Heliyon 10 (2024) e32517

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Predictive modeling for ubiquitin proteins through advanced machine learning technique

Shazia [a, †], Fath U Min Ullah [b, †], Seungmin Rho [c], Mi Young Lee [d, *]

[a] Mardan College of Nursing, Bacha Khan Medical College, Mardan, Pakistan
[b] Deparment of Computing, School of Engineering and Computing, University of Central Lancashire, Preston, United Kingdom
[c] Department of Industrial Security, Chung-Ang University, Seoul 06974, Republic of Korea
[d] Chung-Ang University, Seoul 06974, Republic of Korea

ARTICLE INFO

Keywords:
Machine learning
Predictive modeling
Post-translational modification (PTM)
Ubiquitin-protein
Biological computation

ABSTRACT

Ubiquitination is an essential post-translational modification mechanism involving the ubiquitin protein's bonding to a substrate protein. It is crucial in a variety of physiological activities including cell survival and differentiation, and innate and adaptive immunity. Any alteration in the ubiquitin system leads to the development of various human diseases. Numerous researches show the highly reversibility and dynamic of ubiquitin system, making the experimental identification quite difficult. To solve this issue, this article develops a model using a machine learning approach, tending to improve the ubiquitin protein prediction precisely. We deeply investigate the ubiquitination data that is proceed through different features extraction methods, followed by the classification. The evaluation and assessment are conducted considering Jackknife tests and 10-fold cross-validation. The proposed method demonstrated the remarkable performance in terms of 100 %, 99.88 %, and 99.84 % accuracy on Dataset-I, Dataset-II, and Dataset-III, respectively. Using Jackknife test, the method achieves 100 %, 99.91 %, and 99.99 % for Dataset-I, Dataset-II and Dataset-III, respectively. This analysis concludes that the proposed method outperformed the state-of-the-arts to identify the ubiquitination sites and helpful in the development of current clinical therapies. The source code and datasets will be made available at Github.

1. Introduction

Protein synthesis takes place during the 'translation' process. Protein post-translational modification (PTM), one of the final phases of protein biosynthesis, relates to the irreversible as well as reversible chemical modifications which proteins may experience after translation. It involves in a broad range of physiological and pathological processes inside a cell [1]. It controls a variety of processes, including influencing the protein folding efficiency, activation and deactivation of catalytic activities of protein and directing it to the different cellular compartment [2]. Disruption in PTMs can lead to the dysfunction of vital biological processes and results in various diseases. Thus, to identify and treat those diseases at their early stage, the identification of PTMs sites plays a key role. There are around 200 distinct forms of PTMs that impact several areas of cellular activities [3]. Some of them are sulfation [4], acetylation [5],

* Corresponding author.
E-mail addresses: fath@ieee.org (F.U.M. Ullah), miylee@cau.ac.kr (M.Y. Lee).
† Equal Contributions.

https://doi.org/10.1016/j.heliyon.2024.e32517
Received 29 April 2024; Accepted 5 June 2024
Available online 6 June 2024

methylation [6], glycosylation [7], S-Nitrosylation [8] and ubiquitination [9].

Ubiquitin is a small regulatory protein, found in almost all eukaryotic organisms which contain 76 amino acid proteins. The thousand numbers of ubiquitin are attached to the targeted protein simultaneously during their life. This process is known as ubiquitination [10,11]. Ubiquitin is a small regulatory protein that can bind to other proteins through a process called ubiquitination. These modifications usually indicate degradation of the target protein by the proteasome or regulate its activity, localization, or interaction with other proteins. Therefore, in ubiquitin protein likely refers to a protein that has undergone ubiquitination, rather than a protein that catalyzes ubiquitination. Ubiquitination plays a critical role in many cellular activities such as protein localization, determining amount of protein in the membrane, trafficking, and the regulation of physiological functions of membrane proteins [12]. According to many studies, it is also involved in the shedding of membrane and as well as associated proteins as extracellular vesicles that affect not only the amount of certain membrane proteins on the cell surface, but also their potential transport to neighboring cells. Additionally, several diseases such as muscular dystrophy, metabolic syndrome, nervous system disorders and cancer are several disease which can occur due to any kind of abnormality in ubiquitination process [13]. The multistep process of ubiquitination comprises the ubiquitin activation by E1 enzymes, and to conjugate it to E2 enzymes, and the ligation of ubiquitin to the substrate protein by E3 enzymes. Each of these enzyme categories contributes significantly to ubiquitination and the labeling of proteins for the proteasome, which are discussed below.

Before connecting itself to the active part of the cytisine site, it is crucial to activate the ubiquitin with the help of E1 enzyme. Activation of ubiquitin by E1 enzymes, facilitated by adenosine triphosphate (ATP), precedes its conjugation by E2 enzymes, which transfer ubiquitin to the active site cysteine. E3 ligases then recognize and bind target proteins, marking them with ubiquitin molecules through isopeptide bond formation. These ubiquitinated proteins are then subjected to degradation in the proteasome Activation of ubiquitin by E1 enzymes, facilitated by adenosine triphosphate (ATP), precedes its conjugation by E2 enzymes, which transfer ubiquitin to the active site cysteine. E3 ligases then recognize and bind target proteins, marking them with ubiquitin molecules through isopeptide bond formation. These ubiquitinated proteins are then subjected to degradation in the proteasome [13]. The attachment of the ubiquitin molecule to the cytisine site and production of the intermediate ubiquitin-adenylate can be done with the help of adenosine triphosphate (ATP). The ubiquitin-conjugating enzyme (E2) then fulfills its function to join the two molecules by moving the ubiquitin from E1 to the active cysteine site. Due to the E2 enzyme's unique shape, it may bind to both the ubiquitin and E1 molecules and enable this process to take place. Ubiquitin is then transferred from E1 to the active cysteine site by the ubiquitin-conjugating enzyme (E2), which then performs the function of combining the two molecules. Due to the unique shape of the E2 enzyme, it may bind to both the E1 and ubiquitin molecules and enable this process to take place. Finally, the target protein must be recognized and bound by the ubiquitin protein ligase (E3) in order to mark it with the tiny ubiquitin molecule. The last amino acid, glycine 76, of the ubiquitin molecule forms an iso-peptide bond with a lysine on the target protein to accomplish this. The protein is then marked for degradation in the proteasome by forming a short chain containing many ubiquitin molecules by further enzymatic activity. The process of ubiquitination is illustrated in Fig. 1 [11]

The later of the work can be ordered as follows: Section 2 sheds light the literature review while Section 3 focuses on the proposed method. The experimental results and comparative analysis are provided in Section 4 and Section 5. Section 6 concludes the overall works.

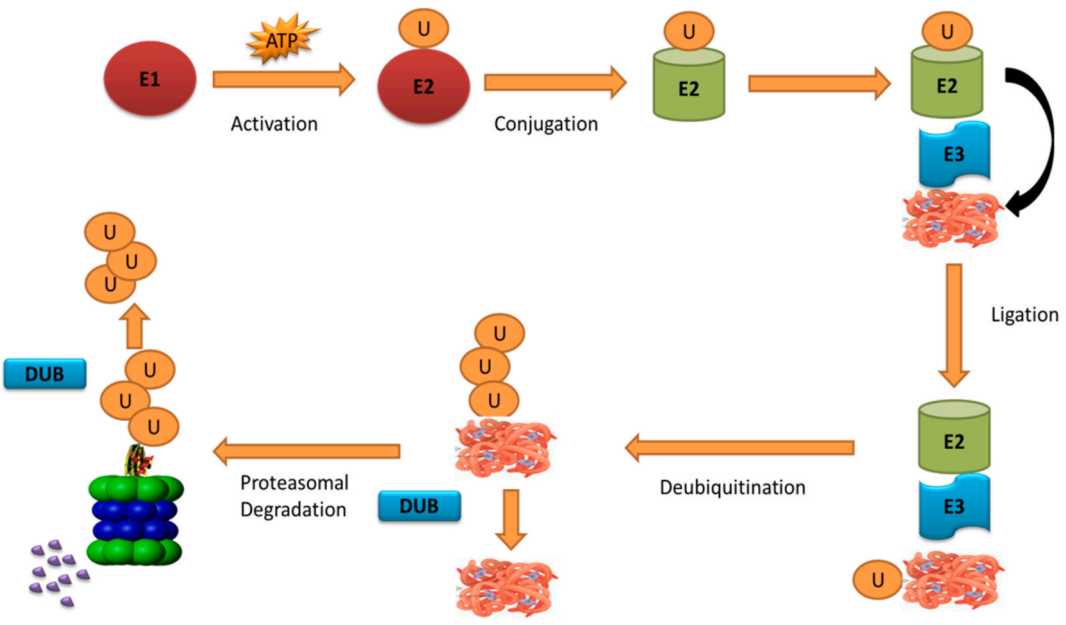

Fig. 1. Ubiquitination involving the sequential action of enzymes to attach ubiquitin to a target protein, signaling various cellular outcomes such as degradation or altered activity

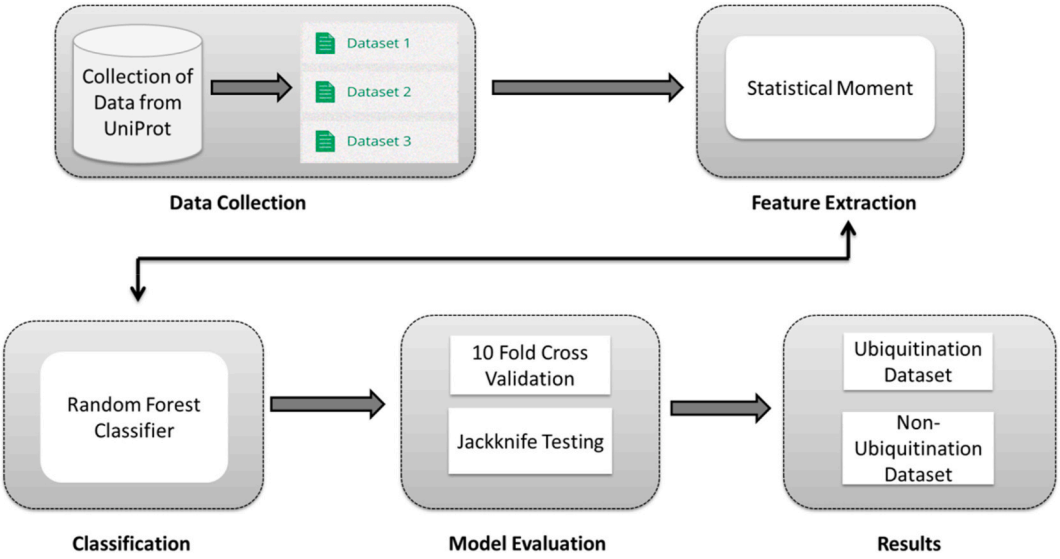

Fig. 2. Overview of the framework of the Proposed Ubiquitination Prediction Method.

2. Literature review

Ubiquitination involved in many biological mechanisms, it is essential to locate them to understand its molecular mechanism and to make it beneficial for researchers [14]. Traditional approaches including the mass spectrometry [15] and the site-directed mutagenesis [16] are utilized to identify these locations. However, due to rapid changes in the ubiquitination process, these methods are less accurate and time consuming [17].

Many researchers have made significant progress in this area [18]. C. W. Tung and S. Y. Ho [19] created a UbiPred model that uses physicochemical property mining technique (IPMA) and Support Vector Machine for classification. Radivojac developed the UbPred model, by integrating the physicochemical features of amino acid components and employed 586 sequences attributes as an input factor [20]. Lee et al. [21] designed a model, by using a kernel named efficient radial basis function (RBF) for the predictions of ubiquitination sites. Next, Cai et al. [22] utilized the KNN technique to predict ubiquitination sites and obtained higher MMC than UbiPred and UbPred. Similarly, Chen et al. [23]created a model CKSAAP and utilized Support Vector Machine as a classification algorithm to predict Ubiquitination sites in yeast. Additionally, hCKSAAP_UbSite is another ubiquitination prediction algorithm with a 0.770 accuracy. Wang et al. [24] created the evolutionary screening algorithm (ESA) to retrieve physicochemical parameters from protein sequences. Working with SVM technique. A Deep Ubi is a model developed by Fu et al. [25] which is used to predict Lysine ubiquitination.

There exists several methods which are used for the ubiquitination prediction purpose. Nguyen et al. [26,27] analyzed protein sequence extraction techniques such as amino acid composition (AAC), amino acid pair composition (AAPC), support vector machine (SVM), and 10-fold cross-validation. The PseAAC was combined with the LASSO as a feature extractor, and three datasets were utilized for training and testing. Many attempts have been made to construct ubiquitination prediction model using various machine learning approaches. However, they are ineffective in terms of performance [19,28–30]. Addressing the challenges posed by the reversibility and dynamic nature of ubiquitin systems, the model employed statistical moment feature extraction on three diverse datasets. This emphasis on capturing the dynamic aspects of ubiquitination enhances the model's robustness, making it adept at identifying crucial sites involved in cell survival, differentiation, and immune responses. The proposed Random Forest-based machine learning model achieves an outstanding accuracy, with a perfect 100 % in 10-fold cross-validation [31], for Dataset-I, and high rates of 99.88 % and 99.84 % for Dataset-II and Dataset-III, respectively. These results signify a significant leap forward in accurately predicting ubiquitin protein sites, surpassing existing models. The model's reliability is confirmed through stringent testing, including 10-fold cross-validation and Jackknife tests [32]. Attaining 100 % accuracy in 10-fold cross-validation for Dataset-I and high rates in Jackknife tests (99.91 % for Dataset-II, 99.99 % for Dataset-III) underscores its robustness. The application of these rigorous testing protocols demonstrates it as a superior technique for ubiquitination site identification compared to prior models.

3. Proposed method

This section focuses on the following steps that are: dataset collection from UniPROT, followed by features extraction of protein sequences or conversion of biological protein sequences into mathematical form and then modeling the extracted dataset for training and classification. Finally, the results evaluation is performed. The proposed method is overviewed in the Fig. 2.

3.1. Data acquisition

In machine learning model, the first step is data collection. It is crucial to gather reliable dataset so that the algorithm used for machine learning can identify the appropriate decision for ubiquitination sites. The precision of the model depends on the quality of the dataset we use [33]. The Ubiquitination dataset, developed by Cai and Jiang for BMC Bioinformatics (2016), can be retrieved through the UniProt database [34]. The three sets of data obtained for the training and testing purpose are given in Table 1.

3.2. Features analysis

For developing machine learning model, researcher uses biological data. Expressing biological data into vector or discrete form without losing its feature and sequence information is a challenging task. Machine learning uses the vector dataset for prediction. To extract feature from the biological sequences, researcher have developed feature extraction methods such Amino Acid Composition (AAC), the Pseudo Amino Acid Composition (PseAAC). These methods are useful for the numerical conversion of sequences into discrete form. In this proposed study, for the statistical description of samples in the benchmark dataset, statistical moment method is adapted. This method is extensively used for the purpose of feature extraction. In this proposed study, we consider Raw, Hahn and central moments for feature extraction which is further discussed below [35,36].

3.2.1. Feature extraction using statistical moment

Statistical moment is widely used for the purpose of extracting features from the protein sequences and to convert them from the biological sequences to numerical form. For the purpose Raw, Hahn and central moment are considered in our proposed study. The function and nature of protein is dependent on the amino acid composition of a particular protein, so it is important that extraction of features should be location variant [3]. As the Raw moment is location variant, which calculates variance, mean and probability distribution of samples in the dataset [37,38]. All of these parameters are computed using central moments as well, although they are location invariant but scale variable as the computations are based on centroid [39]. Hahn polynomials are both scale and location variants, so these are also calculated for each samples of datasets [40]. Here, we calculated a two-dimensional n*n matrix using the protein sequence $\dot{\mathrm{P}}$ as expressed:

$$\dot{\mathrm{P}} = \{c1, c2 \ldots cj\} \tag{1}$$

In the above equation, ci represent the ith amino acid residue component having j residues.

$$\dot{\eta} = \lceil \sqrt{\dot{j}} \rceil \tag{2}$$

$$\dot{p} = \begin{bmatrix} \dot{\mathrm{P}}_{1_1}, & \dot{\mathrm{P}}_{1_2}, & \cdots & \dot{\mathrm{P}}_{1_n} \\ \dot{\mathrm{P}}_{2_1}, & \dot{\mathrm{P}}_{2_2}, & \cdots & \dot{\mathrm{P}}_{2_n} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \dot{\mathrm{P}}_{n_1}, & \dot{\mathrm{P}}_{n_2}, & \cdots & \dot{\mathrm{P}}_{n_n} \end{bmatrix} \tag{3}$$

The above two dimensional matrix is generated by using mapping function ω and turned $\dot{\mathrm{P}}$ into $\dot{p}$, here $\dot{\mathrm{P}}$ represents the main structure of the protein initially used.

$$\omega\,(\dot{\alpha}_{\mathbf{m}}) = \dot{\mathrm{P}}_{ij} \tag{4}$$

Further, the raw moment up to 3° is calculated by the contents of the 2D matrix $\dot{p}$ as presented below

$$\mathrm{RM}_{ij} = \sum_{p=1}^{n} \sum_{q=1}^{n} \mathbf{p}^{i}\mathbf{q}^{j}\,\dot{\mathrm{P}}_{pq} \tag{5}$$

The order of moment is presented by i and j i.e i + j should be ≤ 3; 3° moment is calculated from the areas as RM_{00}, RM_{01}, RM_{10}, RM_{11}, RM_{20}, RM_{21}, RM_{30} and RM_{31}.

The center point from which data emerges in all directions is termed as center of gravity, therefore determining the centroid moment once the raw moment has been obtained is relatively easy. It is given as a point $\overline{k}, \overline{l}$ as defined:

$$\mathbf{k} = \mathrm{RM}_{10}/\mathbf{RM}_{00} \text{ and } \overline{\mathbf{l}} = \mathrm{RM}_{01}/\mathrm{RM}_{00} \tag{6}$$

A central moment's value is calculated using the centroid. It may be calculated mathematically using the following relationship:

$$\dot{\mathrm{C}}_{ij} = \sum_{p=1}^{n} \sum_{q=1}^{n} (p - \overline{k})^{i} (q - \overline{l})^{j}\, \dot{\mathscr{C}}_{pq} \tag{7}$$

In addition, the Hahn polynomial order of n is calculated.

$$\mathbf{H}_{\mathbf{i}}^{\mathbf{y},\mathbf{z}}(\mathbf{r},\mathbf{M}) = (\mathbf{M}+\mathbf{V}-\mathbf{1})_{\mathbf{i}}(\mathbf{M}-\mathbf{1})_{\mathbf{i}} \times \sum_{\mathbf{j}=\mathbf{0}}^{\mathbf{i}} (-\mathbf{1})^{\mathbf{j}} \frac{(-\mathbf{i})_{\mathbf{j}}(-\mathbf{r})_{\mathbf{j}}(\mathbf{2M}+\mathbf{y}+\mathbf{z}-\mathbf{i}-\mathbf{1})_{\mathbf{j}}}{(\mathbf{M}+\mathbf{z}-\mathbf{1})_{\mathbf{j}}(\mathbf{M}-\mathbf{1})_{\mathbf{j}}} \frac{\mathbf{1}}{\mathbf{j}!} \tag{8}$$

The pochhammer symbol in the above equation may be written as:

$$(\mathfrak{b})_{\mathbf{j}} = \mathfrak{b}.(\mathfrak{b}+\mathbf{1})\cdots(\mathfrak{b}+\mathbf{j}-\mathbf{1}) \tag{9}$$

And the Gamma operator is used to simplify it.

$$(\mathfrak{b})_{\mathbf{j}} = \frac{\Gamma(\mathfrak{b}+\mathbf{j})}{\Gamma(\mathfrak{b})} \tag{10}$$

The weighting function and square norm are used to scale the raw values of Hahn moments:

$$\mathbf{H}_{\mathbf{i}}^{\mathbf{y},\mathbf{z}}(\mathbf{r},\mathbf{M}) = \left(\mathbf{H}_{\mathbf{i}}^{\mathbf{y},\mathbf{z}}(\mathbf{r},\mathbf{M})\sqrt{\frac{\mathbf{p}(\mathbf{r})}{\mathbf{d}_{\mathbf{i}}^{\mathbf{2}}}}, \mathbf{n} = \mathbf{0},\mathbf{1},\cdots,\mathbf{M}-\mathbf{1}\right. \tag{11}$$

Meanwhile;

$$\mathbf{p}(\mathbf{r}) = \frac{\Gamma(\mathbf{y}+\mathbf{r}+\mathbf{z})(\mathbf{y}+\mathbf{r}+\mathbf{1})(\mathbf{y}+\mathbf{z}+\mathbf{r}+\mathbf{1})_{\mathbf{M}}}{(\mathbf{y}+\mathbf{z}+\mathbf{2r}+\mathbf{1})\mathbf{n}!(\mathbf{M}-\mathbf{r}-\mathbf{1})!} \tag{12}$$

For 2-dimensional discrete data, the Hahn moment is calculated using the following equation.

$$\mathbf{h}_{\mathbf{ij}} = \sum_{\mathbf{q}=\mathbf{0}}^{\mathbf{N}-\mathbf{1}} \sum_{\mathbf{p}=\mathbf{0}}^{\mathbf{N}-\mathbf{1}} \mathbf{6}_{\mathbf{ij}} \mathbf{H}_{\mathbf{i}}^{\mathbf{y},\mathbf{z}}(\mathbf{q},\mathbf{M}) \mathbf{h}_{\mathbf{j}}^{\mathbf{u},\mathbf{v}}(\mathbf{p},\mathbf{M}), \mathbf{m},\mathbf{n} = \mathbf{0},\mathbf{1},\cdots\mathbf{M}-\mathbf{1} \tag{13}$$

The raw, central, and Hahn moments are computed for each main sequence up to order 3 and a 2-dimensional matrix is formed.

3.2.2. Position relative incidence matrix (PRIM) calculation

It is crucial to measure the relative position of a certain protein sequence, because the primary structure of protein play pivotal role to identify the main characteristics of the protein [41]. It is necessary to quantize the primary sequence of protein in mathematical form to determine the main characteristics. Thus, the PRIM is calculated for the purpose to generate a 20x20 matrix which denotes the information related to their relative position of the ubiquitination sequences for all the samples of datasets.

$$\epsilon_{\mathbf{PRIM}=} \begin{bmatrix} \mathbf{e}_{\mathbf{1}\to\mathbf{1}}, & \mathbf{e}_{\mathbf{1}\to\mathbf{2}},\cdots & \mathbf{e}_{\mathbf{1}\to\mathbf{j}},\cdots & \mathbf{e}_{\mathbf{1}\to\mathbf{20}} \\ \mathbf{e}_{\mathbf{2}\to\mathbf{1}}, & \mathbf{e}_{\mathbf{2}\to\mathbf{2}},\cdots & \mathbf{e}_{\mathbf{2}\to\mathbf{j}},\cdots & \mathbf{e}_{\mathbf{2}\to\mathbf{20}} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{e}_{\mathbf{i}\to\mathbf{1}}, & \mathbf{e}_{\mathbf{i}\to\mathbf{2}},\cdots & \mathbf{e}_{\mathbf{i}\to\mathbf{j}},\cdots & \mathbf{e}_{\mathbf{i}\to\mathbf{20}} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{e}_{\mathbf{n}\to\mathbf{1}}, & \mathbf{e}_{\mathbf{n}\to\mathbf{2}},\cdots & \mathbf{e}_{\mathbf{n}\to\mathbf{j}},\cdots & \mathbf{e}_{\mathbf{n}\to\mathbf{20}} \end{bmatrix} \tag{14}$$

Where each element $\mathbf{e}_{\mathbf{i}\to\mathbf{j}}$ holds the sum of jth residues with respect to the first occurrence of ith residue. For the above matrix 400 coefficients are generated and further statistical moments are calculated to reduce its dimension up to 30 coefficients.

3.2.3. Frequency matrix calculations

The number of amino acids in the ubiquitination sequences determines its frequencies. To precisely determine the frequency matrix the given mathematical equation is calculated:

$$\mathrm{a} = \gamma_{\mathbf{1}}, \gamma_{\mathbf{2}}, \gamma_{\mathbf{3}}, \ldots\ldots\ldots\ldots \gamma_{\mathbf{20}} \tag{16}$$

In the above mentioned equation, γ_{n} represents frequency of occurrence of nth residue. Frequency matrix is calculated to extract compositional information of the sequence.

3.2.4. Accumulative absolute position incidence vector (AAPIV) calculation

After obtaining the compositional information by calculating the frequency, the hidden features related the relative positions of the residues must be calculated, thus a cumulative absolute impact vector of position (AAPIV) of length 20 elements is computed. The sum of all ordinal values for each native amino acid in the original sequence is represented by AAPIV. The following formula is used to calculate AAPIV:

$$\acute{\mathrm{K}} = \bar{\mathrm{w}}_{\mathbf{1}}, \bar{\mathrm{w}}_{\mathbf{2}}, \bar{\mathrm{w}}_{\mathbf{3}}, \ldots\ldots\ldots \bar{\mathrm{w}}_{\mathbf{20}} \tag{17}$$

From this, an arbitrary jth element of AAPIV is computed as:

$$w_j = \sum_{k=1}^{n} p_k \tag{18}$$

3.3. Ubiquitination prediction

The random forest algorithm is a versatile technique used for both classification and regression tasks in machine learning. It works by building multiple decision trees using repeated random sampling called bootstrapping and then combining the results using a method called batch or ensemble learning. This approach uses the power of data collection to outperform individual decision trees and exhibits robustness to overfitting, a common problem, especially in situations characterized by small data sets or noise. This is from a high level. Random forests offer several advantages over independent decision trees, including robustness to outliers and noise, better generalization capabilities, and higher prediction accuracy. Random forest also provide mechanisms for handling missing data and dealing with class imbalance in datasets, making them suitable for real-world applications with heterogeneous and complex data characteristics. Thus, the random forest approach is a powerful and reliable framework for predictive modeling that shows good performance and versatility in various domains, including the proposed ubiquitous classification program.

The approach of using random forest is applied for both classification and regression. This method generates many decision trees by repeating the random sampling (bootstrapping) of predictors and then aggregating the results in a bagging (bootstrap aggregating) or ensemble learning method. The random forest approach beats a single decision tree because it leverages data aggregations, and it is resistant to over fitting. Over fitting may be a concern if the dataset is too small or includes too much noise. The random forest methodology also provides ways for dealing with missing data and balancing faults in datasets with unbalanced classes. In short, random forest models are more accurate and generalizable than standard decision trees. The proposed method also used random forest for the classification of ubiquitination.

4. Results and discussion

After classification, the most important step is to evaluate the performance of the machine learning model. For this, various evaluation tests are carried out, as well as the number of matrices.

4.1. Evaluation Metrics

The assessments of Machine learning models are done using a variety of ways. Accuracy, sensitivity, specificity, and MCC are some model evaluation matrices which are used for the purpose to check the model's stability [42–44]. Mathematical expressions used to represent these matrixes are defined as:

$$\mathbf{Sen} = \mathbf{1} - \frac{\mathbf{A}_{-}^{+}}{\mathbf{A}^{+}} \; 0 \le \mathrm{Sen} \le 1 \tag{19}$$

$$\mathbf{Sp} = \mathbf{1} - \frac{\mathbf{A}_{+}^{-}}{\mathbf{A}_{-}} \; 0 \le \mathrm{Sp} \le 1 \tag{20}$$

$$\mathbf{Accuracy} = \mathbf{1} - \frac{\mathbf{A}_{-}^{+} + \mathbf{A}_{+}^{-}}{\mathbf{A}^{+} + \mathbf{A}^{-}} \; 0 \le \mathrm{Accuracy} \le 1 \tag{21}$$

$$\mathbf{MCC} = \frac{\left(\mathbf{1} - \frac{\mathbf{A}_{-}^{+}+\mathbf{A}_{+}^{-}}{\mathbf{A}^{+}+\mathbf{A}^{-}}\right)}{\sqrt{\left(\mathbf{1} + \frac{\mathbf{A}_{+}^{-}+\mathbf{A}^{+}}{\mathbf{A}^{+}}\right)\left(\mathbf{1} - \frac{\mathbf{A}_{-}^{+}+\mathbf{A}_{+}^{-}}{\mathbf{A}^{-}}\right)}} \; 1 \le \mathrm{Sn} \le 1 \tag{22}$$

A^- represents the negative samples correctly predicted and the A_+^- are the negative samples that the model mistakenly predicted as positive. Similarly, A^+ represents positive predicted dataset and A_-^+ shows positive samples which the model predicted as negative. The formulae shown above are used to calculate accuracy, sensitivity, specificity and MCC, Here the sensitivity will be considered as zero when $A_-^+ = 0$ it means all the positive samples present in the dataset are predicted positive. By $A_-^+ = A^+$, we consider the sensitivity as zero and conclude that the positive dataset is incorrectly predicted as negative samples. Furthermore, $A^+ = 0$ represents the specificity value as zero; it means the all negative samples are accurately predicted. Similarly, $A_+^- = A^-$, the specificity vale is considered as zero when all the negative samples are inaccurately predicted as positive. Moreover, $A_-^+ = A_+^-$, represents that the Accuracy and Mathew's correlation coefficient (MMC) values will be 1 if all the ubiquitination and non-ubiquitination are correctly predicted; The terms $A_-^+ = A^+$ and $A_+^- = A^-$, shows that all the negative samples are inaccurately predicted as positive and positive samples are predicted as negative and here both MMC and accuracy values will be zero. All these measurement matrices are used to test and evaluate the model with different approaches.

Various feature extraction methods are used including statistical moment analysis, position relative incidence matrix (PRIM) calculation, frequency matrix calculation, and overall absolute position incidence vector (AAPIV) calculation. These methods were

chosen to capture different aspects of the protein sequence associated with ubiquitylation prediction. We performed a feature importance analysis. Our analysis revealed that some features obtained in PRIM and frequency matrix calculations play an important role in distinguishing between ubiquitous and non-ubiquitous sites. These features capture the key structural and conformational features of protein sequences that characterize ubiquitylation. Statistical moments analysis allows us to measure the distribution and variability of amino acid residues in protein sequences, providing valuable insight into their structural and conformational properties. PRIM calculations enable quantification of the relative positions of amino acid residues, which is essential for identifying key structural motifs bound to each site. Frequency matrix calculations yield conformational information of protein sequences, highlighting amino acid preferences and biases at specific sites. Finally, AAPIV calculations provide insight into the cumulative effect of amino acid positions on the overall prediction of each locus.

By taking advantage of these diverse feature extraction methods, our approach goes beyond traditional PSSM-based methods to obtain a more comprehensive representation of protein sequence information. This allows us to effectively discriminate between ubiquitous and non-ubiquitous spaces with high accuracy and robustness.

4.2. Testing methods

In machine learning evaluation techniques plays significant role for the estimation of machine learning model's performance. To check the robustness, reproducibility and interpretability of the proposed method many testing methods are used which are discussed below.

4.2.1. Jackknife testing

One of the cross validation techniques is Jackknife testing techniques which is based on resampling. The method perform well for the variances and bias [45]. If there are N numbers of samples in a dataset, then the predictor is trained on N-1 numbers of dataset and testing is done on the remaining one dataset Thus the method is also called leave-one-out cross validation. The process is repeated N times and the predicted labels are noted for iteration. Eventually, all the predicted values are collected and their average represents the overall accuracy results [36]. The Ubi-Pred-RF predictor yields remarkable results while evaluated using jackknife testing and the accuracy values are is 100 %, 99.91 %, and 99.91 for the Dataset-I, Dataset-II, and Dataset-III respectively (see Table 2).

4.2.2. K-fold cross-validation method

K-fold cross-validation is one of the statistical methods which are mainly used for the estimation of machine learning models. Here the parameter k refers to the number of groups that the given data sample is split into. This method is simple to implement, and it is performs less biased evaluation, this feature makes it popular [36–46]. In the proposed method, cross validation performed on the following steps. Firstly, all the datasets i.e Dataset-I, Dataset-II and Dataset III are shuffled and split into 10 numbers of subsamples. In the first iteration, the first subset is used as a test set and the remaining subsamples are used for the training of the machine learning model. In the second iteration the second subset is used as test set while the remaining as training set. This is performed for all the 10 iteration and at the last the overall accuracy is obtained by averaging the output acquired by all the iteration. By 10-fold cross validation our proposed method performed remarkable results for the three datasets that are presented in Table 2.

5. Comparative analysis

The results of Ubi-Pred-RF are compared with previously developed methods using the same benchmark datasets and are reported in Table 3. UbiSitePred [47][48] and Cai et al. [34] uses various machine learning approaches for the ubiquitination sites prediction. Along with that, other ubiquitination prediction models are also compared as illustrated in Table .3. It can be observed by Table 3 that the proposed method Ubi-Pred-RF performs efficiently for all the parameters among all the other methods which clarify the superiority of the model among the existing one.

Additionally, comparison with previously developed methods using the same benchmark dataset provides valuable insight into the performance of Ubi-Pred-RF compared to existing methods. The method proposed by UbiSitePred [47,48] and Cai et al. [34] provide a variety of comparative methods using different machine learning techniques to predict ubiquitous sites. By benchmarking Ubi-Pred-RF against these established methods, we can evaluate its efficiency and robustness across different prediction frameworks. Additionally, comparison with additional ubiquitous prediction models presented in Table 3 further strengthens the performance evaluation of Ubi-Pred-RF. By considering various existing methods based on different algorithmic approaches and feature sets, we obtain a comprehensive understanding of the relative performance and superiority of Ubi-Pred-RF. The results show that Ubi-Pred-RF consistently outperforms other methods across a variety of evaluation parameters, reaffirming its efficiency and superiority for ubiquitous site prediction. This highlights the importance of our proposed approach and its potential impact on the advancement of the

Table 1
Curated datasets for Ubiquitination.

Dataset Type	Dataset-I [23]	Dataset-II [19]	Dataset-III [22]
Ubiquitination	150	3419	6118
Non Ubiquitination			
Total	300	6838	12,236

Table 2
Comprehensive evaluation of Ubi-Pred-RF: Insights from 10-fold Cross-Validation results in terms of Accuracy (ACC), Specificity (SP), Sensitivity (Sen), and Matthews correlation coefficient (MCC). [Proposed Method: PM, Ubi-Pred-RF: UBP].

PM	Datasets along with their results											
UBP	Dataset-I				Dataset-II				Dataset-III			
K-Fold	ACC	SP	SEN	MCC	ACC	SP	SEN	MCC	ACC	SP	SEN	MCC
1	100	100	100	100	100	100	100	100	99.92	100	100	100
2	100	100	100	100	99.71	0.99	100	0.99	99.84	100	100	100
3	100	100	100	100	99.85	100	100	100	100	100	100	100
4	100	100	100	100	99.85	100	100	100	99.92	100	100	100
5	100	100	100	100	99.85	100	100	100	100	100	100	100
6	100	100	100	100	99.71	100	100	0.99	100	100	100	100
7	100	100	100	100	99.85	100	100	100	100	100	100	100
8	100	100	100	100	100	100	100	100	99.92	100	100	100
9	100	100	100	100	100	100	100	100	100	100	100	100
10	100	100	100	100	100	100	100	100	99.84	100	100	100
	Final score = 100				Final score = 99.9				Final score = 99.84			

Table 3
Analytical assessment: Contrasting Ubi-Pred-RF against Alternative ubiquitination prediction methods. [Ubi-Pred-RF: UBP, D:Dataset].

Model		Positive samples	ACC (%)	SEN (%)	SP (%)	AUC	MCC
UBP	DI	150	100	100	100		100
	DII	3419	99.9	100	0.99		0.998
	DIII	6118	99.84	100	100		100
DeepUbi		53,999	88.98	89.80	88.10	0.91	0.78
ESA-UbiSite		85	94.0	96.0	92.0		0.92
hCKSAAP_UbSite		9535				0.77	
CKSAAP_UbSite		263	73.4	69.85	76.96	0.81	0.47
UbSite		385	74.5	65.5	74.8		
		265	72.0			0.79	
		151	84.44	83.44	85.43	0.85	0.69

field of ubiquitous forecasting. Through rigorous benchmarking and comparison with existing methods, our study provides strong evidence for the outstanding performance of Ubi-Pred-RF and highlights its potential as a valuable tool in bioinformatics research and biomedical applications.

6. Conclusions

In this study for the accurate prediction of ubiquitination prediction three datasets are utilized. Ubiquitination sequences are converted into discrete form using statistical moment is used and further the model is developed using machine learning algorithm that are Random Forest and SVM. Different evaluation techniques such as 10-fold Cross Validation and Jackknife testing are used. The proposed Random Forest-based machine learning model achieve an outstanding accuracy, with a perfect 100 % in 10-fold cross-validation, for Dataset-I, and high rates of 99.88 % and 99.84 % for Dataset-II and Dataset-III, respectively and for Jackknife tests (99.91 % for Dataset-II, 99.99 % for Dataset-III). Our results and findings demonstrate the effectiveness of the Ubi-Pred-RF model in predicting ubiquity, surpassing the performance of previous models documented in the literature. This increased accuracy holds significant promise for a number of applications, particularly in drug development and diagnosis of serious diseases. Although our study represents an important advance in the field of ubiquitous spatial prediction, it has few limitations. One such limitation is the reliance on selective datasets that may not fully capture the diversity and complexity of ubiquitous processes. Furthermore, our model requires further validation to generalize to real-world scenarios, particularly in the context of different cellular environments and disease states. In the future, we will explore synergistic approaches combining well-established deep learning and machine learning techniques with statistical inference. By investigating more complex post-translational modification (PTM) sites, we aim to further refine predictive models and uncover deeper insights into cellular processes.

CRediT authorship contribution statement

Shazia: Funding acquisition, Project administration, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft. **Fath U Min Ullah:** Investigation, Methodology, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Chung-Ang University Research Grants in 2024.

References

[1] H. Wang, L. Yang, M. Liu, J. Luo, Protein post-translational modifications in the regulation of cancer hallmarks, . 30, n, Cancer Gene Ther. 304 (4) (2022) 529–547, https://doi.org/10.1038/s41417-022-00464-3. Apr. 2022.
[2] J.M. Lee, H.M. Hammarén, M.M. Savitski, S.H. Baek, Control of protein stability by post-translational modifications, Nat. Commun. 14 (1) (2023) 1–16, https://doi.org/10.1038/s41467-023-35795-8.
[3] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, K.C. Chou, iPhosY-PseAAC: identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC, Mol. Biol. Rep. 45 (6) (Dec. 2018) 2501–2509, https://doi.org/10.1007/s11033-018-4417-z.
[4] D. Dai, et al., Enhanced tyrosine sulfation is associated with chronic kidney disease-related atherosclerosis, BMC Biol. 21 (1) (2023) 1–17, https://doi.org/10.1186/s12915-023-01641-y.
[5] W. Bao, B. Yang, Protein acetylation sites with complex-valued polynomial model, Frontiers (Boulder). 18 (3) (2024).
[6] S. Akbar, M. Hayat, iRNA-PseTNC : Identification of RNA 5-methylcytosine Sites Using Hybrid Vector Space of Pseudo Nucleotide Composition, 2020. August 2019.
[7] H. Li, A.W.T. Chiang, N.E. Lewis, Artificial intelligence in the analysis of glycosylation data, Biotechnol. Adv. 60 (2022), https://doi.org/10.1016/j.biotechadv.2022.108008.
[8] C. ki Oh, et al., Targeted protein S-nitrosylation of ACE2 inhibits SARS-CoV-2 infection, Nat. Chem. Biol. 19 (3) (2023) 275–283, https://doi.org/10.1038/s41589-022-01149-6.
[9] Z. Zhang, et al., Elucidation of E3 ubiquitin ligase specificity through proteome-wide internal degron mapping, Mol. Cell 83 (18) (2023) 3377–3392.e6, https://doi.org/10.1016/j.molcel.2023.08.022.
[10] T. Sun, Z. Liu, Q. Yang, The role of ubiquitination and deubiquitination in cancer metabolism, Mol. Cancer 19 (1) (2020) 1–19, https://doi.org/10.1186/s12943-020-01262-x.
[11] A. Ciechanover, K. Iwai, The ubiquitin system: from basic mechanisms to the patient bed, IUBMB Life 56 (4) (2004) 193–201, https://doi.org/10.1080/15216540420000223616.
[12] The Ubiquitin-Proteasome Proteolytic Pathway_ Destruction for the Sake of Construction _ Enhanced Reader.pdf.".
[13] C.M. Pickart, Mechanisms underlying ubiquitination, Annu. Rev. Biochem. 70 (2001) 503–533, https://doi.org/10.1146/annurev.biochem.70.1.503.
[14] Z. Chen, Y. Zhou, Z. Zhang, J. Song, Towards more accurate prediction of ubiquitination sites: a comprehensive review of current methods, tools and features, Brief. Bioinform. 16 (4) (Jul. 2014) 640–657, https://doi.org/10.1093/bib/bbu031.
[15] E. de Hoffmann, Mass spectrometry, in: Kirk-Othmer Encyclopedia of Chemical Technology, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2005, https://doi.org/10.1002/0471238961.1301191913151518.a01.pub2.
[16] M.S. Gentry, C.A. Worby, J.E. Dixon, Insights into Lafora disease: malin is an E3 ubiquitin ligase that ubiquitinates and promotes the degradation of laforin, Proc. Natl. Acad. Sci. U. S. A. 102 (24) (Jun. 2005) 8501–8506, https://doi.org/10.1073/pnas.0503285102.
[17] C. Kannicht, B. Fuchs, Post-translational modifications of proteins, Mol. Biomethods Handb (2008) 427–449, https://doi.org/10.1007/978-1-60327-375-6_28. Second Ed.
[18] W.R. Qiu, X. Xiao, W.Z. Lin, K.C. Chou, IUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model, J. Biomol. Struct. Dyn. 33 (8) (Aug. 2015) 1731–1742, https://doi.org/10.1080/07391102.2014.968875.
[19] C.W. Tung, S.Y. Ho, Computational identification of ubiquitylation sites from protein sequences, BMC Bioinf. 9 (1) (Jul. 2008) 1–15, https://doi.org/10.1186/1471-2105-9-310.
[20] P. Radivojac, et al., Identification, analysis, and prediction of protein ubiquitination sites, Proteins Struct. Funct. Bioinforma. 78 (2) (2010) 365–380, https://doi.org/10.1002/prot.22555.
[21] T.-Y. Lee, S.-A. Chen, H.-Y. Hung, Y.-Y. Ou, V. Uversky, Incorporating Distant Sequence Features and Radial Basis Function Networks to Identify Ubiquitin Conjugation Sites, 2011, https://doi.org/10.1371/journal.pone.0017331.
[22] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, Y. Li, Prediction of lysine ubiquitination with mRMR feature selection and analysis, Amino Acids 42 (4) (Apr. 2012) 1387–1395, https://doi.org/10.1007/s00726-011-0835-0.
[23] Z. Chen, Y. Zhou, J. Song, Z. Zhang, HCKSAAP-UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties, Biochim. Biophys. Acta, Proteins Proteomics 1834 (8) (Aug. 2013) 1461–1467, https://doi.org/10.1016/j.bbapap.2013.04.006.
[24] J.-R. Wang, W.-L. Huang, M.-J. Tsai, K.-T. Hsu, H.-L. Huang, and S.-Y. Ho, "ESA-UbiSite: Accurate Prediction of Human Ubiquitination Sites by Identifying a Set of Effective Negatives", doi: 10.1093/bioinformatics/btw701.
[25] H. Fu, Y. Yang, X. Wang, H. Wang, Y. Xu, DeepUbi: a deep learning framework for prediction of ubiquitination sites in proteins, BMC Bioinf. 20 (1) (2019) 1–10, https://doi.org/10.1186/s12859-019-2677-9.
[26] V.N. Nguyen, K.Y. Huang, C.H. Huang, K.R. Lai, T.Y. Lee, A new scheme to characterize and identify protein ubiquitination sites, IEEE ACM Trans. Comput. Biol. Bioinf 14 (2) (Mar. 2017) 393–403, https://doi.org/10.1109/TCBB.2016.2520939.
[27] Z. Chen, Y. Zhou, Z. Zhang, and J. Song, "Towards More Accurate Prediction of Ubiquitination Sites: a Comprehensive Review of Current Methods, Tools and Features", doi: 10.1093/bib/bbu031.
[28] W.R. Qiu, X. Xiao, W.Z. Lin, K.C. Chou, IUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model, J. Biomol. Struct. Dyn. 33 (8) (Aug. 2015) 1731–1742, https://doi.org/10.1080/07391102.2014.968875.
[29] X.B. Wang, L.Y. Wu, Y.C. Wang, N.Y. Deng, Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs, Protein Eng. Des. Sel. 22 (11) (2009) 707–712, https://doi.org/10.1093/protein/gzp055.
[30] Z. Ju, H. Gu, Predicting pupylation sites in prokaryotic proteins using semi-supervised self-training support vector machine algorithm, Anal. Biochem. 507 (2016) 1–6, https://doi.org/10.1016/j.ab.2016.05.005.
[31] OpenML." https://www.openml.org/a/estimation-procedures/7 (accessed February. 2, 2021).
[32] Jackknife Test - an overview | ScienceDirect Topics." https://www.sciencedirect.com/topics/nursing-and-health-professions/jackknife-test (accessed January 26, 2021).
[33] K. Haglund, I. Dikic, Ubiquitylation and cell signaling, EMBO J, EMBO J. 24 (19) (2005) 3353–3359, https://doi.org/10.1038/sj.emboj.7600808. Oct. 05.
[34] B. Cai, X. Jiang, Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences, BMC Bioinf. 17 (1) (2016) 1–12, https://doi.org/10.1186/s12859-016-0959-z.
[35] statistical moment - Google Search." https://www.google.com/search?q=statistical+moment&oq=statistical+moment&aqs=chrome.69i57j0l5j0i20i263i395j69i60.5630j1j7&sourceid=chrome&ie=UTF-8 (accessed February. 2, 2021).

[36] I. Technology, K.A. Aziz, S. Arabia, "PREDICTION OF SAUDI ARABIA SARS-COV 2 DIVERSIFICATIONS IN PROTEIN STRAIN AGAINST CHINA STRAIN," 8 (1) (2020) 64–73.
[37] Z. Ju, S.Y. Wang, Prediction of citrullination sites by incorporating k-spaced amino acid pairs into Chou's general pseudo amino acid composition, Gene 664 (Jul. 2018) 78–83, https://doi.org/10.1016/j.gene.2018.04.055.
[38] Y.D. Khan, N. Rasool, W. Hussain, S.A. Khan, K.C. Chou, iPhosT-PseAAC: identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC, Anal. Biochem. 550 (Jun. 2018) 109–116, https://doi.org/10.1016/j.ab.2018.04.021.
[39] G. Gerig, "Lecture: Shape Analysis Moment Invariants," Cs 7960 (38) (2010). Available: http://www.sci.utah.edu/~gerig/CS7960-S2010/handouts/CS7960-AdvImProc-MomentInvariants.pdf.
[40] Y.D. Khan, S.A. Khan, F. Ahmad, S. Islam, Iris recognition using image moments and k-Means algorithm, Sci. World J. 2014 (2014), https://doi.org/10.1155/2014/723595.
[41] M.A. Akmal, N. Rasool, Y. Daanial Khan, Prediction of N-Linked Glycosylation Sites Using Position Relative Features and Statistical Moments, 2017, https://doi.org/10.1371/journal.pone.0181966.
[42] Evaluation Metrics Definition | DeepAI." https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics (accessed February. 2, 2021).
[43] Machine Learning Classifier: Basics and Evaluation | by James Le | Data Notes | Medium." https://medium.com/cracking-the-data-science-interview/machine-learning-classifier-basics-and-evaluation-44dd760fea50 (accessed February. 2, 2021).
[44] Evaluation Metrics Machine Learning." https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/(accessed February. 2, 2021).
[45] Jackknife - an overview | ScienceDirect Topics." https://www.sciencedirect.com/topics/mathematics/jackknife (accessed February. 2, 2021).
[46] What is Cross Validation in Machine learning? Types of Cross Validation." https://www.mygreatlearning.com/blog/cross-validation/(accessed February 11, 2021).
[47] X. Cui, Z. Yu, B. Yu, M. Wang, B. Tian, Q. Ma, UbiSitePred: a novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components, Chemometr. Intell. Lab. Syst. 184 (2019) 28–43, https://doi.org/10.1016/j.chemolab.2018.11.012.
[48] IU. Haq, F.U. Ullah, K. Muhammad, S.W. Baik, Deep learning techniques for oral cancer diagnosis. InComputational Intelligence in Cancer Diagnosis, 1, Academic Press, 2023, pp. 175–193. Jan.