positive unlabelled N-linked glycosylation prediction using Stacking techniques (Pitti *et al*., 2019). To integrate the result various base predictors are assembled using ensemble techniques called stacking. It increases the model's capacity and scalability which is unachievable with just one predictor. Ensemble approaches are based on base models and meta models where base models are machine learning classifiers that operate independently to produce predictions which are then combined to produce an integrated prediction result. The final machine learning classifier or meta-model generates the final prediction result by utilizing the input as an integrated prediction result.

In this study, we have used Support Vector Machine (SVM) developed by Cortes and Vapnik (1995) Random Forest (RF) (Breiman, 2001; Hamby and Hirst, 2008) Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) and Logistic Regression (LR) (Tolles and Meurer, 2016) machine learning predictive model for evaluating the performance of the proposed model. For the selection of a classifier, we have reviewed research articles on machine learning classifiers as well as the previous machine learning predictors that were mentioned in the Introduction for N-linked glycosylation.

SVM: Because SVM clearly defines decision boundaries it excels at binary classification. They are particularly helpful when working with datasets that have a lot of features since they are less prone to overfitting than other techniques, especially in high-dimensional data situations. The SVM polynomial kernel uses this high-dimensional feature space to linearly distinguish the glycosylated and non-glycosylated classes.

LR: We have decided to use the Logistic Regression (LR) classifier as our regression technique of choice. Using a sigmoid function the linear regression model converts the continuous value output of the linear regression function into a categorical value output. Though its intrinsic linearity logistic regression can be expanded to accommodate non-linear correlations by using interaction terms or polynomial features.

XGBoost: An ensemble learning technique called XGBoost aggregates many decision trees to get a single prediction. The primary goal is to build new trees that rectify the errors of the ones that already exist.

RF: By building an ensemble of decision trees Random Forest generates a class that is the mean of the classes that each individual tree predicted. Multiple decision trees are constructed and high dimensional features are handled by Random Forest. Ensemble of decision trees trained on bootstrapped data with randomly selected features. In regards to the dataset features hold numerous amino acids and their combinations interact in protein sequence.

Ensemble Stacking Approach with Cross Validation

Stacking is an enhanced ensemble learning technique that is intended to enhance the predictive performance of machine learning models by combining several base learners. The main concept is to minimize each model's unique weaknesses while utilizing its strengths (Dey and Mathur, 2023). Base classifiers (Level 1) such as Support Vector Machine Logistic Regression XGBoost and Random Forest are initially trained on the same dataset in a stacking framework. Predictions are generated by each base classifier and used as input features by a meta-classifier XGBoost (Level 2). The meta-classifier is trained to produce final predictions based on these inputs. By using a two-level strategy, the meta-classifier can learn how to optimally integrate the base classifiers predictions hence enhancing overall accuracy and capturing complex patterns. But only the stacking approach might lead to overfitting if any one of the base models over-fittings and is sensitive to data. There are two techniques that can help to solve meta learner overfitting problem (1) Hold out method which is commonly known as blending (Wu *et al*., 2021), and 2 k-fold cross-validation (Nti *et al*., 2021). We have selected k-fold cross-validation techniques with an ensemble stacking method to ensure the data will not overfit to the selected classifier. Cross-validation is the method that divides the dataset into training and testing data using k-fold cross-validation. To assess the cross-validation performance k experiments were carried out after the dataset was divided at random. The kth partition data is used to average accuracy data over the experiments which can be used for testing and training. To ensure that the training set remains unseen during model training cross-validation is done to each base model to create out-of-fold predictions instead of using their predictions directly. Then a meta-model also known as a meta-learner uses the new dataset which is the result from the base learner and learns to give the prediction on the test dataset. To prevent overfitting this meta-model is trained on a different fold of the data which improves the generalization and robustness of the ensemble model.

Figure (4) represents the entire architecture of stacking with cross-validation. In this architecture, the training set is divided into k-folds at the start of the process. The training set is made up of the remaining folds in each fold, with one portion designated as the validation fold. Level-1 predictions (P1-P4) are generated by training each classifier (C1-C4) on the training folds and testing it on the validation fold. To guarantee that every instance in the dataset is used for both training and validation, this process is repeated over all folds. This helps to produce reliable level-1 predictions that are less likely to overfit.

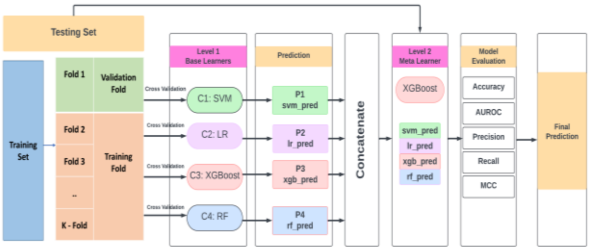

Fig. 4: Stacking ensemble based on cross-validation

Evaluation Criteria

Model performance was evaluated using several criteria to judge the quality of the model. Rainio *et al.* (2024) described evaluation metrics and statistical test for binary classification problem in machine learning. The most common binary classification indicators to express correctly classified instances are (1) Accuracy (2) Specificity (3) Sensitivity commonly known as Recall and (4) precision. There are other evaluation criteria F1-score that depends on all values of confusion metrics and gives the same weight to accuracies within the positive and negative instances and Matthews' Correlation Coefficient (MCC) measures the correlation between the real and the predicted values of instances Cohen's kappa which is applied to measure the agreement between the predicted and actual classes and ROC curve obtained by plotting sensitivity against the false positive rate (Rainio *et al*., 2024). The most significant evaluation indicator for machine learning models is Accuracy (ACC) Precision Recall MCC and F-Score which are evaluated on TP FP FN and TN stand for true positives false positives false negatives and true negatives respectively. Evaluation criteria for possible value measurement are defined below:

$$Accuracy \in [0,1],\ Specificity \in [0,1], Precision \in [0,1], Sensitivity = Recall \in [0,1], F1 \in [0,1], MCC \in [-1,1],\ ROC \in [0,1] \quad (4)$$

Results and Discussion

Hyperparameter Optimization for Classifier

For tuning the parameter of the selected classifier, we have implemented Grid Search techniques with the 182 selected features on the selected classifier. The details of the parameters are mentioned in Table (6).

Table 6: Hyperparameter Tunning for SVM, LR, XGBoost and RF

SVM	Kernel	Linear, poly, rbf, sigmoid	Poly
	C	0.1,1,10	1
	gamma	Scale, Auto	Scale
LR	C	0.1,1,10	10
	max_Iter	1000	1000
	Solver	liblinear, lbfgs	liblinear
RF	n-estimator	50,100,150	150
	max_depth	None,10,20	10
	min_sample_split	2,5,10	10
XGBoost	learning_rate	0.1,0.2,0.3	0.3
	max_depth	3,4,5	3
	n_estimators	50,100,150	50

Performance of Base Model Classifiers on Cross-Validation

To improve the predictive performance of our machine learning models we utilized a complex model selection process in this study that included stacking and cross-validation techniques. First, we put into practice four base models: Support Vector Machine (SVM) Logistic Regression (LR) XGBoost, and Random Forest (RF). Each base model was thoroughly trained and validated using cross-validation to guarantee reliable and objective performance estimates. Next, we used stacking techniques to combine the strengths of these various base models. Specifically, predictions from the SVM LR XGBoost and RF models were used as input features for a meta-model. For the selection of the meta-model, we have performed cross-validation on the individual base model and the base model with the highest performance was selected as the meta-model for the ensemble approach.

To improve the predictive performance of our machine learning models we utilized a complex model selection process in this study that included stacking and cross-validation techniques. First, we put into practice four base models: Support Vector Machine (SVM) Logistic Regression (LR) XGBoost, and Random Forest (RF). Each base model was thoroughly trained and validated using cross-validation to guarantee reliable and objective performance estimates. Next, we used stacking techniques to combine the strengths of these various base models. Specifically, predictions from the SVM LR XGBoost and RF models were used as input features for a meta-model. For the selection of the meta-model, we have performed cross-validation on the individual base model and the base model with the highest performance was selected as the meta-model for the ensemble approach.

The k-fold cross-validation with k = 2-k = 20 with the gap 2 result in Table (7) shows that Initial folds (0 and 2) showed significant variability but from fold 4 onwards the classifiers' performance became more consistent and reliable. In fold 8-20 accuracy stabilized with minor variations and all classifiers consistently performed well, especially from fold 12 onwards where accuracy remained high and stable. The findings show that XGBoost performs consistently across several folds achieving the highest mean accuracy (0.96) with a low standard deviation (0.04). SVM Random Forest and Logistic Regression all indicate robust performance with mean accuracies of 0.94 0.95 and 0.93 respectively. The models that indicate the least variability are Random Forest and XGBoost while the standard deviation values indicate that all models offer consistent predictions.

To analyze the central tendency and variability of model accuracies and to check the stability of the model we have plotted the Gaussian distribution of the model accuracy which is defined in Fig. (5).

Table 7: Accuracy of SVM, LR, XGBoost, and RF Over k-Fold cross-validation with mean accuracy and standard deviation

	Accuracy			
k- Fold	SVM	LR	XGBoost	RF
0	0.99	0.98	1	0.99
2	0.81	0.80	0.83	0.83
4	0.89	0.86	0.94	0.92
6	0.92	0.90	0.96	0.94
8	0.94	0.92	0.95	0.95
10	0.95	0.93	0.97	0.95
12	0.96	0.97	0.98	0.96
14	0.96	0.97	0.98	0.96
16	0.96	0.98	0.98	0.97
18	0.96	0.98	0.98	0.97
20	0.96	0.98	0.98	0.97
Mean Accuracy	0.94	0.93	0.96	0.95
Standard Deviation	0.05	0.06	0.04	0.04

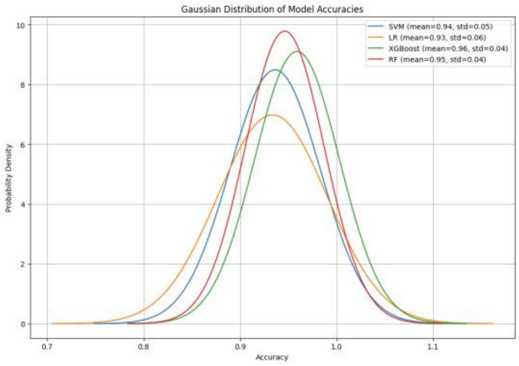

Fig. 5: Accuracy density Plots for SVM, LR, XGBoost and RFclassifier

Figure (5) Accuracy Density Plots for SVM LR, XGBoost and RF classifier. From the Gaussian distribution plot, we observed accuracies of all four models satisfy Gaussian distributions meaning that the performance of each model is distributed normally around its mean. The peaks of these distributions fit the accuracy levels that are most observed during the testing of models. From the cross-validation result and Gaussian distribution plot, we have selected XGBoost as a meta-model (Level 1). XGBoost was chosen as the meta-model due to its proven ability to handle complex patterns and interactions effectively.

Performance Metrics Comparison: Base Classifiers vs. Ensemble Stacking with Cross-Validation

When different machine learning models were compared to predict protein N-linked glycosylation it was shown that the Ensemble Stacking model with Cross-Validation (CV) performed significantly better than the other individual classifiers. This model is the most robust and reliable choice for this prediction task with nearly perfect performance metrics across all evaluation criteria. Most significantly the Ensemble Stacking model achieved an outstanding 0.9999 accuracy demonstrating nearly perfect prediction ability. The model demonstrated the capacity to accurately detect all real positive glycosylation sites without producing any false negatives as evidenced by its specificity recall and sensitivity both reaching 1.0000. Additionally, the 0.9987 of the models indicates its ability to minimize false positives and ensure a high level of confidence in discriminating between glycosylated and non-glycosylated locations. Its ability to predict positive instances with accuracy is further supported by its precision of 0.9998. The model performs even better as proven by the ROC AUC value of 0.9994 which shows nearly perfect classification abilities between the two groups. With a Matthews Correlation Coefficient (MCC) of 0.9996 the model can manage class imbalances exceptionally well and produce accurate predictions. Furthermore, the model's balanced performance in terms of recall and precision is confirmed by the F-Score of 0.9999. All the base models' kappa scores are above 0.8 which denotes strong to extremely strong agreement between the predicted and real labels. This implies that every base model performs satisfactorily on the dataset on its own. The stacking model achieves an exceptional performance by combining the advantages of the different base models as evidenced by its nearly perfect kappa score. This suggests that the stacking method successfully captures and improves the basic models' predictive ability.

Together the results shown in Table (8) indicate that the Ensemble Stacking model with Cross-validation is the optimal solution to predicting protein N-linked glycosylation from human protein sequences. It provides surpassed accuracy robustness and reliability when compared to other models that were assessed including SVM Logistic Regression XGBoost and Random Forest. Because of its exceptional performance, it is a very useful tool for both researchers and practitioners who study glycosylation prediction.

Table 8: Base classifier and ensemble stacking performance metrics for N-linked glycosylation prediction

Model	SVM	LR	XGBoost	RF	Ensemble Stacking With CV
ACC	0.96	0.97	0.98	0.96	0.9999
Specificity	0.96	0.91	0.98	0.97	0.9987
Pre	0.99	0.99	0.99	0.99	0.9998
Recall	0.96	0.98	0.98	0.96	1.0000
ROC AUC	0.96	0.94	0.98	0.97	0.9994
MCC	0.83	0.86	0.93	0.86	0.9996
F - Score	0.98	0.98	0.99	0.98	0.9999
Cohen's Kappa	0.82	0.86	0.93	0.86	0.9996

A complete list of abbreviations is listed in Appendix I.

Appendix I

Sr. No.	Abbreviation	Description
1	PTM	Post-translation modification
2	MMSeq2	Many-against-many sequence searching
3	UniProtKB	UniProt knowledgebase
4	ANOVA	Analysis of variance
5	SVM	Support vector machine
6		RF random forest
7	LR	Logistic regression
8	XGBoost	Extreme gradient boosting
9	CV	Cross-validation
10	ACC	Accuracy
11	MCC	Matthews’correlation coefficient
12	ROC	Receiver operating characteristic
13	AUC	Area under the curve

Conclusion

In this study, we developed a comprehensive approach to predict protein N-linked glycosylation from human protein sequences. The methodology integrates multiple stages starting from data preparation and feature extraction to feature selection and ensemble learning model construction. We obtained our data from reliable protein databases such as UniProtKB dbPTM and N-Glycosite Atlas considering the diversity of the dataset. To ensure a broad and representative collection of protein sequences we carefully selected an independent dataset using MMseq2 and a 30% similarity threshold. To effectively predict N-linked glycosylation sites it is essential to incorporate structural and physicochemical properties alongside sequencing features. These features offer complementary information about the protein's location confirmation and chemical composition. These characteristics improve the model's capacity to capture complex interactions and amino acid dependencies that affect glycosylation beyond simple sequence patterns. The feature selection step is used to enhance the model's efficiency and predictive power. An ensemble feature selection technique was employed using ANOVA f-score CHI2 and Mutual Information which have different efficiency to select features. As a result, we got 182 optimal features from 9040 features that were extracted. The model was constructed using a stacking technique that combines the results of SVM Logistic Regression XGBoost and Random Forest classifiers with 12-fold cross-validation. We experimented with folds ranging from 1-20 with a gap of 2 and selected 12 folds for optimal performance. XGBoost was chosen as the meta-model due to its superior performance in our experiments. When compared to individual models the suggested ensemble stacking approach performed better obtaining almost perfect metrics for all evaluation criteria with accuracy 99.99% Precision 99.98% Recall 100% ROC AUC 99.94% MCC 99.96% and F-Score 99.99%. With this method, glycosylation site prediction is made robust and accurate with implications for proteomics and bioinformatics research. Abnormality in N-linked glycosylation is detected in many diseases such as diabetes cancer Inflammation and Alzheimer's disease. Future work will focus on the clinical diagnosis of such diseases and consequently their drug development.

Acknowledgment

We extend our heartfelt thanks to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided, which have allowed us to share our findings with a broader audience. We sincerely appreciate the efforts of the editorial team in reviewing and refining our work. Additionally, we acknowledge the valuable dataset resources provided by UniProtKB dbPTM and nGlycositeAtlas, which were instrumental in the success of our research. We are thankful for the opportunity to contribute to the scientific community through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Author Contributions

Mubina Malik: Instrumental in this research, responsible for conceptualizing and designing the study, coordinating the entire research process, and integrating data from sources such as UniProtKB, dbPTM, and N-Glycosite Atlas. She identified the classifier, performed cross-validation to develop the ensemble stacking model, conducted cross-validation experiments, and drafted and revised the manuscript.

Jaimin Undavia: As the supervisor, Dr. Undavia provided expert guidance, methodological oversight, and critical feedback throughout the research process.

Ethics

This article is original and unpublished. The corresponding author confirms that authors have thoroughly reviewed and approved the manuscript and there are no ethical concerns associated with the research.

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

Schulz B.L, (2012). Beyond the Sequon: Sites of N-Glycosylation. *InTech*, 21–40.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

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), 1–13. https://doi.org/10.1186/1471-2105-8-438

Chauhan, J. S., Bhat, A. H., Raghava, G. P. S., & Rao, A. (2012). GlycoPP: A Webserver for Prediction of N- and O-Glycosites in Prokaryotic Protein Sequences. *PLoS ONE*, *7*(7), e40155. https://doi.org/10.1371/journal.pone.0040155

Chauhan, J. S., Rao, A., & Raghava, G. P. S. (2013). In silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences. *PLoS ONE*, *8*(6), e67008. https://doi.org/10.1371/journal.pone.0067008

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, W., Enck, S., Price, J. L., Powers, D. L., Powers, E. T., Wong, C.-H., Dyson, H. J., & Kelly, J. W. (2013). Structural and Energetic Basis of Carbohydrate–Aromatic Packing Interactions in Proteins. *Journal of the American Chemical Society*, *135*(26), 9877–9884. https://doi.org/10.1021/ja4040472

Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y.-Z., Akutsu, T., Daly, R. J., Webb, G. I., Zhao, Q., Kurgan, L., & Song, J. (2021). iLearnPlus: a Comprehensive and Automated Machine-Learning Platform for Nucleic Acid and Protein Sequence Analysis, Prediction and Visualization. *Nucleic Acids Research*, *49*(10), e60. https://doi.org/10.1093/nar/gkab122

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

Chuang, G.-Y., Boyington, J. C., Joyce, M. G., Zhu, J., Nabel, G. J., Kwong, P. D., & Georgiev, I. (2012). Computational prediction of N-linked glycosylation incorporating structural properties and patterns. *Bioinformatics*, *28*(17), 2249–2255. https://doi.org/10.1093/bioinformatics/bts426

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Culyba, E. K., Price, J. L., Hanson, S. R., Dhar, A., Wong, C.-H., Gruebele, M., Powers, E. T., & Kelly, J. W. (2011). Protein Native-State Stabilization by Placing Aromatic Side Chains in N-Glycosylated Reverse Turns. *Science*, *331*(6017), 571–575. https://doi.org/10.1126/science.1198461

Dey, R., & Mathur, R. (2023). Ensemble Learning Method Using Stacking with Base Learner, A Comparison. In N. Chaki, N. D. Roy, P. Debnath, & K. Saeed (Eds.), *Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023* (Vol. 727, pp. 159–169). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3878-0_14

Edgar, R. C. (2010). Search and Clustering Orders of Magnitude Faster Than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Eisenhaber, B., & Eisenhaber, F. (2010). Prediction of Posttranslational Modification of Proteins from Their Amino Acid Sequence. In C. Oliviero & E. Frank (Eds.), *Data Mining Techniques for the Life Sciences* (Vol. 609, pp. 365–384). Humana Press. https://doi.org/10.1007/978-1-60327-241-4_21

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gattani, S., Mishra, A., & Hoque, M. T. (2019). StackCBPred: A Stacking Based Prediction of Protein-Carbohydrate Binding Sites from Sequence. *Carbohydrate Research*, *486*, 107857. https://doi.org/10.1016/j.carres.2019.107857

Gavel, Y., & Heijne, G. von. (1990). Sequence Differences Between Glycosylated and Non-Glycosylated Asn-X-Thr/Ser Acceptor Sites: Implications for Protein Engineering. *Protein Engineering, Design, and Selection*, *3*(5), 433–442. https://doi.org/10.1093/protein/3.5.433

Gupta, R., & Brunak, S. (2001). Prediction of Glycosylation Across the Human Proteome and the Correlation to Protein Function. *Biocomputing 2002*, 310-322 https://doi.org/10.1142/9789812799623_0029

Hamby, S. E., & Hirst, J. D. (2008). Prediction of Glycosylation Sites Using Random Forests. *BMC Bioinformatics*, *9*(1), 1–13. https://doi.org/10.1186/1471-2105-9-500

Hasan, K. A., & Hasan, Md. A. M. (2020). Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques Resolving Class Imbalance. *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 1–6. https://doi.org/10.1109/iccit51783.2020.9392694

Hauser, M., Steinegger, M., & Söding, J. (2016). MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets. *Bioinformatics, 32*(9), 1323–1330. https://doi.org/10.1093/bioinformatics/btw006

Jorge, V., & Pablo, E. (2014). A Review of Feature Selection Methods Based on Mutual Information. In *Neural Computing and Applications*.

Li, F., Li, C., Revote, J., Zhang, Y., Webb, G. I., Li, J., Song, J., & Lithgow, T. (2016). GlycoMinestruct: A New Bioinformatics Tool for Highly Accurate Mapping of the Human N-linked and O-Linked Glycoproteomes by Incorporating Structural Features. *Scientific Reports*, *6*(1), 1–16. https://doi.org/10.1038/srep34595

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, W., & Godzik, A. (2006). Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Liu, H., & Setiono, R. (1995). Chi2: Feature Selection and Discretization of Numeric Attributes. *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, 388–391. https://doi.org/10.1109/tai.1995.479783

Malik, M., & Undavia, J. N. (2022). Approach and Techniques for Precise Prediction of N-linked Glycosylation from Human Protein using Artificial Intelligence. *International Journal of Engineering Trends and Technology*, *70*(12), 118–126. https://doi.org/10.14445/22315381/ijett-v70i12p213

Moremen, K. W., Tiemeyer, M., & Nairn, A. V. (2012). Vertebrate Protein Glycosylation: Diversity, Synthesis and Function. *Nature Reviews Molecular Cell Biology*, *13*(7), 448–462. https://doi.org/10.1038/nrm3383

Muthu, M., Chun, S., Gopal, J., Anthonydhason, V., Haga, S. W., Jacintha Prameela Devadoss, A., & Oh, J.-W. (2020). Insights into Bioinformatic Applications for Glycosylation: Instigating an Awakening towards Applying Glycoinformatic Resources for Cancer Diagnosis and Therapy. *International Journal of Molecular Sciences*, *21*(24), 9336. https://doi.org/10.3390/ijms21249336

Nita-Lazar, M., Wacker, M., Schegg, B., Amber, S., & Aebi, M. (2005). The N-X-S/T Consensus Sequence is Required But not Sufficient for Bacterial N-linked Protein Glycosylation. *Glycobiology*, *15*(4), 361–367. https://doi.org/10.1093/glycob/cwi019

Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. *International Journal of Information Technology and Computer Science*, *13*(6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05

Pakhrin, S. C., Aoki-Kinoshita, K. F., Caragea, D., & KC, D. B. (2021). DeepNGlyPred: A Deep Neural Network-Based Approach for Human N-Linked Glycosylation Site Prediction. *Molecules*, *26*(23), 7314. https://doi.org/10.3390/molecules26237314

Pakhrin, S. C., Pokharel, S., Aoki-Kinoshita, K. F., Beck, M. R., Dam, T. K., Caragea, D., & KC, D. B. (2023). LMNglyPred: Prediction of Human N-Linked Glycosylation Sites Using Embeddings from A Pre-trained Protein Language Model. *Glycobiology*, *33*(5), 411–422. https://doi.org/10.1093/glycob/cwad033

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

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation Metrics and Statistical Tests for Machine Learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Ramírez, A. S., & Locher, K. P. (2023). Structural and Mechanistic Studies of the N-Glycosylation Machinery: from Lipid-Linked Oligosaccharide Biosynthesis to Glycan Transfer. *Glycobiology*, *33*(11), 861–872. https://doi.org/10.1093/glycob/cwad053

Steinegger, M., & Söding, J. (2017). MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Suresh, S., & Naidu, V. (2022). Mahalanobis-ANOVA Criterion for Optimum Feature Subset Selection in Multi-Class Planetary Gear Fault Diagnosis. *Journal of Vibration and Control*, *28*(21–22), 3257–3268. https://doi.org/10.1177/10775463211029153

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

Tolles, J., & Meurer, W. J. (2016). Logistic Regression: Relating Patient Characteristics to Outcomes. *JAMA*, *316*(5), 533–534. https://doi.org/10.1001/jama.2016.7653

Wang, Y., Wang, D., Ye, X., Wang, Y., Yin, Y., & Jin, Y. (2019). A Tree Ensemble-Based Two-Stage Model for Advanced-Stage Colorectal Cancer Survival Prediction. *Information Sciences*, *474*, 106–124. https://doi.org/10.1016/j.ins.2018.09.046

Wu, T., Zhang, W., Jiao, X., Guo, W., & Alhaj Hamoud, Y. (2021). Evaluation of Stacking and Blending Ensemble Learning Methods for Estimating Daily Reference Evapotranspiration. *Computers and Electronics in Agriculture*, *184*, 106039. https://doi.org/10.1016/j.compag.2021.106039

Xiao, J. (2019). Xiao SVM and KNN ensemble learning for traffic incident detection. *Physica A: Statistical Mechanics and Its Applications*, *517*, 29–35.

Ząbczyńska, M., & Pochec, E. (2015). The Role of Protein Glycosylation in Immune System. *Postepy Biochemii*, *61*(2), 129–137.

Zhang, G., Hou, J., Wang, J., Yan, C., & Luo, J. (2020). Feature Selection for Microarray Data Classification Using Hybrid Information Gain and a Modified Binary Krill Herd Algorithm. *Interdisciplinary Sciences: Computational Life Sciences*, *12*(3), 288–301. https://doi.org/10.1007/s12539-020-00372-w