membranes

MDPI

Article

iAcety–SmRF: Identification of Acetylation Protein by Using Statistical Moments and Random Forest

Sharaf Malebary [1], Shaista Rahman [2,*], Omar Barukab [1], Rehab Ash'ari [1] and Sher Afzal Khan [2]

[1] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21911, Saudi Arabia; smalebary@kau.edu.sa (S.M.); obarukab@kau.edu.sa (O.B.); rashary@kau.edu.sa (R.A.)
[2] Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan; sher.afzal@awkum.edu.pk
* Correspondence: shista123rahman@gmail.com

Abstract: Acetylation is the most important post-translation modification (PTM) in eukaryotes; it has manifold effects on the level of protein that transform an acetyl group from an acetyl coenzyme to a specific site on a polypeptide chain. Acetylation sites play many important roles, including regulating membrane protein functions and strongly affecting the membrane interaction of proteins and membrane remodeling. Because of these properties, its correct identification is essential to understand its mechanism in biological systems. As such, some traditional methods, such as mass spectrometry and site-directed mutagenesis, are used, but they are tedious and time-consuming. To overcome such limitations, many computer models are being developed to correctly identify their sequences from non-acetyl sequences, but they have poor efficiency in terms of accuracy, sensitivity, and specificity. This work proposes an efficient and accurate computational model for predicting Acetylation using machine learning approaches. The proposed model achieved an accuracy of 100 percent with the 10-fold cross-validation test based on the Random Forest classifier, along with a feature extraction approach using statistical moments. The model is also validated by the jackknife, self-consistency, and independent test, which achieved an accuracy of 100, 100, and 97, respectively, results far better as compared to the already existing models available in the literature.

Keywords: acetylation; random forest; probabilistic neural network; statistical movement; post-translational modification; machine learning; membrane proteins

check for updates

Citation: Malebary, S.; Rahman, S.; Barukab, O.; Ash'ari, R.; Khan, S.A. iAcety–SmRF: Identification of Acetylation Protein by Using Statistical Moments and Random Forest. *Membranes* **2022**, *12*, 265. https://doi.org/10.3390/membranes12030265

Academic Editor: Olga Vinogradova

Received: 10 December 2021
Accepted: 1 February 2022
Published: 25 February 2022

1. Introduction

Proteins are the basic and key part of the human body that perform many kinds of major functions in and outside a cell. The proteins are translated or synthesized from messenger RNA which is first codified into ribosomes and makes a chain of amino acid or polypeptide. After the translation process, certain amino acids can experience chemical changes at the protein's C-termini or N-termini or in amino acid side chains, known as post-translation modifications (PTLM or PTM). The PTM can modify or may introduce the new functional group to the protein, such as in Acetylation, an example of Acetyl-lysine is shown in the Figure 1) [1]. It plays a key role in making protein products [2–4]. Each protein in the proteome may be altered either before or after it is translated. The charge state, hydrophobicity, conformation, and stability of a protein are all affected by various changes, which, in turn, influence its function. Protein modification has a variety of functions in different organs: (1) It ensure the fast and complex response of cells to regulate intra-cellular communication, division, and growth of cells (2) also pivotal for various physiological and pathological mechanism.

Protein acetylation can be achieved using a variety of methods, this adds the acetyl functional group into a chemical compound, which make another ester, the acetate.

Another form of lysine residue is usually acetylated [1–4]. The active substance, acetic anhydride, is commonly used to react with free hydroxyl groups as an acetylating agent.

It is used in, for example, aspirin, heroin, and THC-O-acetate synthesis. Thousands of acetylated mammalian proteins have been identified [1], in addition to protein analysis. The acetylation takes place, for example, via a co-translation and a post-translational adaptation of proteins, histone, and p53, in addition to tubulins. Among these proteins, there is a high representation of chromatin proteins and metabolic enzymes, suggesting that acetylation, in addition to digestion, has an extremely important influence on the appearance of the genetic material. Among the microbes, 90 percent of the proteins that are surrounded by the central metabolism of Salmonella Centrica are acetylated [1,2].

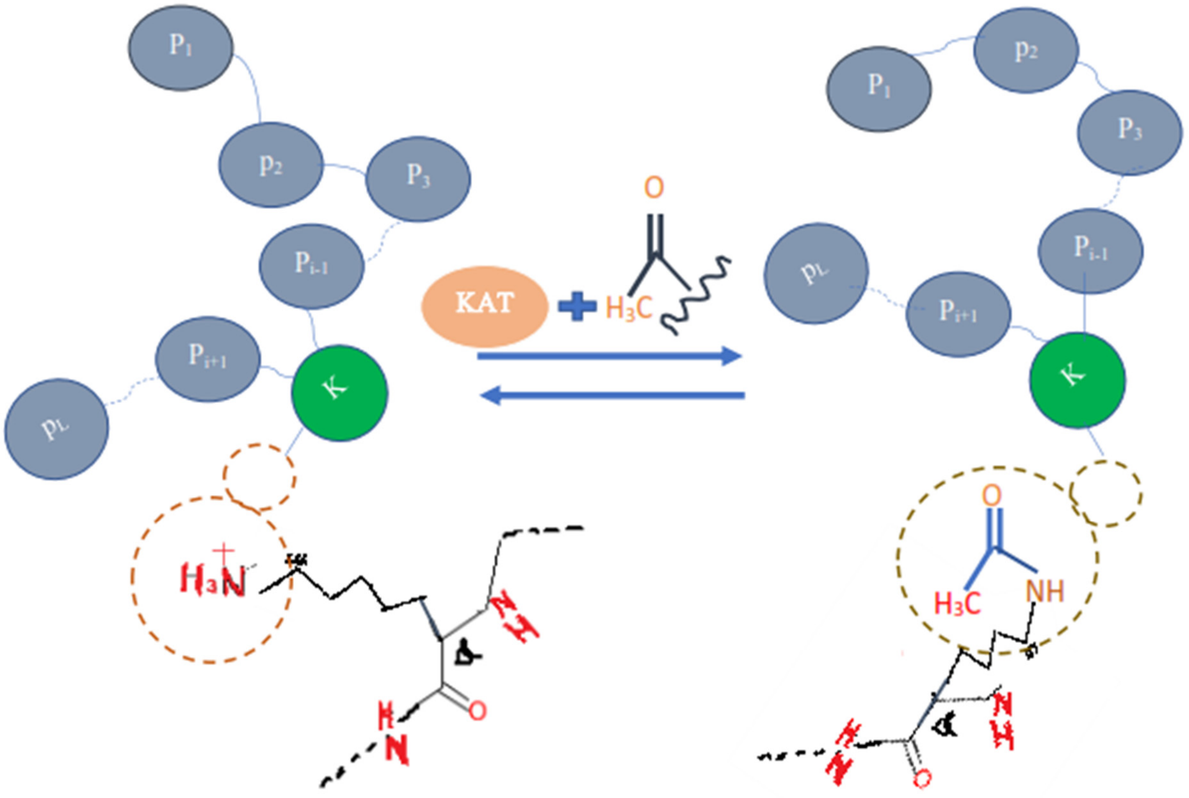

Figure 1. Acetylation protein.

Further, acetylation sites also play an important role in regulating membrane protein functions of multiple families, as documented in Reference [5]. It is supported by examples that acetylation is significantly enhanced in membrane-binding regions, where it is often located directly in critical membrane-binding pockets ideally positioned to modulate membrane interactions. Moreover, it was found that acetylation and acetylation mimetics strongly affected the membrane interaction of proteins, resulting in decreased membrane affinity and, in the case of amphiphysin and EHD2, altered membrane remodeling [6]. In cells, mimicking even a single acetylation event within the membrane interaction region reduced the binding affinity to membranes, resulting in cytoplasmic dispersal. In another report, acetylation affected the effects on membrane interaction, as well as membrane remodeling [7]. Similarly, the ability to control the membrane-binding activity of C2 domains via acetylation could allow the cell to further regulate Ca-dependent transmembrane transport and signaling events. It has also been documented that acetylation is present on the membrane-binding surface of the phosphatase domain of K163/K164, as it appears to be that Alanine mutations reduce membrane binding [8]. In addition, two reports on proteins with PH domains indicate that acetylation has opposite effects on membrane localization in cells (either increased or decreased) [9,10]. Much has been learned about the acetyltransferases and deacetylases that regulate protein-DNA-protein-protein interactions [2,5,11]. Some of these enzymes may also be involved in controlling protein membrane interactions.

Consistent with this idea, localization of acetyltransferase in the cytoplasm and deacetylases in cell membranes have been observed [12–14].

Acetylation also plays a prominent role in numerous important cellular processes, such as stability and localization of protein [4,5]. In addition, modification of S/T/Y sites by acetylation, glycosylation, sulfation, and nitration has been reported [5,6]. Moreover, it also plays a role in the modulation of gene expression by histone alteration, as well as is a very significant function in controlling cellular metabolism and protein folding [15–18].

From the above discussion it revealed that Acetylation is an important post-translation modification, and it is necessary to correctly identify them; however, it remains a major challenge to understand the functions and regulations of the molecular acetylation mechanism. Many traditional approaches are in use for their identification, including high-throughput mass spectrometry (MS) [19,20]. However, since the acetylation mechanism is complicated, rapid, and reversible, such methods remain time-consuming, expensive, and laborious [21,22].

To overcome the existing problems in its identification, many researchers have developed a computational model for fast and inexpensive prediction of PTM sites [15,16,23,24], such as the ubiquitination [17,18,24], the phosphorylation [15,16,19,25–27], sumoylation [28–30], and the acetylation [31–34], etc. The important step for the PTM prediction model to correctly transform the biological sequences into their equivalent numerical form, for this purpose, many feature extraction methods are developed which are documented like the amino acids composition (AAC), the dipeptides composition (DPC), and Pseudo Amino Acid Composition (PseAAC) [35]. For such feature extractions many methods are discussed in Reference [20].

Reference [1] proposed a novel measurement procedure iAcet-PseFDA, a classification model for acetylation proteins by extracting features come from sequence conservation information using a gray structure model and KNN scoring based on functional domain annotation databases including GO [36] and subcellular localization for acetylation protein recognition. The authors achieved 77.10 percent accuracy using 5-fold cross-validation on three datasets, with a significant amount of attribute analysis and the discovery algorithm for relief functionality.

Reference [37] proposed a method ProAcePred to predict prokaryote-specific lysine acetylation sites, using SVM, 10-fold cross-validation, and the elastic net mathematical approach for optimizing the dimensionality of feature vectors, which greatly increased prediction accuracy and yielded promising results.

Wuyun et al. [38], developed a model, KA-predictor, to predict species-specific lysine acetylation sites using the classifier SVM. They achieved highly competitive for the majority of species as compared with other methods.

Hou [39] suggested a predictor for lysine acetylation prediction called LAceP, based on logistic regression classifiers and various biological characteristics. Using Random Forest classifiers, Li [40] developed SSPKA, a tool for species-specific lysine acetylation prediction.

From the above discussion, it has been observed that many predictors have been developed for the identification of acetylation sites; however, the maximum prediction accuracy established in all previous models was 77.10%, which is very poor for a correct identification of the acetylation sites.

Therefore, in the present study, we use statistical moments as feature extractions and Random Forest and PNN as a classifier. Further, the model evaluation is done by 10-fold cross-validation, self-consistency, independent, and by jackknife testing. We obtained dominant results as compared to the existing models that were developed earlier.

Therefore, to improve the predictor model, we use statistical moments as feature extractions and Random Forest and PNN as classifiers. The model evaluation is carried out through 10-fold cross-validation, self-consistency, independently, and the jackknife tests. We achieved dominant results as compared to the existing models developed earlier.

2. Materials and Methods

In this review, we follow the 5-step procedure mentioned in References [16,41] to establish a predictive predictor for biological sequences. It consists of the following steps: (1) choosing or generating an appropriate benchmark dataset to be used for training and testing; (2) transforming a biological sequence into its equivalent mathematical form, which returns the basic correlation with the biological sequence, and transforming a biological sequence into its equivalent mathematical form, which returns the required correlation with the biological sequence; (3) developing or using existing classifier for the required predictor; (4) using cross-validation experiments to determine the accuracy of the suggested indicator; and, finally, (5) developing an influential website/GitHub resource for public use for the benefit of future research and development. All the above steps are presented in Figure 2. Further, details of the above can be found in the following subsections.

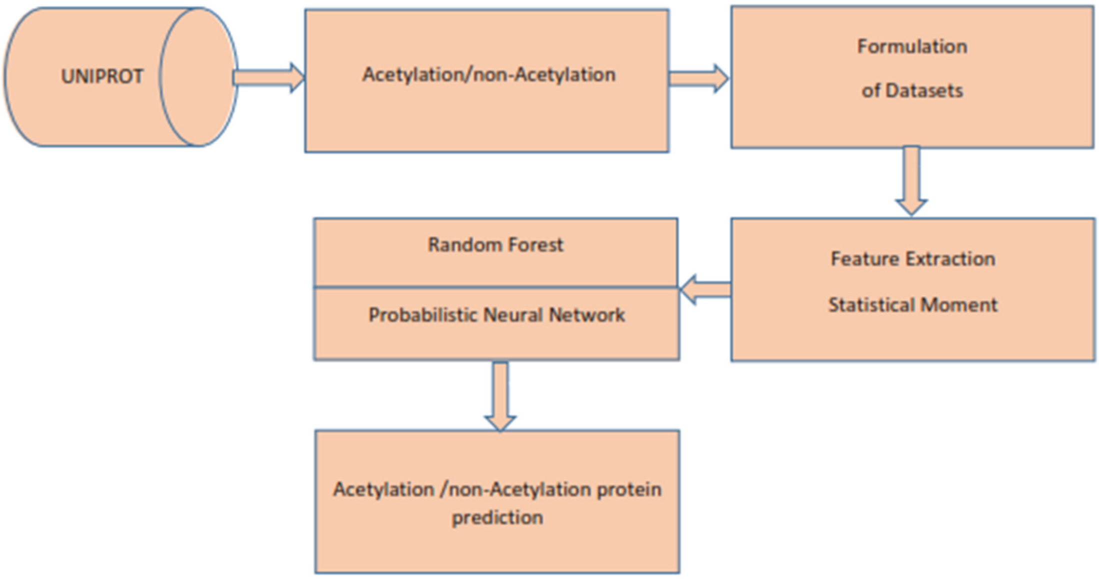
UNIPROT
Acetylation/non-Acetylation
Formulation of Datasets
Feature Extraction
Statistical Moment
Random Forest
Probabilistic Neural Network
Acetylation /non-Acetylation protein prediction

Figure 2. Flowchart of the proposed predictor.

3. Benchmark Dataset

We begin with collection of a valid benchmark dataset for training and testing, which is the first step in the 5-step rule [1]. The dataset is collected from the well-known data repository, the UNIPROT http://www.uniprot.org, retrieved dated 8 December 2021. It contains 2900 protein samples, of which 725 were positive denoted by S_{posi}, and 2175 were negative denoted by S_{negt}. Further, for the effectiveness of the proposed predictor, the set of negative data samples is equally divided in three sets, $\overline{S}_1$, $\overline{S}_2$, and $\overline{S}_3$ as in Reference [1] such that,

$$\overline{S}_i \cap \overline{S}_j = \phi, \quad (i \neq j;, \ i, \ j = 1, \ 2, \ 3),$$

where the sign $\cap$ represents the intersection of sets. Furthermore, we individually combine these three negative datasets $\overline{S}_1$, $\overline{S}_2$ and $\overline{S}_3$ with S_posi and created three new datasets with the same number of positive and negative samples as expressed in Equation (1) below:

$$(S_{posi} \cup \overline{S}_1), \ (S_{posi} \cup \overline{S}_2), \ (S_{posi} \cup \overline{S}_3). \tag{1}$$

The symbol $\cup$ denotes the union of two sets. It is important to note that the dataset under discussion was used for prediction by [1], whereas the positive dataset was collected based on the given three steps: (1) The possible proteins to be acetylated are identified by a single fixed keyword, i.e., {N acetylcysteine, N acetylserine, N acetylglutamate, N acetylglycine, N acetylcholine, N acetylthreonine, N acetylcholine, N acetylmethionine, N acetylmethionylc, N acetyltyrosine, O-acetylserine, or N6-acetyllysine O-acetyltheronine

O}; (2) here, protein collection was validated using some assertion technique; and (3) the 30, or additional, amino acids in proteins, along with the redundant proteins, were removed as discussed in [1].

Whereas the negative data was generated in a similar way to the positive data, except those proteins are not a member being searched by the above keywords? As a result, it produced a large number of negative samples, moreover, random selection is made from those that were balanced in size to positive samples.

4. Feature Extraction

The existing traditional classifiers, such as SVM, KNN, ANN, and many others, are not as powerful in classifying the biological data and making the required prediction. Therefore, a medium is needed to convert biological data into the necessary numerical form to make it suitable for traditional classifiers. For this purpose, many models are developed to extract the required characteristics from biological data, e.g., PseAAC, AAC, Pse-in-One, Pse-Analysis, and many more [27–29]. In feature extraction, the emphasis is on preserving the critical properties of the protein, its location, and functions. The statistical moment [42] is used to derive features in this study, which is discussed in detail below.

4.1. Statistical Moments

In statistics and probability distributions, some form remains beneficial when performing analysis of a particular sequence. The study of such configuration of data collection in pattern matching is known as moments [25]. There are useful moments when there are various pattern recognition problems related to feature development that do not depend on the pattern or sequence parameters provided [27,29–31,43]. Particular moments are used to calculate data size, data alignment, and data eccentricity. In this study, we extract the necessary features of acetylation proteins using Hahn, raw, and central moments. The raw moment is used to estimate the probability distribution by using mean, variance, and asymmetry, these moments are neither location invariant nor scale invariant [32]. Similarly, the same procedure is used in case of the central moment, but the calculation is based on the data centroid. This moment is a scale variant and location invariant. The Hahn moments, on the other hand, are dependent on Hahn polynomials, it is neither scale invariant nor location variant [33,34,44]. These moments are very important for extracting obscure features from protein sequences, as they contain complex orderly details about biotic sequences. In the proposed work, a linearly planned structural of a protein sequence is used, as given in Equation (2).

$$\mathrm{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3 \ldots \mathrm{R}_\mathrm{L}, \tag{2}$$

where R_1 is the 1st amino acid, represented in proteins P, the last amino acid is R_L, and total length is 'L'. Transforming the information of the protein linear structure as given in Equation (2) into 2D matrix representation of dimension k as computed by the following equation.

$$\mathrm{k} = [\sqrt{\mathrm{P}}] \tag{3}$$

where "P" represents the protein sequence length, and k represents the dimension of the obtained 2D square matrix.

Hence the Equation (4), represents the matrix denoted by N' is constructed by using the order obtained from Equation (3) that is $\mathrm{k} \times \mathrm{k}$

$$\mathrm{N}' = \begin{bmatrix} \mathrm{T}_{1\to1} & \mathrm{T}_{1\to2} & \cdots & \mathrm{T}_{1\to j} & \cdots & \mathrm{T}_{1\to k} \\ \mathrm{T}_{2\to1} & \mathrm{T}_{2\to2} & \cdots & \mathrm{T}_{2\to j} & \cdots & \mathrm{T}_{2\to k} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \mathrm{T}_{i\to1} & \mathrm{T}_{i\to2} & \cdots & \mathrm{T}_{i\to j} & \cdots & \mathrm{T}_{i\to k} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathrm{T}_{k\to1} & \mathrm{T}_{k\to2} & \cdots & \mathrm{T}_{k\to j} & \cdots & \mathrm{T}_{k\to k} \end{bmatrix} \tag{4}$$

The raw moment R (a, b) is computed by the values of N′, which is a continuous 2D function of order (a + b), as shown in Equation (5):

$$\mathrm{R_{ab}} = \sum\nolimits_{\mathrm{p}} \cdot \sum\nolimits_{\mathrm{q}} \cdot \mathrm{p^a q^b N'(p, q)}. \tag{5}$$

Up to order 3, the equation calculates the raw moments. This raw moments are measured using the data's roots as a starting point [45–48]. $\mathrm{R_{00}}$, $\mathrm{R_{01}}$, $\mathrm{R_{10}}$, $\mathrm{R_{11}}$, $\mathrm{R_{02}}$, $\mathrm{R_{20}}$, $\mathrm{R_{21}}$, $\mathrm{R_{30}}$, and $\mathrm{R_{03}}$ are the raw moment's characteristics, weighed up to order 3rd.

The centroid is a point from where all points are equivalently dispersed in all directions with a weighted average [45,48–50]. The following equation, which uses the centroid, calculates the special characteristics of central moments up to order 3 (6).

$$\mathrm{C_{ab}} = \sum\nolimits_{\mathrm{p}} \cdot \sum\nolimits_{\mathrm{q}} \cdot (\mathrm{p} - \overline{\mathrm{p}})^{\mathrm{a}} (\mathrm{q} - \overline{\mathrm{q}})^{\mathrm{b}} \mathrm{N'(p, q)}. \tag{6}$$

The unique features are calculated up to 3rd order as: $\mathrm{C_{00}}$, $\mathrm{C_{01}}$, $\mathrm{C_{10}}$, $\mathrm{C_{11}}$, $\mathrm{C_{02}}$, $\mathrm{C_{20}}$, $\mathrm{C_{30}}$, and $\mathrm{C_{31}}$. Further, the centroids are calculated, as given by Equations (7) and (8), as $\overline{\mathrm{p}}$ and $\overline{\mathrm{q}}$.

$$\overline{\mathrm{p}} = \frac{\mathrm{R_{10}}}{\mathrm{R_{00}}}, \tag{7}$$

$$\overline{\mathrm{q}} = \frac{\mathrm{R_{01}}}{\mathrm{R_{00}}}. \tag{8}$$

The Hahn moments must be converted from 1D notation to a square matrix before they can be calculated. Discrete Hahn's moments, also known as 2D moments, necessitate square matrix input data in a 2D structure [51]. Since these moments are orthogonal possess inverse properties, therefore, the construction of original data can be constructed using the inverse discrete Hahn moment. The aforementioned remains observed, and the positional and compositional features are somehow preserved in the measured moments [25,32–34,44,52] Two-dimensional input data in the form N′ is used to calculate the orthogonal Hahn moments, as seen in Equation (9).

$$\mathrm{h_m^{x,y}}\ (\mathrm{r, M}) = (\mathrm{M} + \mathrm{y} - 1)_{\mathrm{m}} (\mathrm{M} - 1)_{\mathrm{m}} \sum\nolimits_{\mathrm{s}=0}^{\mathrm{m}} -1^{\mathrm{s}} \times \frac{(-\mathrm{m})_{\mathrm{s}}\ (-\mathrm{r})_{\mathrm{s}}\ (2\mathrm{M} + \mathrm{x} + \mathrm{y} - \mathrm{m} - 1)_{\mathrm{s}}}{(\mathrm{M} + \mathrm{y} - 1)_{\mathrm{s}} (\mathrm{M} - 1)_{\mathrm{s}}} \cdot \frac{1}{\mathrm{s}!}, \tag{9}$$

where 'p' and 'q' (p > −1, q > −1) controlling the shape of polynomials by using the adjustable parameters. The Pochhammer symbol is defined by Equation (10), as follows:

$$\left(\rho\right)_{\mathrm{s}} = \rho\ \left(\rho + 1\right)\ \ldots\ldots \left(\rho + \mathrm{s} - 1\right). \tag{10}$$

The equation is further simplified by the Gamma operator:

$$\left(\rho\right)_{\mathrm{s}} = \frac{\Gamma\left(\rho + \mathrm{s}\right)}{\rho\left(\rho\right)}. \tag{11}$$

The raw values of Hahn moments are usually scaled using a square norm and weighting formula, as seen in Equation (12):

$$h_m^{x\sim,y}(r,M) = h_m^{x,y}\ (r,M)\sqrt{\frac{\rho\,(r)}{s_m^2}},\ m\ = 0,1,\ldots,M-1. \tag{12}$$

Meanwhile, in Equation (13),

$$\rho\,(r) = \frac{\Gamma\ (p+r+q)\Gamma\ (q+r+1)(p+q+r+1)_M}{(p+q+2r+1)_m\ !(M-r-1)!}. \tag{13}$$

Hahn moments are computed for the discrete 2D data up to the 3rd order through the following, Equation (14):

$$H_{pq}\ \sum_{j=0}^{M-1} \cdot \sum_{i=0}^{M-1} \cdot N'_{i,j}\ h_p^{x\sim,y}(j,M)\ h_q^{x\sim,y}(i,M)\ ,\ p,\ q\ = 0,1,\ldots,M-1. \tag{14}$$

The special features based on the Hahn moments are represented by H_{00}, H_{01}, H_{10}, H_{11}, H_{02}, H_{20}, H_{12}, H_{21}, H_{30}, and H_{03}. For each protein sequence up to the third order, we produced 10 central, 10 raw, and 10 Hahn moments and added them to the miscellaneous Super Feature Vector at random (SFV).

4.2. Position Relative Incident Matrix (PRIM)

The amino acids' order and location in a protein sequence have crucial importance for the recognition of protein characteristics [47,50,53]. In any protein sequence, the relative position of an amino acid remains an essential pattern for understanding its physical properties. The Position Relative Incident Matrix (PRIM) uses a square matrix of order 20 to depict the relative location of amino acids in protein sequences, which is expressed by Equation (15):

$$N_{PRIM} = \begin{bmatrix} O_{1\to1} & O_{1\to2} & \cdots & O_{1\to j} & \cdots & O_{1\to20} \\ O_{2\to1} & O_{2\to2} & \cdots & O_{2\to j} & \cdots & O_{2\to20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ O_{i\to1} & O_{i\to2} & \cdots & O_{i\to j} & \cdots & O_{i\to20} \\ \vdots & \vdots & & \vdots & & \vdots \\ O_{k\to1} & O_{k\to2} & \cdots & O_{k\to j} & \cdots & O_{k\to20} \end{bmatrix} \tag{15}$$

$O_{i\to j}$ represents the position of the jth amino acid for the first occurrence of the ith amino acid in the chain.

The score is of biological evolutional process accomplished by amino-acid of type 'J'. The matrix, N_{PRIM} has 400 coefficients based on the relative position of amino acids occurrence.

Ten central moments, 10 raw moments, and 10 Hahn moments are calculated using the 2D N_{PRIM} and 30 additional special features randomly applied to the miscellaneous SFV.

4.3. Reverse Position Relative Incident Matrix (R-PRIM)

There are several instances of cell biology where biochemical sequences are homologous in origin. This normally occurs where a single ancestor is involved in the evolution process, and several sequences are derived from it. In such situations, using these homologous sequences has a significant impact on the classifier's output. For the purpose of obtaining correct results, successful and efficient sequence similarity searching is carried out. In machine learning, efficiency and accuracy are urgently needed for the preciseness of feature extraction algorithms through which the most relevant features are extracted from biological data [43,47,50,53].

The methods used in R-PRIM and PRIM computations are the same, but R-PRIM is only useful for reverse protein sequence ordering. The R-PRIM computations revealed hidden trends in the data and removed ambiguities between homologous sequences. R-PRIM was created as a 20 × 20 matrix containing 400 hundred coefficients, as seen in Equation (16):

$$N_{R-PRIM} = \begin{bmatrix} B_{1\to 1} & B_{1\to 2} & \cdots & B_{1\to j} & \cdots & B_{1\to 20} \\ B_{2\to 1} & B_{2\to 2} & \cdots & B_{2\to j} & \cdots & B_{2\to 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ B_{i\to 1} & B_{i\to 2} & \cdots & B_{i\to j} & \cdots & B_{i\to 20} \\ \vdots & \vdots & & \vdots & & \vdots \\ B_{k\to 1} & B_{k\to 2} & \cdots & B_{k\to j} & \cdots & B_{k\to 20} \end{bmatrix} \quad (16)$$

The $N_{\text{R-PRIM}}$ 2D matrix is used to measure 10 raw, 10 central, and 10 Hahn moments up to 3rd order, as well as more than 30 special features that are randomly applied to the SFV range.

4.4. Frequency Distribution Vector (FDV)

A frequency distribution vector (FDV) can be generated by using the distribution rate of each amino acid in a protein chain, as expressed in Equation (17).

$$\mu = \{a_i, a_2, a_3, a_4, \ldots\ldots\ldots\ldots\ldots a_{20}\}. \quad (17)$$

Here a_i is the occurrence of frequency of ith ($1 \leq i \leq 20$) amino acid in each protein chain. Twenty more special functions have been randomly added to the SFV's miscellany.

4.5. AAPIV (Accumulative Absolute Position Incidence Vector)

The AAPIV was used to retrieve relevant amino acid positional information, which retrieves and stores amino acid positional information for 20 native amino acids in a protein sequence [50,53]. This creates 20 critical features associated with each amino acid in a sequence, as expressed by Equation (18). These 20 new features were thrown into the SFV at random.

$$\text{AAPIV} = \{\mu_1, \mu_2, \mu_3, \ldots, \mu_{20}\}, \quad (18)$$

where Φ_i is expressed by Equation (19).

$$\mu_i = \sum\nolimits_{x=1}^{n} R_x. \quad (19)$$

The μ_i comes from the protein sequence R_x, which has a cumulative amino acid count of 'n', which can be determined using Equation (19).

4.6. R-AAPIV (Reverse Accumulative Absolute Position Incidence Vector)

R-AAPIV uses reverse sequence ordering to extract and store positional information of amino acids with respect to 20 native amino acids in a protein sequence, which is in reverse order relative to AAPIV [50,53].

This creates 20 critical features associated with each amino, as expressed by Equation (20).

$$R - \text{AAPIV} = \{\Phi_i, \Phi_2, \Phi_3, \ldots\ldots\ldots\ldots, \ \Phi_{20}\}, \quad (20)$$

where Φ_i is expressed by Equation (21).

$$\Phi_i = \sum_{x=1}^{n} \text{Reverse } (R)_x. \quad (21)$$

where $R_1, R_2, R_3, \ldots R_n$ are the ordinal locations at which the residue of protein sequence occurs in the reverse sequence? The values of an arbitrary element of Φ_i are given by Equation (21).

5. Machine Learning Classifiers

The Random Forest and Probabilistic Neural Network are used as a training model to predict acetylation and non-acetylation sites in this research. The following sections go into these classification algorithms in greater depth.

5.1. Probabilistic Neural Network

In this paper, we used a Probabilistic Neural Network (PNN) as a classifier. The Probabilistic Neural Network (PNN) is a powerful algorithm that is mostly employed for classification problems. PNN was first introduced by Specht in 1990 [54]. It is a feed-forward neural network that works on the principle of Kernel Fisher Discriminant Analysis. PNN uses the probability density function and generates better prediction results compared to other neural network algorithms [55,56].

The PNN operates on four layers, i.e., (1) input layer, (2) hidden layer, (3) pattern layer/summation layer, and (4) output layer [57]. The input layer accepts the training samples as input; further, the input data is forwarded to the hidden layer to operate the computational functions. In the hidden layer, an individual node is a computational unit that has some weighted input connections. The third layer received the probability results along with its given classes [58]. The output layer takes the decision and assigns the respective class label to the unknown sample. The schematic view of the PNN is shown in Figure 3.

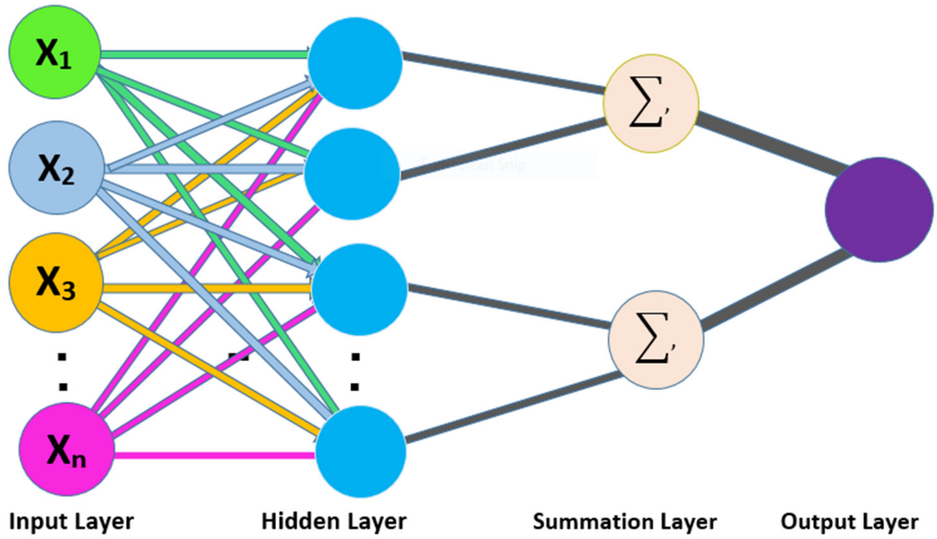

Figure 3. Schematic view of PNN.

5.2. Random Forest

In this paper, we used Random Forest (RF) as a classifier [59]. Random Forest (RF) is an ensemble learning method for classification and regression problems that is widely used [60–62]. RF, as its name implies, contains a large number of individual decision trees that operate as an ensemble. Each individual tree in the Random Forest splits the dataset into training and testing subsets. The training subset is used for training the model, and the testing subset is employed for testing the prediction performance of the trained model classification. During the classification, the class label is assigned to the testing sample by receiving the majority votes of all trees. The variation or bias of a single tree has little effect on the overall prediction accuracy because of the majority voting principle. RF also implements the concept of the weight model by providing a weight value that is low when a particular tree consumes a high error rate. Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random Forest has nearly the same hyper parameters as a decision tree or a bagging classifier. Random Forest adds additional randomness to the model while growing the trees [59]. The working of the Random Forest algorithm is illustrated by the following procedure.

The first step starts with the selection of random samples from a given dataset; in the second step, this algorithm will construct a decision tree for every sample, and, in the last step, voting will be performed for every predicted result [63].

We tested the proposed model performance using the different numbers of trees and found the best results when the number of trees is 500. The RF offers several advantages, including its optimal accuracy, it works efficiently with large datasets, and its detection of outliers. Many researchers used RF for solving several biological problems and achieve better performance. The working mechanism of the RF is depicted in Figure 4.

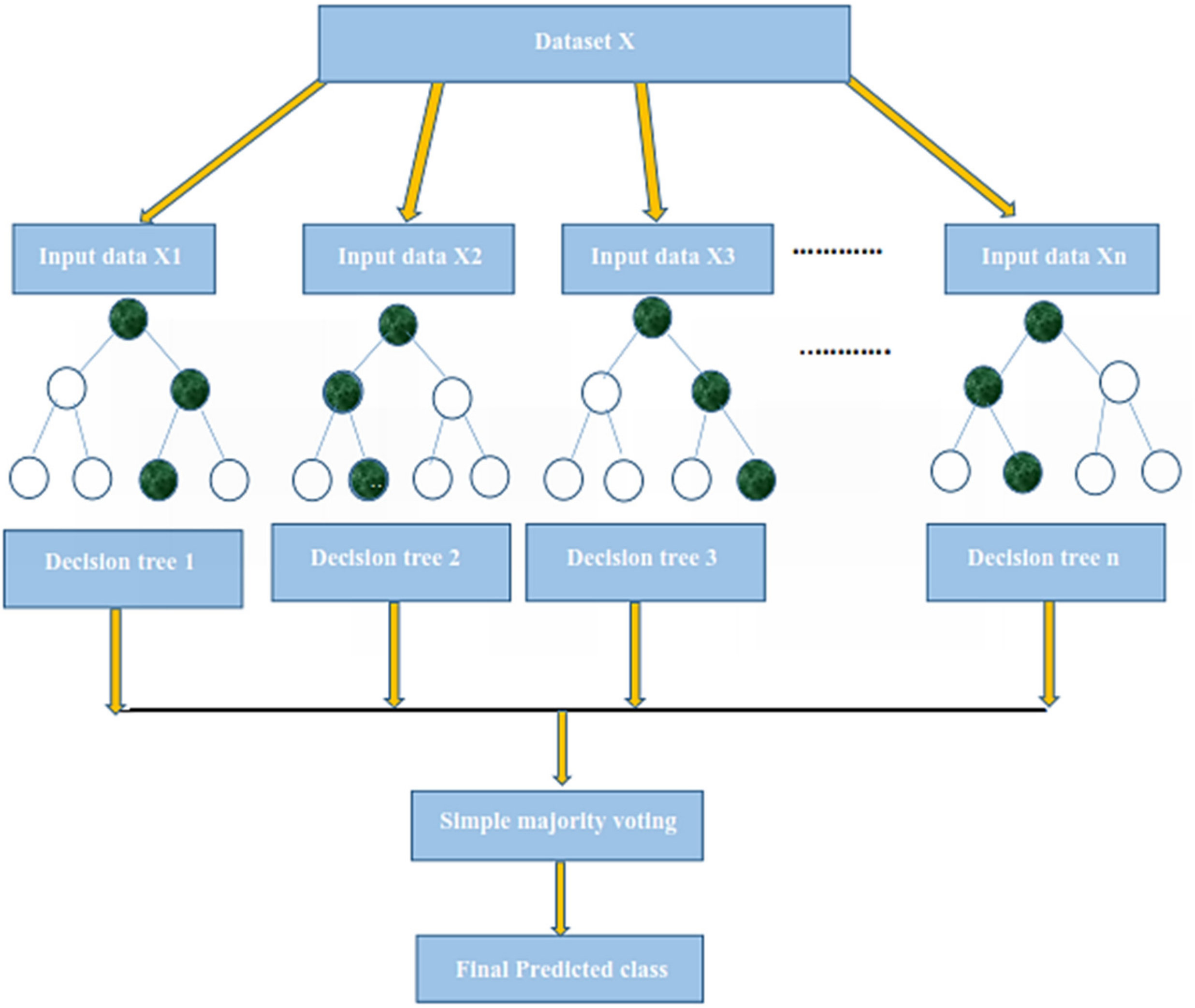

Figure 4. The working mechanism of the Random Forest classifier.

6. Performance Evaluation Parameters and Testing Methods

6.1. Performance Evaluation Parameters

The performance of the proposed model can be measured by effective evaluation metrics. We use the subsequent four metrics to measure the forecast quality: (1) Overall Accuracy (ACC), (2) Sensitivity (Sn), (3) Specificity (Sp), and (4) Mathews Correlation Coefficient (MCC). We computed these parameters using a binary confusion matrix. These metrics remain the most common metrics used to measure efficiency of the proposed model. We computed these parameters using the following equations.

$$\begin{cases} \mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{S_n} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{S_p} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \\ \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ \mathrm{F.measure} = 2\times\frac{\mathrm{Recall}\times\mathrm{Precision}}{\mathrm{Recall}+\mathrm{Precision}} \end{cases} \tag{22}$$

where TP is True Positive, FN is False Negative, TN is True Negative, and FP is False Positive, respectively.

As per the given confusion matrix as shown in the Table 1, we subsequently calculate the following:

Table 1. Confusion matrix.

Status Person	Predicted Patient (1)	Predicted Healthy Person (0)
Actual patient (1)	TP	FN
Actual healthy person (0)	FP	TN

TP: Production prognosis, such as True Positive (TP), where we found that acetylation subject stays properly categorized, as well as classified, then the subjects have acetylation proteins.

TN: Production forecast, such as True Negative (TN), where we found that an non-acetylation protein remains properly classified, and then the subject remains non-acetylation protein.

FP: Production prognosis, such as false positive (FP), where we found that non-acetylation protein remains inaccurately classified as containing acetylation proteins known as "type 1 error".

FN: Forecast of production, such as false negative (FN), where we found that acetylation proteins remain inaccurately classified and that the subject has non- acetylation proteins, this is the "type 2 error".

ROC and AUC: The optimistic receiver curves evaluate the predictability of the machine learning classifiers at various threshold settings. The ROC exam remains a graphical demonstration that relates the "true positive rate" to the "false positive rate" in the grouping results of the machine learning algorithm. AUC describes a classifier of the ROC. The higher value of the AUC more than 0.5, suggest discrimination, whereas the value of 0.5 doesn't suggest any discrimination in true positive and true negative of classifier, more the AUC value more the efficiency in the performance of a classifier.

6.2. Testing Methods

Several cross-validation techniques have been used to examine the statistical forecaster's results in the literature. The jackknife test, independent dataset test, and k-fold cross-validation test are three experiments that are commonly used by various researchers.

When testing a forecaster designed for its efficiency, we use the following cross-validation methods in this paper to estimate the expected accuracy of the forecaster, self-consistency, independent, K-fold cross-validation, and jackknife testing for the assessment of the proposed model.

The following sub-sections contain the details.

6.3. K-Fold (10-Fold) Cross-Validation Test

The K-fold cross-validation (KFCV) test is a technique to estimate predictive models by partitioning the original dataset A into disjoint k-folds $\{A_1, A_2, A_3, A_i \ldots, A_k\}$, where it uses A-$A_i$ folds for training, and the remaining ith fold A_i for testing, where i = 1, 2, 3, … k as shown in the Figure 5. The method iterates the process for each i, and calculate the performance that is the accuracy, sensitivity, precession, recall, F-Measure, and MCC. Further, for the overall result, the average is taken of all the iterations performed for each fold. This technique has many benefits, such as the fact that it validates the model based on multiple datasets to reduce the bias and reach to a stable evaluation that how the model performs. This technique is much more powerful compared to other cross-validation techniques. In literature, the K-fold method is quite popular for k = 10 and 5. In this research, a 10-fold cross-validation is used: the overall result obtained is 100% with random forest classifier as presented in Table 2, and through ROC curves shown in Figure 6.

The results obtained from three dataset using 10-folds Cross validation with the Random forest classifier, and achieved the best result with ACC 100%, MCC, Sn, Precision, and F-measure all are 1.

We also used a 10-fold cross-validation test to evaluate the PNN-based model on the three datasets, and obtained the ACC 66.83, MCC 0.36, Sn 0.72, Precision 0.65, and F-measure 0.72 as presented in the Table 3 and by the ROC curve in the Figure 7.

6.4. Jackknife Test

Several cross-validation tests are extensively useful to estimate the performance of the statistical predictors. Amongst these, the jackknife test is considered to be the supreme in being consistent and reliable. Consequently, the jackknife test is comprehensively applied by researchers to estimate the performance of the predictor model. In this test, if the dataset has N records in the dataset, then it trains the model for $N - 1$ records and tests the model for the remaining one record, which is why it is also called leave-one-out cross-validation. Further, this process is repeated N-times, and the label of each record is predicted. Finally, we accumulate all the results to make the overall prediction based on accuracy, sensitivity, precession, recall, F-measure, and MCC.

In present research work, the jackknife test is used to measure the performance of models by using the classifiers Random Forest and PNN, and we achieved the result of 100% through RF but got 66.87 through PNN, as presented in Tables 4 and 5, and by the ROC curves in Figures 8 and 9.

In this evaluation process, three datasets were used which give the overall accuracy, sensitivity, specificity, precession, MCC, Recall, and F-Measure, given in detail in Table 4.

6.5. Self-Consistency Test

Self-consistency test is a technique referred to as the ultimate test for the validation of efficiency and efficacy of the prediction model, this method uses the same data for both training and testing A representation of these proposed parameters, by conducting the self-consistency testing, the results for the acetylation protein prediction based on the Random Forest classifier as presented in the Table 6 and by the ROC curve is shown in the Figure 10.

Similarly, the evaluation results based on the PNN for the three datasets, S1, S2, and S3, based on RF and PNN as presented in the Table 7 and by the ROC curve as shown in the Figure 11.

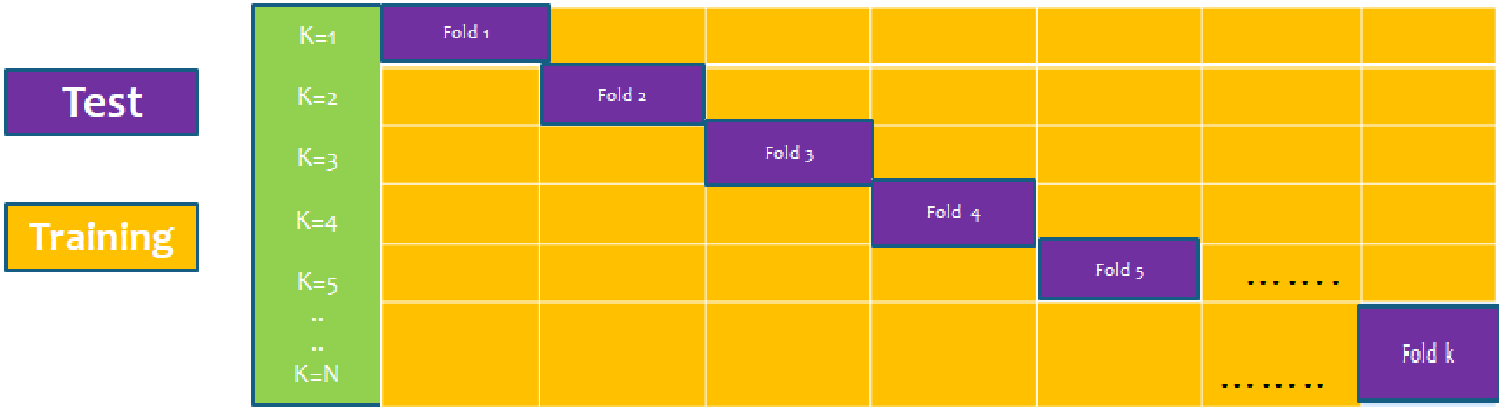

Figure 5. K-fold cross-validation (KFCV).

Table 2. Result of 10-fold cross-validation based on the Random Forest classifier.

10-Fold Cross-Validation Random Forest																					
	Dataset 1							Dataset 2							Dataset 3						
K-Folds	**ACC**	**Sn**	**Sp**	**Prec**	**MCC**	**Recall**	**F.m**	**ACC**	**Sn**	**Sp**	**Prec**	**MCC**	**Recall**	**F.m**	**ACC**	**Sn**	**Sp**	**Prec**	**MCC**	**Recall**	**F.m**
1	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
2	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
3	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
4	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
5	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
6	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
7	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
8	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
9	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
10	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1
result	100	1	1	1	1	1	1	100	1	1	1	1	1	1	100	1	1	1	1	1	1

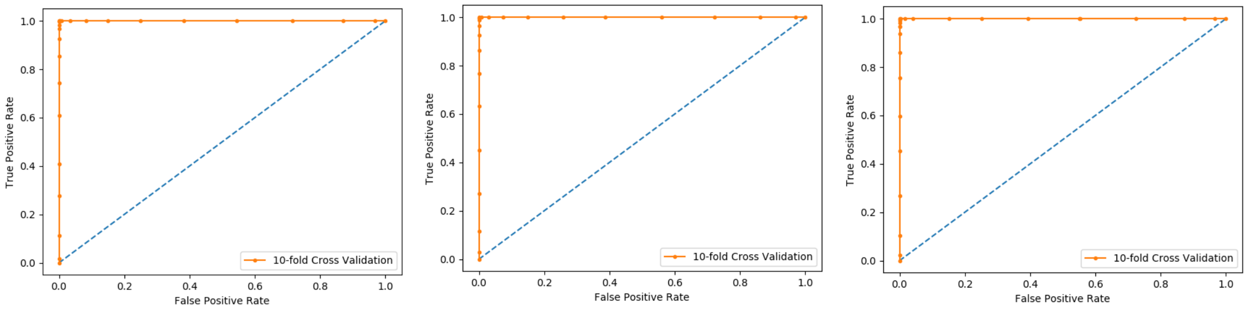

Figure 6. 10-fold Random Forest ROC curve.

Table 3. 10-fold cross-validation Result for Probabilistic neural network.

10-Fold Cross-Validation Probabilistic Neural Network																					
	Dataset 1							**Dataset 2**							**Dataset 3**						
K-Folds	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
Final Score	66.83	0.72	0.60	0.65	0.36	0.95	0.72	60	0.26	0.93	0.81	0.21	0.26	0.40	57.17	0.92	0.22	0.54	0.36	0.92	0.72

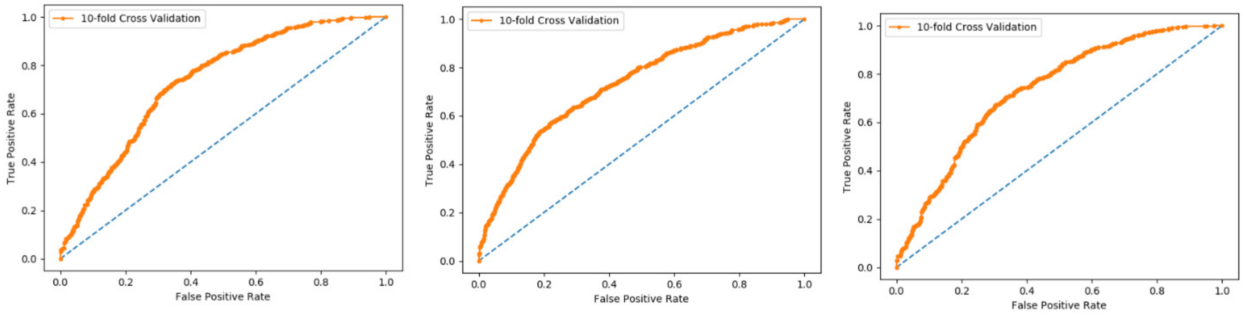

Figure 7. Ten-fold Probabilistic Neural Network ROC curve.

Table 4. Jackknife test score based on Random Forest.

Predicton	**Jackknife Random Forest**																				
	Dataset 1							**Dataset 2**							**Dataset 3**						
Fold	ACC	Sn	Sp	Pre	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
Result	100	0.55	0.5	0.55	0.05	0.003	0.01	99.86	0.55	0.5	0.55	0.05	0.54931	0.55	99.86	0.55	0.5	0.55	0.05	0.54931	0.5

Table 5. Jackknife score based on the Probabilistic Neural Network.

Prediction	**Jackknife Probabilistic Neural Network**																				
	Dataset 1							**Dataset 2**							**Dataset 3**						
Jackknife	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
Final Score	66.87	0.55	0.5	0.5	0.6	0.41	0.42	59.77	o.5	0.5	0.5	0.6	0.13	0.20	57.41	0.55	0.4	0.5	0.6	0.50	0.48

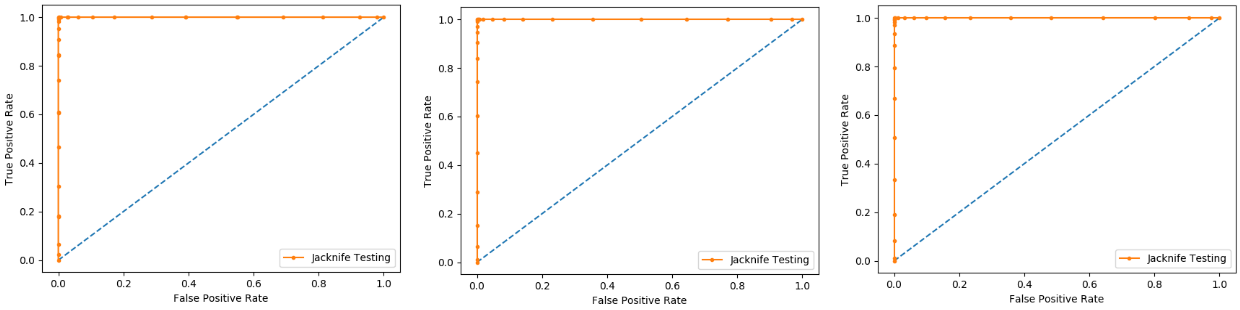

Figure 8. Jackknife Random Forest ROC curve.

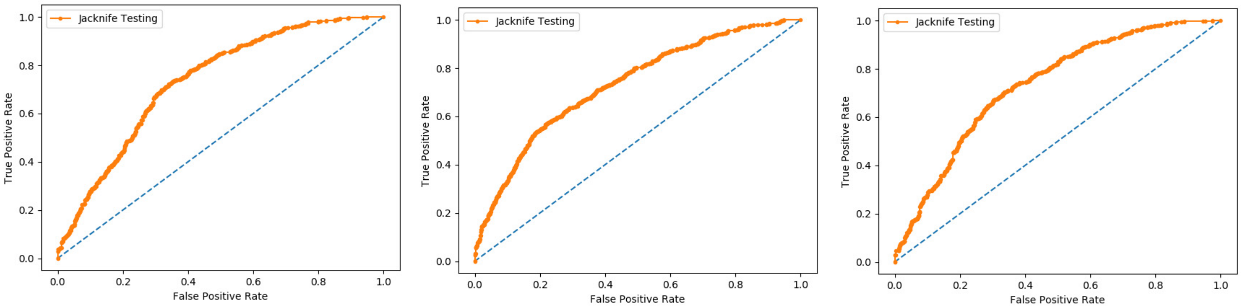

Figure 9. Jackknife PNN ROC curve.

Table 6. Results of self-consistency based on Random Forest.

Self-Consistency Random Forest																				
Dataset 1							Dataset 2							Dataset 3						
ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
100	1	0.99	1	0.99	0.997	1	100	1	1	1	0.99	0.997	1	100	1	1	1	1	1	1

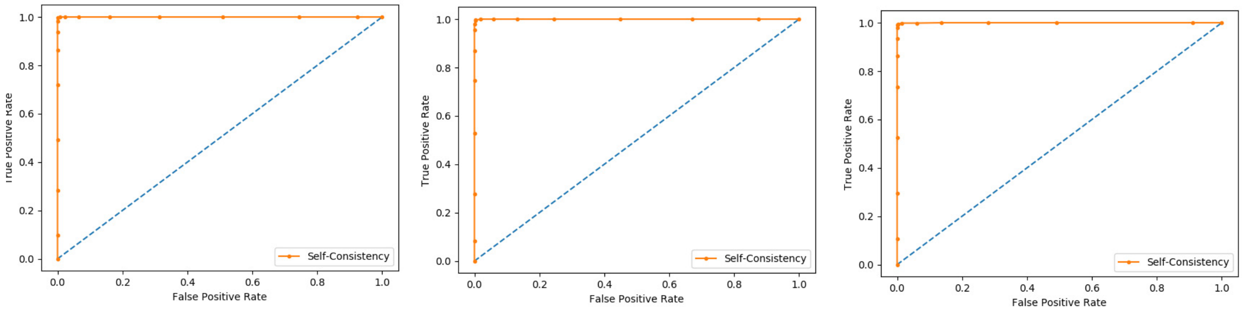

Figure 10. Self- consistency Random Forest ROC curve.

Table 7. Self-consistency test result for probabilistic neural network.

Self-Consistency Probabilistic Neural Network																				
Dataset 1							Dataset 2							Dataset 3						
ACC	Sn	Sp	Prec	MCC	Recall	F.m	AC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
66.83	0.72	0.60	0.64	0.36	0.72	0.68	60	0.26	0.93	0.80	0.20	0.26	0.39	57.17	0.92	0.22	0.54	0.36	0.92	1.84

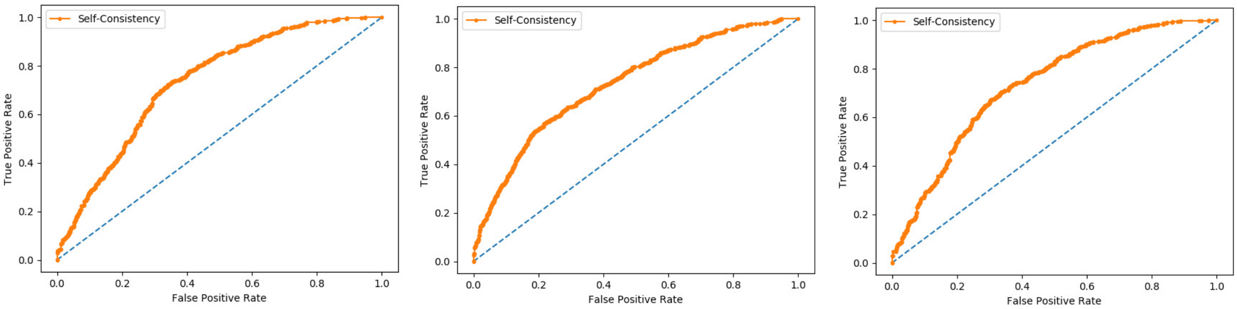

Figure 11. Self- consistency Probabilistic Neural Network ROC curve.

6.6. Independent Test

An independent test is a cross-validation method that objectively finds out the predictor's performance metrics of the planned model, which obtains values from a confusion matrix to evaluate the performance of the model. In this method, the dataset is divided into two parts, training, and the testing part. In the proposed work the data is split in two parts that is 70% of the data for the training and the remaining 30% for the testing as shown in Figure 12. The method used to train and test the models based on the classifiers, the Random Forest and the PNN, and we obtained the result of 98% through RF, and 50.8

through PNN. The area under the curve (AUC), obtained by Random Forest and PNN, is 98% and 50.8%, respectively. The remaining detailed results, based on the two classifiers, is presented in Tables 8 and 9 and by the ROC curve in Figures 13 and 14, respectively.

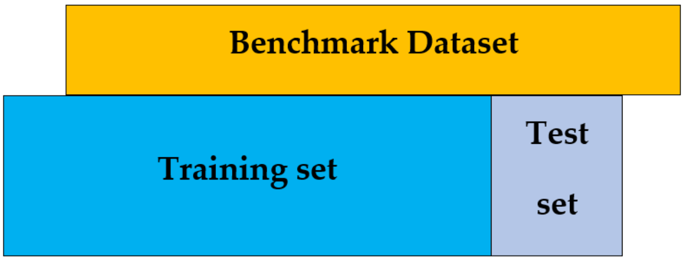

Figure 12. Independent test.

Table 8. Results of independent test based on the Random Forest.

Independent Test Results Random Forest																				
Dataset 1							**Dataset 2**							**Dataset 3**						
Training Dataset							Training Dataset							Training Dataset						
ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
97	1	1	1	1	0.969	1	96	1	1	1	1	0.97	1	96	1	1	1	1	0.95	1
Testing Dataset							Testing Dataset							Testing Dataset						
ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
98	1	1	1	1	0.96	1	95	1	1	1	1	0.93	1	97	1	1	1	1	0.96	1

Table 9. Result of independent test based on Probabilistic Neural Network.

Independent Test Result Probabilistic Neural Network																			
Dataset 1						**Dataset 2**							**Dataset 3**						
Independent Training Dataset Confusion Matrix						**Independent Training Dataset Confusion Matrix**							**Independent Training Dataset Confusion Matrix**						
ACC	Sn	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
50.8	0.1	1	0.6	0.115	0.20	52.6	0.7	0.3	0.53	0.05	0.75	0.62	51.33	1	0.1	1	0.5	0.97	0.98
Independent Testing Dataset Confusion Matrix						**Independent Testing Dataset Confusion Matrix**							**Independent Testing Dataset Confusion Matrix**						
ACC	Sn	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
50.8	1	1	0.6	0.92	0.96	54.0	0.2	0.9	0.54	0.06	0.19	2.86	51.03	1	0.9	1	0.6	0.07	0.13

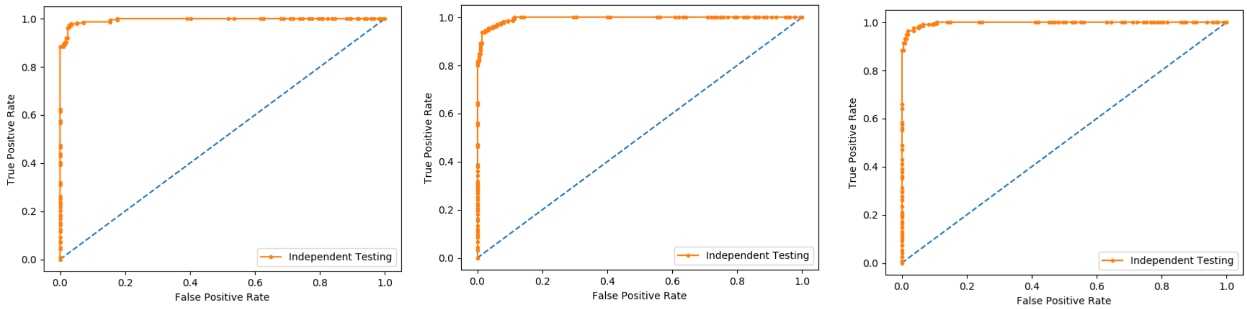

Figure 13. Independent Random Forest ROC curve.

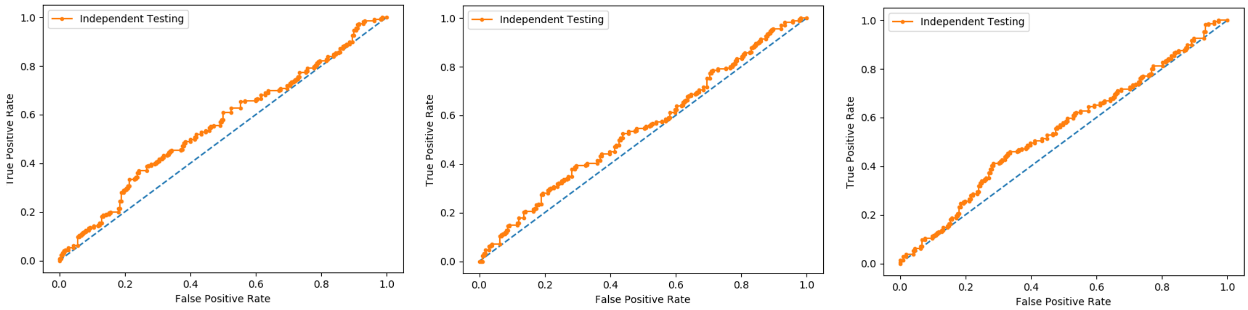

Figure 14. Independent Probabilistic Neural Network ROC curve.

7. Results and Discussions

The most critical thing is to compare the proposed novel model to other state-of-the-art models in order to assess its prediction accuracy. When compared to well-known current classifiers, the RF and the PNN. In this work, the model with RF achieved significantly higher accuracy and efficiency in predicting acetylation from non-acetylation, as seen in Table 10 and by Figures 15 and 16.

Table 10. Performance of proposed model based on RF and PNN.

Prediction	**Comparative Analysis of RF and PNN**																				
	Dataset 1							**Dataset 2**							**Dataset 3**						
Classifier	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m	ACC	Sn	Sp	Prec	MCC	Recall	F.m
RF	100	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
PNN	66.83	0.72	0.6	0.65	0.36	0.95	0.72	60	0.26	0.93	0.81	0.21	0.26	0.40	57.17	0.92	0.22	0.54	0.36	0.92	0.72

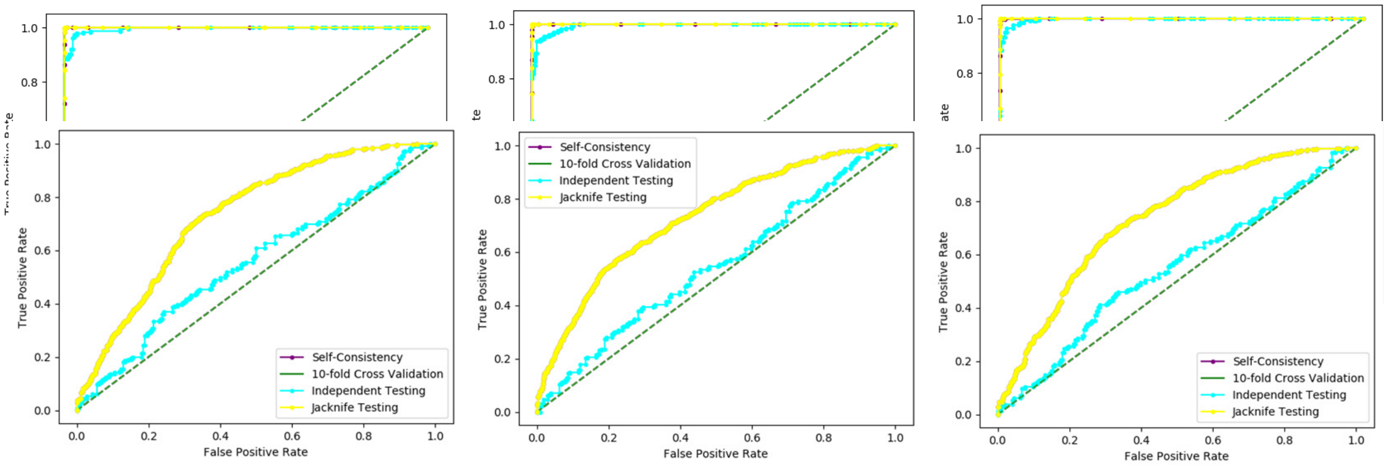

Figure 16. ROC curves through Probabilistic Neural Network.

Comparison to Existing Models

To show the efficiency of our proposed model, iAcety–SmRF, which achieved the highest accuracy of 100% with Sensitivity, Precision, and MCC, is 1, as shown in Section 4.2 (B) (i), using a 10-fold cross-validation. The proposed model was compared to several models, including the latest iAcet-PseFDA by Reference [1]; they used the same dataset and developed a method for predicting acetylation proteins by extracting features from sequence conservation information using a gray frame model and an ANN score based on information from the annotation of the functional domains and subcellular localization. For model validation, they used a 5-fold cross-validation for all three datasets and achieved an average accuracy of 77.10%. The All-Mean JK model by Nakai and Horton [36] achieved an accuracy of 74.64%. Hunter constructed a predictive model InterPro, with accuracy of 68.25% reference [64]. Table 11 lists all comparative analyses of the proposed study, based on our two models iAcety-SmRF and iAcety-SmPNN, in which iAcety-SmRF achieved the superior accuracy as compared to all existing models.

Table 11. Comparative analysis of the proposed acetylation model with the existing models.

Prediction Models	ACC%	MCC%	Sn%	Sp%	Prec%	F.m%
All-Mean JK	74.64%	0.4980	81.38%	67.91%	71.78%	76.24%
iAcet-PseFDA	77.55%	0.5883	96.41%	71.26%	52.79%	68.23%
InterPro	68.25%	0.3658	71.40%	65.10%	67.17%	69.22%
iAcety–SmRF	100	1.0	1.0	1.0	1.0	1.0
iAcety–SmPNN	66.83	0.36	0.72	0. 60	0.65	0.72

8. Conclusions

A computational model for predicting Acetylation sites from non-Acetylation sites is developed in this paper. The model contained Statistical Moments used for extraction of features into the equivalent numerical form of the original biological data. Further, Random Forest and Probabilistic Neural Network were applied for its classification to predict acetylation from non-acetylation. Furthermore, independent testing, 10-fold cross-validation, self-consistency test, and jackknife testing are used to evaluate accuracy, with results of 97%, 100%, and 100%, respectively, based on the Random Forest. Further, the model was compared with the already existing relevant models available in the literature, which revealed the remarkable performance of our work. Finally, the final step is to develop an influential website/GitHub resource for public use for the benefit of future research and development which can be accessed by the following link: https://github.com/shaistarahmanmcs/My-Website-identifying-Proteins-Acetylation-.git (accessed on 8 December 2021).

Author Contributions: Conceptualization, S.M. and S.R.; methodology, O.B.; software, S.R.; validation, S.A.K. and R.A.; formal analysis, investigation, resources, data curation, S.A.K.; writing—original draft preparation, S.R.; writing—review and editing, S.M. and O.B.; supervision, S.A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and classification code is available on the GitHub: https://github.com/shaistarahmanmcs/My-Website-identifying-Proteins-Acetylation-.git (accessed on 8 December 2021).

Acknowledgments: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number IFPHI-359-830-2020 and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qiu, W.-R.; Xu, A.; Xu, Z.-C.; Zhang, C.-H.; Xiao, X. Identifying Acetylation Protein by Fusing Its PseAAC and Functional Domain Annotation. *Front. Bioeng. Biotechnol.* **2019**, *7*, 311. [CrossRef]
2. Chunaram, C.; Chanchal, K.; Florian, G. Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions. *Science* **2009**, *325*, 834–840. [CrossRef]
3. Drazic, A.; Myklebust, L.M.; Ree, R.; Arnesen, T. The world of protein acetylation. *Biochim. Biophys. Acta—Proteins Proteom.* **2016**, *1864*, 1372–1401. [CrossRef] [PubMed]
4. Zhang, K.; Tian, S.; Fan, E. Protein lysine acetylation analysis: Current MS-based proteomic technologies. *Analyst* **2013**, *138*, 1628. [CrossRef] [PubMed]
5. Choudhary, C.; Weinert, B.T.; Nishida, Y.; Verdin, E.; Mann, M. The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 536–550. [CrossRef] [PubMed]
6. Yang, Y.; Rao, R.; Shen, J.; Tang, Y.; Fiskus, W.; Nechtman, J.; Atadja, P.; Bhalla, K. Role of Acetylation and Extracellular Location of Heat Shock Protein 90α in Tumor Cell Invasion. *Cancer Res.* **2008**, *68*, 4833–4842. [CrossRef]
7. Bozelli, J.C.; Kamski-Hennekam, E.; Melacini, G.; Epand, R.M. α-Synuclein and neuronal membranes: Conformational flexibilities in health and disease. *Chem. Phys. Lipids* **2021**, *235*, 105034. [CrossRef]
8. Okada, A.K.; Teranishi, K.; Ambroso, M.R.; Isas, J.M.; Vazquez-Sarandeses, E.; Lee, J.Y.; Melo, A.A.; Pandey, P.; Merken, D.; Berndt, L.; et al. Lysine acetylation regulates the interaction between proteins and membranes. *Nat. Commun.* **2021**, *12*, 6466. [CrossRef]
9. Sundaresan, N.R.; Pillai, V.B.; Wolfgeher, D.; Samant, S.; Vasudevan, P.; Parekh, V.; Raghuraman, H.; Cunningham, J.M.; Gupta, M.; Gupta, M.P. The Deacetylase SIRT1 Promotes Membrane Localization and Activation of Akt and PDK1 During Tumorigenesis and Cardiac Hypertrophy. *Sci. Signal.* **2011**, *4*, ra46. [CrossRef]
10. Fischer, A.; Mühlhäuser, W.W.D.; Warscheid, B.; Radziwill, G. Membrane localization of acetylated CNK1 mediates a positive feedback on RAF/ERK signaling. *Sci. Adv.* **2017**, *3*, e1700475. [CrossRef]
11. Gräff, J.; Tsai, L.-H. Histone acetylation: Molecular mnemonics on the chromatin. *Nat. Rev. Neurosci.* **2013**, *14*, 97–111. [CrossRef]
12. Sadoul, K.; Wang, J.; Diagouraga, B.; Khochbin, S. The Tale of Protein Lysine Acetylation in the Cytoplasm. *J. Biomed. Biotechnol.* **2011**, *2011*, 970382. [CrossRef]
13. Longworth, M.S.; Laimins, L.A. Histone deacetylase 3 localizes to the plasma membrane and is a substrate of Src. *Oncogene* **2006**, *25*, 4495–4500. [CrossRef]

14. Budayeva, H.G.; Cristea, I.M. Human Sirtuin 2 Localization, Transient Interactions, and Impact on the Proteome Point to Its Role in Intracellular Trafficking. *Mol. Cell. Proteom.* **2016**, *15*, 3107–3125. [CrossRef]
15. Zhang, Z.-H.; Wang, Z.-H.; Zhang, Z.-R.; Wang, Y.-X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.* **2006**, *580*, 6169–6174. [CrossRef]
16. Shi, S.-P.; Qiu, J.-D.; Sun, X.-Y.; Suo, S.-B.; Huang, S.-Y.; Liang, R.-P. A method to distinguish between lysine acetylation and lysine methylation from protein sequences. *J. Theor. Biol.* **2012**, *310*, 223–230. [CrossRef]
17. Jiao, Y.-S.; Du, P.-F. Predicting protein sub mitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol.* **2017**, *416*, 81–87. [CrossRef]
18. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef]
19. Liu, B.; Wu, H.; Chou, K.-C. Pse-in-One 2.0: An Improved Package of Web Servers for Generating Various Modes of Pseudo Components of DNA, RNA, and Protein Sequences. *Nat. Sci.* **2017**, *9*, 67–91. [CrossRef]
20. Liu, B.; Liu, F.; Wang, X.; Chen, J.; Fang, L.; Chou, K.-C. Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* **2015**, *43*, W65–W71. [CrossRef]
21. Chou, K.-C. Impacts of Bioinformatics to Medicinal Chemistry. *Med. Chem.* **2015**, *11*, 218–234. [CrossRef]
22. Chou, K.-C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Genet.* **2001**, *43*, 246–255. [CrossRef]
23. Kaur, H.; Raghava, G.P.S. A neural network method for prediction of -turn types in proteins using evolutionary information. *Bioinformatics* **2004**, *20*, 2751–2758. [CrossRef]
24. Chen, Z.; Chen, Y.-Z.; Wang, X.-F.; Wang, C.; Yan, R.-X.; Zhang, Z. Prediction of Ubiquitination Sites by Using the Composition of k-Spaced Amino Acid Pairs. *PLoS ONE* **2011**, *6*, e22930. [CrossRef]
25. Papademetriou, R.C. Reconstructing with moments. *Proc. Int. Conf. Pattern. Recognit.* **1992**, *3*, 476–480.
26. Butt, A.H.; Khan, S.A.; Jamil, H.; Rasool, N.; Khan, Y.D. A Prediction Model for Membrane Proteins Using Moments Based Features. *Biomed. Res. Int.* **2016**, *2016*, 8370132. [CrossRef]
27. Butt, A.H.; Rasool, N.; Khan, Y.D. A Treatise to Computational Approaches Towards Prediction of Membrane Protein and Its Subtypes. *J. Membr. Biol.* **2017**, *250*, 55–76. [CrossRef]
28. Han, Z.-J.; Feng, Y.-H.; Gu, B.-H.; Li, Y.-M.; Chen, H. The post-translational modification, SUMOylation, and cancer (Review). *Int. J. Oncol.* **2018**, *52*, 1081–1094. [CrossRef]
29. Butt, A.H.; Rasool, N.; Khan, Y.D. Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC. *Mol. Biol. Rep.* **2018**, *45*, 2295–2306. [CrossRef]
30. Butt, A.H.; Rasool, N.; Khan, Y.D. Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC. *J. Theor. Biol.* **2019**, *473*, 1–8. [CrossRef]
31. Butt, A.H.; Khan, Y.D. Prediction of S-Sulfenylation Sites Using Statistical Moments Based Features via CHOU'S 5-Step Rule. *Int. J. Pept. Res. Ther.* **2020**, *26*, 1291–1301. [CrossRef]
32. Khan, Y.D.; Khan, S.A.; Ahmad, F.; Islam, S. Iris Recognition Using Image Moments and k-Means Algorithm. *Sci. World J.* **2014**, *2014*, 723595. [CrossRef] [PubMed]
33. Zhu, H.; Shu, H.; Zhou, J.; Luo, L.; Coatrieux, J.L. Image analysis by discrete orthogonal dual Hahn moments. *Pattern Recognit. Lett.* **2007**, *28*, 1688–1704. [CrossRef]
34. Yap, P.-T.; Paramesran, R.; Ong, S.-H. Image Analysis Using Hahn Moments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2057–2062. [CrossRef]
35. Kumar, R.; Panwar, B.; Chauhan, J.S.; Raghava, G.P. Analysis and prediction of cancerlectins using evolutionary and domain information. *BMC Res. Notes* **2011**, *4*, 237. [CrossRef]
36. Harris, M.A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258–D261.
37. Chen, G.; Cao, M.; Luo, K.; Wang, L.; Wen, P.; Shi, S. ProAcePred: Prokaryote lysine acetylation sites prediction based on elastic net feature optimization. *Bioinformatics* **2018**, *34*, 3999–4006. [CrossRef]
38. Wuyun, Q.; Zheng, W.; Zhang, Y.; Ruan, J.; Hu, G. Improved species-specific lysine acetylation site prediction based on a large variety of features set. *PLoS ONE* **2016**, *11*, e0155370. [CrossRef]
39. Hou, T.; Zheng, G.; Zhang, P.; Jia, J.; Li, J.; Xie, L.; Wei, C.; Li, Y. LAceP: Lysine Acetylation Site Prediction Using Logistic Regression Classifiers. *PLoS ONE* **2014**, *9*, e89575. [CrossRef]
40. Li, T.; Du, Y.; Wang, L.; Huang, L.; Li, W.; Lu, M.; Zhang, X.; Zhu, W.G. Characterization and Prediction of Lysine (K)-Acetyl-Transferase Specific Acetylation Sites. *Mol. Cell. Proteom.* **2012**, *11*, M111.011080. [CrossRef]
41. Nawaz, S.; Fatima, K.; Ashraf, A. Prediction of Allergen and Non-Allergen Proteins Sequence via Chou's 5-Step Rule. *VFAST Trans. Softw. Eng.* **2021**, *9*. [CrossRef]
42. Ashraf, A.; Ashraf, R.A.R. A Technique for Prediction Cytokines based On Statistical Moments and a Random Forest Classifier. *VFAST Trans. Softw. Eng.* **2021**, *9*. [CrossRef]
43. Albugami, N. Prediction of Saudi Arabia SARS-COV 2 diversifications in protein strain against China strain. *VAWKUM Trans. Comp. Sci.* **2020**, *8*. [CrossRef]

44. Goh, H.-A.; Chong, C.-W.; Besar, R.; Abas, F.S.; Sim, K.-S. Translation and scale invariants of HAHN moments. *Int. J. Image Graph.* **2009**, *9*, 271–285. [CrossRef]
45. Khan, Y.D.; Rasool, N.; Hussain, W.; Khan, S.A.; Chou, K.-C. IPhosYPseAAC: Identify phosphotyrosine sites by incorporating sequence statistical moments into PseAAC. *Mol. Biol. Rep.* **2018**, *45*, 2501–2509. [CrossRef] [PubMed]
46. Yang, C.-H.; Lin, Y.-D.; Chuang, L.-Y. TRNAfeature: An algorithm for tRNA features to identify tRNA genes in DNA sequences. *J. Theor. Biol.* **2016**, *404*, 251–261. [CrossRef]
47. Akmal, M.A.; Rasool, N.; Khan, Y.D. Prediction of N-linked glycosylation sites using position relative features and statistical moments. *PLoS ONE* **2017**, *12*, e0181966. [CrossRef]
48. Khan, Y.D.; Jamil, M.; Hussain, W.; Rasool, N.; Khan, S.A.; Chou, K.-C. PSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *J. Theor. Biol.* **2019**, *463*, 47–55. [CrossRef]
49. Khan, Y.D.; Batool, A.; Rasool, N.; Khan, S.A.; Chou, K.-C. Prediction of nitrosocysteine sites using position and composition variant features. *Lett. Org. Chem.* **2019**, *16*, 283–293. [CrossRef]
50. Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. SPrenylC–PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *J. Theor. Biol.* **2019**, *468*, 1–11. [CrossRef]
51. Reiss, T.H. Features invariant to linear transformations in 2D and 3D. *Proc. Int. Conf. Pattern Recognit.* **1992**, *3*, 493–496.
52. Pawlak, M.; Liao, X. On image analysis by orthogonal moments. *Proc. Int. Conf. Pattern Recognit.* **1992**, *3*, 549–552.
53. Awais, M.; Hussain, W.; Khan, Y.D.; Rasool, N.; Khan, S.A.; Chou, K.-C. IPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**. *to be published*. [CrossRef] [PubMed]
54. Specht, D.F. Probabilistic neural networks. *Neural Netw.* **1990**, *3*, 109–118. [CrossRef]
55. Khan, Z.U.; Hayat, M.; Khan, M.A. Discrimination of acidic and alkaline enzyme using pseudo amino acid composition in conjunction with probabilistic neural network model. *J. Theor. Biol.* **2014**, *365*, 197–203. [CrossRef]
56. Paliwal, M.; Kumar, U.A. Neural networks and statistical techniques: A review of applications. *Expert Syst. Appl.* **2009**, *36*, 2–17. [CrossRef]
57. Huang, Y.; Li, Y. Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. *Neural Process. Lett.* **2004**, *19*, 211–226. [CrossRef]
58. Hayat, M.; Khan, A. Prediction of membrane protein types by using dipeptide and pseudo amino acid composition based composite features. *IET Commun.* **2012**, *6*, 3257–3264. [CrossRef]
59. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
60. Dai, Q.; Ma, S.; Hai, Y.; Yao, Y.; Liu, X. A segmentation based model for subcellular location prediction of apoptosis protein. *Chemometr. Intell. Lab. Syst.* **2016**, *158*, 146–154. [CrossRef]
61. Kabir, M.; Hayat, M. iRSpot-GAEnsC: Identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. *Mol. Genet Genom.* **2016**, *291*, 285–296. [CrossRef]
62. Farman, A.; Maqsood, H. Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition. *J. Theor. Biol.* **2015**, *384*, 78–83. [CrossRef]
63. Ashraf, A.; Muhammad, S.R.; Muhammad, S.A. Identifying Key Genes of Liver Cancer by Using Random Forest Classification. *VFAST Trans. Softw. Eng.* **2021**.
64. Hunter, S.; Jones, P.; Mitchell, A.; Apweiler, R.; Attwood, T.K.; Bateman, A.; Bernard, T.; Binns, D.; Bork, P.; Burge, S.; et al. InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res.* **2012**, *40*, 4725. [CrossRef]