BMC Bioinformatics

RESEARCH

Open Access

PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features

Salman Khan[1], Salman A. AlQahtani[2], Sumaiya Noor[3] and Nijad Ahmad[4*]

*Correspondence:
Nijad@khurasan.edu.af

[1] Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, KPK, Pakistan
[2] New Emerging Technologies and 5G Network and Beyond Research Chair, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[3] Business and Management Sciences Department, Purdue University, West Lafayette, IN, USA
[4] Department of Computer Science, Khurasan University Jalalabad, Jalalabad, Afghanistan

Abstract

Post-translational modifications (PTMs) are fundamental to essential biological processes, exerting significant influence over gene expression, protein localization, stability, and genome replication. Sumoylation, a PTM involving the covalent addition of a chemical group to a specific protein sequence, profoundly impacts the functional diversity of proteins. Notably, identifying sumoylation sites has garnered significant attention due to their crucial roles in proteomic functions and their implications in various diseases, including Parkinson's and Alzheimer's. Despite the proposal of several computational models for identifying sumoylation sites, their effectiveness could be improved by the limitations associated with conventional learning methodologies. In this study, we introduce pseudo-position-specific scoring matrix (PsePSSM), a robust computational model designed for accurately predicting sumoylation sites using an optimized deep learning algorithm and efficient feature extraction techniques. Moreover, to streamline computational processes and eliminate irrelevant and noisy features, sequential forward selection using a support vector machine (SFS-SVM) is implemented to identify optimal features. The multi-layer Deep Neural Network (DNN) is a robust classifier, facilitating precise sumoylation site prediction. We meticulously assess the performance of PSSM-Sumo through a tenfold cross-validation approach, employing various statistical metrics such as the Matthews Correlation Coefficient (MCC), accuracy, sensitivity, specificity, and the Area under the ROC Curve (AUC). Comparative analyses reveal that PSSM-Sumo achieves an exceptional average prediction accuracy of 98.71%, surpassing existing models. The robustness and accuracy of the proposed model position it as a promising tool for advancing drug discovery and the diagnosis of diverse diseases linked to sumoylation sites.

Keywords: Post-translation modification, Sumoylation, Pseudo position-specific score matrix, Sequential forward selection, Deep neural network

Introduction

The PTM process is highly critical for the regulation of protein function. It is fundamental for protein initiation, protection, and maintenance; loci identification also plays a role in genomic replication. Some PTM processes include sumoylation, wherein

the SUMO proteins are covalently conjugated to the target protein lysine residues [1, 2]. This modification has been linked with controlling cellular processes such as nucleus-cytoplasmic transport, transcription control, DNA repair, and protein interaction [3]. This is also shown by its participation in the distinctive mechanisms of pathogenesis for neurodegenerative disorders, including Parkinson's and Alzheimer's diseases. These impairing syndromes have been linked with different irregularities in the process of sumoylation, such as misfolded protein aggregation and disruption of homeostasis. The emphasis on the sumoylation sites will provide insights into the modern sphere, including genetics [4–6].

Within the past, a list of 450 individual protein modifications has been found, which includes ubiquitination [7], acetylation [8], phosphorylation [9], and sumoylated as well. These alterations regulate the protein–protein interactions, subcellular localization of the target proteins, their enzymatic activity, and also PTM processes [10]. As far as this protection diversity, a lot about the sumoylation is covered by researchers since they appear most frequently in the regulation of post-translational proteins [11], extending to the amino acid code, recycling and degradation processes inside cells, intercellular protein localization distribution, and control over cell physiology or transport transformers enzyme. More recent studies have revealed a strong relationship between sumoylation sites and numerous diseases as well as disorders, including Parkinson's disease and Alzheimer's. Several bioinformatics tools have been developed for structural retrieval at these sumoylation sites [12–14]. Given the significance of sequential and structural bioinformatics in drug discovery modeling, computational biology has emerged as a vital contributor in this field [15]. Therefore, identifying protein sumoylation sites has significant implications for understanding the fundamental biological processes and even developing therapeutic drugs that can be viable inhibitors against cancer or other related diseases [16].

The importance of sumoylation sites has significantly fuelled computational biology and bioinformatics research efforts towards their prediction and characterization. The current models such as SUMOsp [17], SUMOsp2.0 [18] and GPS-SUMO [19], are based on the group phosphorylation scoring algorithm (GPS) that uses clustered known peptide groups, calculates similarities of peptides to determine closeness among them according in their closest belonging Another type, called SUMO_LDA [20], introduces unique sequence Although it shows good potential, the SUMO_LDA heavily uses linear discriminant analysis (LDA), which has difficulties with high-dimensional feature vectors and struggles to deal with nonlinearly separable classes. In some of the recent studies, Xu et al. [21] and Chen et al. [22] proposed SUMOPre and SUMOHydro models that apply a Support Vector Machine (SVM) used for the classification of the SUMO sites. Analogously, Jia et al. [23] presented the pSumo-CD predictor using a universal PseAAC approach and SVM together. Later, Sharma et al. [24] proposed HseSUMO, employing the half-sphere exposure technique and a decision tree algorithm for sumoylation site prediction. Recently, khan et al. proposed Deep-Sumo [25] a deep learning-based model for predicting sumoylation sites using efficient feature representation and principal component analysis. The model achieves an average accuracy of 96.47%, outperforming existing methods. Nonetheless, while the aforementioned existing models show good potential, they require further improvement. Additionally, most of these models rely

on traditional learning approaches that demand significant human expertise in feature extraction and are limited to linear datasets [26].

In this study, we propose PSSM-Sumo, an advanced computational model that integrates state-of-the-art deep learning methods and efficient feature extraction methods to accurately identify sumoylation sites. The PSSM-Sumo model is designed to overcome the limitations of conventional machine learning approaches by leveraging the power of deep neural networks and innovative feature representation strategies, thereby enhancing the accuracy and reliability of sumoylation site prediction. The model performance through an extensive evaluation applying different metrics in a tenfold cross-validation. From an experimental perspective, the proposed predictor attains a high 98.71% average prediction accuracy rate. Comparative analysis of the recently published predictors shows the superiority of the proposed Deep Sumo, especially in accuracy and other performance metrics. These findings support the marked improvements and efficiency achieved by the proposed model in improving predictive precision compared to the current predictors. The significant contributions of the paper are as follows.

- Development of an intellectual and strong computational model based on a multi-layer DNN model, incorporating automatic weight optimization through a standard learning procedure.
- Utilization of the Pseudo position-specific scoring matrix method, enabling the efficient transformation of peptide sequences into a feature vector.
- Implementation of an efficient feature extraction technique, Sequential forward selection using Support Vector Machine (SFS-SVM), to eliminate noisy and irrelevant features.
- Comprehensive performance evaluation metrics, demonstrating the model's robustness and effectiveness in accurately predicting sumoylation sites.

The remainder of the paper is structured as follows: "Framework model" section explains the materials and methods, which include the benchmark dataset, feature extraction, and classification algorithms. The performance evaluation metrics are presented in "Performance evaluation" section. "Experimental results and analysis" section discusses the experimental findings and provides discussions. Finally, "Conclusions" section includes the paper's conclusion and future work.

Framework model

In this section, we introduces the proposed model's design, as illustrated in Fig. 1. A comprehensive explanation of each component comprising the model is provided to offer a detailed understanding.

Benchmark dataset

In deep learning and bioinformatics, choosing an appropriate training dataset is essential for developing an intelligent prediction model. The choice of benchmark dataset has a major effect on the performance of a computational model. In this study, we use a highly reliable dataset to validate our computational model [27, 28]. Therefore, the selection

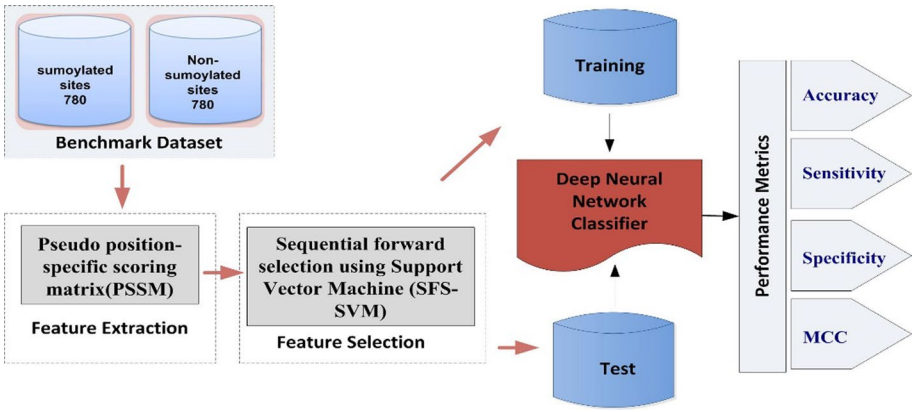

Fig. 1 The framework proposed computational model

is a trustworthy benchmark dataset from [24] for the training and validation of the proposed model. The visualization below presents the benchmark set, demonstrating our commitment to transparency and credibility in building the model.

$$S = S^{P} \cup S^{N} \tag{1}$$

where S represents both sumoylation S^{P} and non-sumoylation S^{N} site sequences. The used dataset is from the Compendium of Protein Lysine Modification (CPLM) [29], a database describing 12 types of lysine PTMs in detail. The dataset comprised 780 sumoylation site sequences (positive examples) and 21,353 non-sumoylation site sequences (i.e., negative samples). Recognizing the imbalance in the dataset as a standard classification challenge is essential. The literature discusses several methods, including data-centric approaches such as oversampling and under sampling, which are very common [30]. To overcome the disparity in our reference dataset, we applied underrepresentation methods with the Near Miss algorithm [31] developed by Python. Under-sampling entails resizing the number of majority class members, such as non-sumoylation sites, to use a balance distribution. Under-sampling, which performs a moderate solution, is important because the over-sample methods that duplicate instances of the minority class result in model over-fitting from repeated data samples. After running the under-sampling method, 780 positive and negative samples are achieved. This balance ensures an appropriate representation of the sumoylation and non-sumoylation sites in our following analyses and model training, making our computational method reliable. To objectively express the model generalizability, one-fifth of the original samples was separated as the testing dataset to perform independent test. To ensure the generalization of the training model, none of the sequences from the training data were repeated in the independent dataset.

Feature formulation method

The majority of predictive machine learning models handle numerical vectors, making it challenging to express a peptide sequence with numerical values or discrete models while preserving sequence information. Feature extraction can address this issue; however, selecting appropriate features is crucial for designing highly accurate predictive models,

as the success of the model depends on the features chosen during training. A notable and complex stage is encoding a discrete model for an input sequence [32]. The model should include features that preserve structural information and key characteristics. While PSSM provides evolutionary information, machine-learning algorithms like SVM, RF, and KNN are limited by variable-length protein sequences. Additionally, PSSM does not consider order information and correlation factors. To solve these problems, PsePSSM is used, which incorporates sequence-order details for calculating residue frequencies. PsePSSM applications include bioinformatics, proteomics, DNA-binding protein systems, and predictions on the structure of non-protein classes.

However, the datasets used in our work are inconsistent with a diverse length of peptide sequences, creating a hurdle in the classifier training process. The dataset used in this work comprises peptide sequences ranging from 4 to 53 amino acids in length. Most sequences have readily available PSSM profiles, while a few require additional parameters such as the number of iterations, and E-value threshold for retrieval. For the remaining short sequences, we augmented the peptide sequence by adding dummy alphabets like hypon (-), which do not correspond to standard amino acids. These hypons (-) have no impact on the function or physical structure of the peptide. Furthermore, we ensured that the extended sequences retained the original peptides without altering their sequence. For example, in the sequence "--- HDEF---", the original "HDEF" remains at the center. In this study, for a protein sequence with length L, PsePSSM has a size L*20, that is formulated via the PSI-BLAST tool, searching the Swiss-Prot database [33]. Hence, Pseudo-PSSM (PsePSSM) generates a uniform vector length from diverse peptide samples. PsePSSM is used for calculating the mean score of each amino acid in the PSSM matrix by determining the correlation between residues separated by 'd' amino acids. The PsePSSM vector for a peptide sample can be represented as:

$$P_{pse} = \left[p_{1,}p_{2,},\ldots,p_{20},p_1^{\varepsilon},p_2^{\varepsilon},\ldots,p_{20}^{\varepsilon}\right]^T \tag{2}$$

$$P_j^{\varepsilon} = \frac{1}{L-\varepsilon}\sum_{i=1}^{L-\varepsilon}\left[P_{i,j} - P_{i+\varepsilon,j}\right]^2,\ \left(j = 1,2,\ldots,20;\ \varepsilon < L\right). \tag{3}$$

Correspondingly, the initial exuberant correlation factor p_j^1 comes from the consecutive amino acid deposits of type j in a given protein sequence using their respective scores obtained through the PSSM. The second subsequent neighbouring PSSM scores are represented as the p_j^2 and so on. The ε defines the vigorous association feature value, which must be less than L straight protein sequence length in the data set.

Feature selection

Feature selection is crucial for achieving optimal classifier performance, as noisy or irrelevant information can significantly affect the results. To address this, we use the SFS-SVM technique, known for its computational efficiency in reducing complexity and improving accuracy. SFS-SVM trains the classifier by introducing features from a qualifying training dataset, starting with an empty feature subset. Superior features are added to the subset through recursive testing. This process identifies the best feature set for a predictive model, minimizing classification errors and computational

time by eliminating redundant or less effective features. The result is a highly effective feature set that significantly enhances the efficiency and accuracy of classification.

Deep architecture

The network topology of a Deep Neural Network, an algorithm based on machine learning or artificial intelligence encouraged by the human brain [34, 35] includes input and output layers as well as multiple hidden layers. The mechanism of neuron Transmission and activation function in DNN as shown in Fig. 2.

Each neuron in the input layer processes a feature x_i and produced output y_k by using a weight vector W_i, bias vector Bi, and activation function fi as shown in Eq. (4 & 5). The output value y of the neurons is feed to the next layer and so forth. This process is continued until it reaches the output layer. In Eq. (4), the weight vector W was calculated using the Xavier initialization [36] method and it optimized using back-propagation and stochastic gradient descent.

$$y_k = f\left(B_i + \sum_{i=1}^{n} x_i w_i^n\right) \tag{4}$$

$$f(i) = \frac{e^i}{1 + e^i} \tag{5}$$

Hidden layers play a really essential role in the learning process, but expanding them can also cause increased computational costs and high model complexity [37, 38]. Contrastingly to traditional processing techniques, DNNs can self-learn and automatically acquire pertinent features from unstructured or raw data. Domains in which DNN has been successfully implemented include speech recognition, NLP and bioengineering, and image [39].

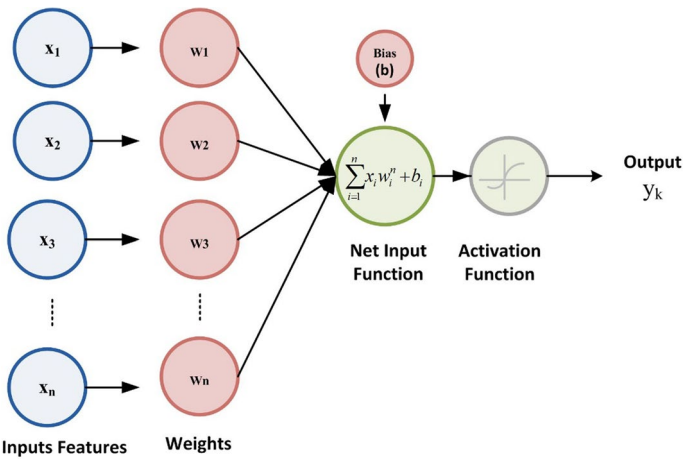

Fig. 2 Mechanism of neuron Transmission and activation function in DNN

Model training

Deep learning algorithms require multiple learning layers to train a complex and non-linear function. As the dataset used in the study consist of biological sequences that are difficult for the classification algorithm to train the model. Therefore, at first, feature encoding methods were used to convert the biological sequence into numerical form to efficiently train the DNN model. The extracted feature vectors are then provided to the DNN model for classification purposes. In the training phase, our DNN model was configured with an input layer, four hidden layers, and an output layer as shown in Fig. 3.

At first, the input layer has multiple neurons that receive input features vector and compute the process and forward it to the first hidden layer. Secondly, the output of the first hidden layer is given as input to the second hidden layer and so forth. This process was continued until we reached the output layer. Finally, the output layer contained a single neuron with a Softmax activation function [40]., which learns the mapping from the hidden layers to the output class labels [0, 1]. The final output either predicts '0' for representing sumoylation sites or a "1" for non-sumoylation sites.

Performance evaluation

Evaluation of the performance metrics is required for any statistical machine learning models before deploying in a natural production environment. While accuracy is critical, more is needed. Various measurements have been proposed, including the SN, SP, ACC, AUC ROC, and MCC. The excellent of metrics depends on the particular applicability and problem domain, according to [41, 42]. These popular metrics research uses them to assess the performance of the proposed Deep Sumo, like other publications. These presentation measurement metrics are calculated as follows:

$$\mathrm{Acc} = 1 - \frac{S_{-}^{+} + S_{+}^{-}}{S^{+} + S^{-}} \tag{6}$$

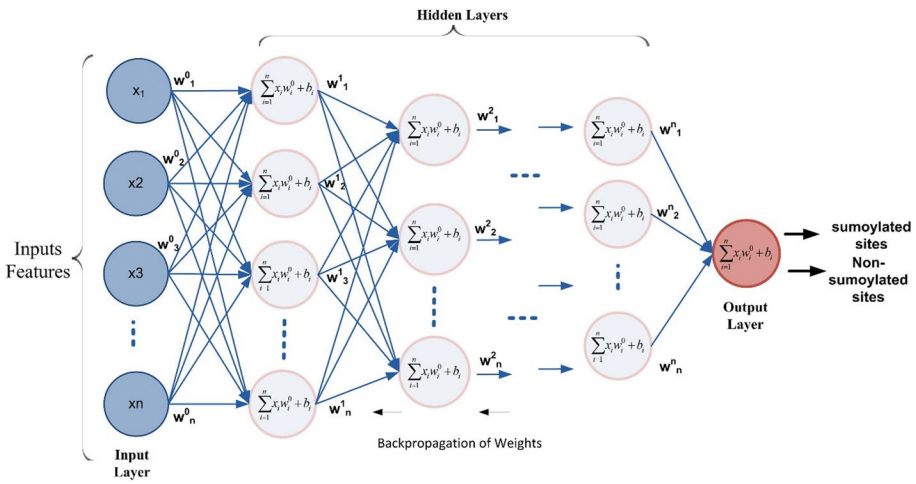

Fig. 3 DNN configuration topology

$$SP = 1 - \frac{S_{-}^{+}}{S^{+}} \tag{7}$$

$$SN = 1 - \frac{S_{+}^{-}}{S^{-}} \tag{8}$$

$$MCC = \frac{1 - \left(\frac{S_{-}^{+} + S_{+}^{-}}{S^{+} + S^{-}}\right)}{\sqrt{(1 + \frac{S_{-}^{+} + S_{+}^{-}}{S^{+}})\left(1 + \frac{S_{-}^{+} + S_{+}^{-}}{S^{-}}\right)}} \tag{9}$$

Experimental results and analysis

Hyper-parameters analysis

The objective of this section is to determine the optimal configuration values for the hyper-parameters employed in the topology of the Deep Neural Network (DNN). The critical hyper-parameters encompass the number of layers and neurons, seed, regularization techniques (L1 & L2), activation function, weight initialization, momentum, dropout, updater, iterations, learning rate, and optimizer, as indicated in Table 1. These parameters significantly influence the performance and behavior of the neural network. For instance, the number of layers and neurons per layer directly impacts the network's learning capacity and its ability to fit the training data. The seed is a predefined starting point for initializing or controlling random processes, ensuring reproducibility in computations or experiments. Regularization techniques like L1 and L2 regularization contribute to preventing over-fitting by introducing penalty terms to the loss function. Activation functions introduce non-linearity into each neuron, while the initialization of weights sets the initial values for the parameters (weights and biases) of the network's neurons before training.

Additionally, momentum enhances the optimization process by incorporating past gradient information to expedite convergence and improve stability during training. Dropout, a regularization technique, randomly drops out a fraction of neurons during

Table 1 DNN model optimum hyper-parameters values

List of parameters	Optimal values
LR	0.1
Weight initialization function	XAVIER function
Seed	12345L
Training iterations	500
Updater	ADAGRAD function
Momentum	0.9
Dropout	0.35
Number of hidden layers	4
Regularization l2	0.001
Activation functions	Tanh & Softmax
Optimizer	SGD method
Neurons at hidden layers	90-60-40-26

each training iteration. The "updater" is responsible for adjusting model parameters, "iteration" represents a single cycle through the training data, "learning rate" controls the size of weight updates, and the "optimizer" serves as the overarching algorithm guiding the optimization process by determining how weights are updated in each iteration. To assess the DNN's performance under various hyper-parameters, a grid search technique was employed, exploring different combinations of parameters. Specifically, the analysis focused on the hyper-parameters that significantly influence the performance of the DNN model, including the activation function, learning rate, and number of iterations.

We conducted experiments to examine the effects of activation functions and LR. The result of the experiments is given in Table 2. From the table, it can be observed that the highest accuracy i.e. 98.71% is obtained by the DNN classifier at a learning rate value of 0.1 using Tanh as an activation function. Furthermore, the DNN model is continuously improved by decreasing the learning rate, however, after reducing the LR (i.e. 0.09 and 0.08), the DNN model accuracy could not significantly be improved. Hence, the DNN model presented a high accuracy at a learning rate 0.1 with the Tanh activation function.

Secondly, we also carried out many experiments to test the performance of DNN with different iteration counts for the model training. Tanh, ReLU, and Sigmoid activation functions results show that after 500 epoch, error losses reach stabilization. In our study, we used two activation functions Tanh, which is used at the hidden layers and Softmax, which is used at the output layer, for predicting the input instance in the Sumoylation or non- Sumoylation site class. The specific optimal parameters for the DNN used in this study are shown in Table 1.

Performance analysis of cross-validation scheme

In computational biology and bioinformatics, statistical learning models experience difficult testing through validation methods like jackknife, k-fold, and sub-sampling. Among these, k-fold cross-validation is particularly prevalent due to its unbiased results. Its systematic approach partitions data, ensuring thorough evaluation and enhancing the reliability of statistical learning models in diverse biological applications. This study analyzed the proposed method's performance using fivefold and tenfold CV

Table 2 Performance comparison of DNN model with different grid search of DNN model

LR	ReLU	Sigmoid	Tanh
0.08	95.01	92.23	98.71
0.09	95.01	92.23	98.71
0.1	95.01	92.23	98.71
0.2	94.71	91.78	97.92
0.3	93.21	90.80	96.47
0.4	92.84	90.17	95.94
0.5	91.94	89.46	94.41
0.6	91.14	88.74	93.95
0.7	90.34	88.03	93.45
0.8	89.54	87.31	92.95
0.9	88.74	86.60	92.45

Table 3 Evaluating the performance of the PSSM-Sumo model via both the feature set and an optimized subset of features

Method	ACC (%)	SP (%)	SN (%)	F1	MCC
PSSM-Sumo (fivefold)	95.91	95.59	96.22	0.915	0.916
PSSM-Sumo (tenfold)	95.94	95.58	96.24	0.916	0.918
	After feature selection				
PSSM-Sumo (fivefold)	98.20	98.81	97.61	0.969	0.972
PSSM-Sumo (tenfold)	98.71	99.68	97.72	0.974	0.974

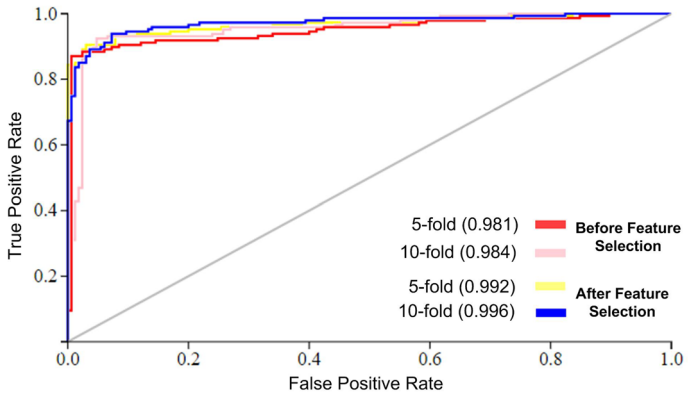

Fig. 4 Comparison of AUC using different cross-validation schemes

tests. Results in Table 3 indicate that the PSSM-Sumo model achieved higher accuracy (95.91%) with a tenfold CV compared to a fivefold CV (95.94%).

The feature vector obtained through the PsePSSM method contained 1,090 features, which may include inappropriate, redundant, and noisy features. To obtain efficient features and reduce the dimensionality, we employed the feature selection method discussed in "Feature selection" section. We reduced the feature vector dimension from 1090×1560 to 120×1560. The evaluation of the proposed model includes the assessment of its performance using both comprehensive and optimized feature sets, ensuring a thorough analysis of its capabilities. The experimental results of this evaluation are shown in Table 3.

Table 3 shows that the proposed model's performance is superior when using an optimized feature set compared to the entire feature set. For instance, using tenfold cross-validation, the proposed model achieved an accuracy of 95.94% with the entire feature set, while it achieved an average accuracy of 98.71% with the optimized feature set. Similar improvements are reported for other performance metrics: specificity (99.68%), sensitivity (97.72%), F1 score (0.974), and MCC (0.974) using the optimized feature set compared to the entire feature set. Given the significance of the optimized feature vector and its prediction results via tenfold testing, we select the optimized vector and DNN classification as our training model.

Moreover, the performance of the PSSM-Sumo model was assessed using the Area Under the Curve (AUC) metric, a measure of the accuracy of binary classifiers, where higher values correspond to improved performance. As depicted in Fig. 4, the

PSSM-Sumo model demonstrated the highest AUC values of 0.996 with tenfold cross-validation and 0.992 with fivefold cross-validation, leveraging the efficient feature set. These outcomes validate the superior predictive capabilities of the proposed model, particularly when utilizing the tenfold cross-validation approach and the selected features. Additionally, a confusion matrix is presented in Fig. 5 to further explore the behavior of the proposed DNN in prediction using the optimized features vector on tenfold.

Performance comparison of different classifiers

In this section, we provide an analysis of the DNN model in comparison to well-known machine learning algorithms such as K-Nearest Neighbor (KNN) [36], Random Forest (RF) [34], and Support Vector Machine (SVM) [25, 42]. The KNN algorithm, often used in image processing, is an instance-based learning technique that identifies instances based on distances [34, 43]. RF is a popular supervised learning method for regression and classification tasks, creating a large number of decision tree models based on random samples using the bootstrap algorithm. This comparative study highlights the specific strengths and weaknesses of each approach. Additionally, the SVM algorithm, known for its effectiveness in bioinformatics, determines an optimal hyper-plane to differentiate groups both linearly and nonlinearly. To ensure a fair assessment of all the learning algorithms, we used a similar effective feature set, standard measurements, and validation methods.

Table 4 presents the performance evaluations of different algorithms. The DNN model performed significantly better than the other models. For instance, the DNN model achieved an average accuracy of 98.71%, while the SVM achieved only 95.32%. Similarly, regarding the MCC criterion for model stability, the DNN achieved a top

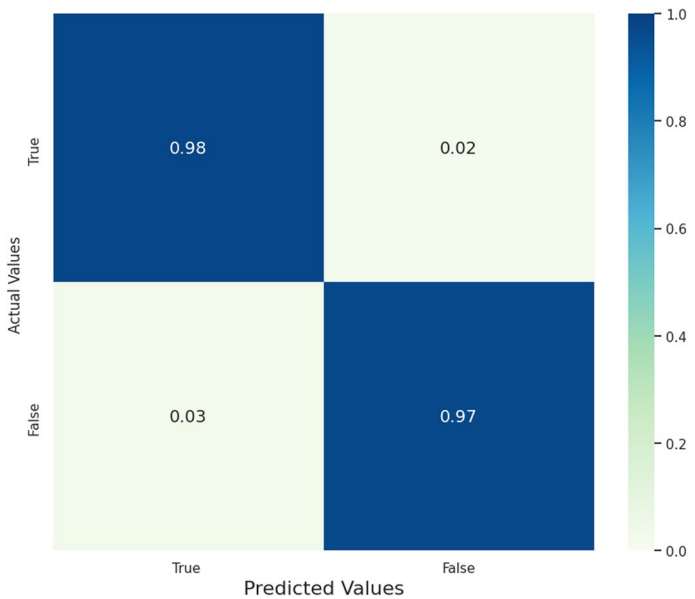

Fig. 5 DNN model Confusion matrix using optimized features

Table 4 A comparison of the proposed model with machine learning models has been considered

ML	ACC (%)	SP (%)	SN (%)	MCC
PSSM-Sumo	98.71	99.68	97.72	0.974
SVM	95.32	96.08	94.62	0.911
RF	95.11	95.82	94.38	0.908
KNN	92.76	93.91	91.52	0.881

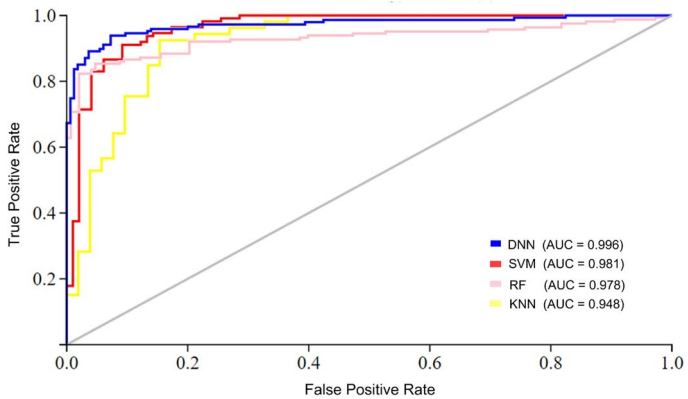

Fig. 6 AUC performance comparison of ML algorithms

rate of 0.974, standing high above the SVM's highest value of 0.911. According to all performance measures, the KNN model performed very poorly.

The analysis findings suggest that our proposed model outperforms traditional learning algorithms. Due to the high similarity between sumoylation and non-sumoylation sites, traditional machine learning algorithms struggle to classify them accurately. These traditional methods often rely on a single processing layer, which may be insufficient for handling nonlinear datasets, significantly affecting their performance. In contrast, our DNN model utilizes multiple hidden layers to perform layer-by-layer sampling on the input data. This layered approach enhances its ability to distinguish between similar sites, leading to superior performance compared to traditional methods. Additionally, Fig. 6 demonstrates a comparison between the performance of the DNN model and traditional learning algorithms using AUC. The figure shows that the proposed DNN model outperforms all other models in terms of obtaining the highest AUC value. For example, the DNN model scored an AUC value of 0.996, while SVM, RF, and KNN algorithms recorded AUC values of 0.981, 0.978, and 0.948, respectively.

Existing models performance comparison

In this section, we compare our proposed model with the existing benchmark methods i.e. [23–25]. The mentioned latest methods build prediction models based on machine learning algorithms. The performance of our proposed model and the existing benchmark models are evaluated on benchmark datasets by using tenfold cross-validation. For facilitating comparison, Table 5 shows the corresponding results obtained by the existing state of the art methods. It can be observed from Table 5 that our proposed PSSM-Sumo model performs overwhelmingly better than the existing

Table 5 Comparison performance of the existing models

Method	ACC (%)	SP (%)	SN (%)	MCC
pSumo-CD [23]	72.80	92.10	53.60	0.494
HseSUMO [24]	89.50	89.50	89.50	0.790
Deep-Sumo [25]	96.47	96.25	96.71	0.929
PSSM-Sumo	98.71	99.68	97.72	0.974

Table 6 A comparison of the proposed model with machine learning models on independent dataset

ML	ACC (%)	SP (%)	SN (%)	MCC
PSSM-Sumo	94.45	92.03	96.87	0.912
SVM	92.53	91.82	93.23	0.861
RF	91.01	89.21	92.81	0.859
KNN	90.74	88.74	92.73	0.853

model. For instance, the proposed model yielded the highest accuracy of 98.71%, and the current predictor (Deep-Sumo) got the second-highest success rate, which is equal to 96.47%. Likewise, the PSSM-Sumo had a 0.974 MCC, which was the highest score achieved and far more significant than Deep-Sumo's result of 0.929. These outcomes emphasize the superior performance of PSSM-Sumo compared to the existing models, with an average success rate increasing up by 10.46%.

Performance analysis of classification learners using an independent dataset

In most cases, the generalization capability of a prediction model is examined using unseen data. Therefore, to test our proposed model we used an independent dataset (i.e. 80% Training and 20% Testing dataset). The performance outcome of the independent dataset is given in Table 6. Among the traditional algorithms, SVM achieved an improved accuracy of 92.53% with sensitivity, specificity, and MCC of 93.23%, 91.82%, and 0.861. On the other hand, DNN obtained higher prediction outcomes with an accuracy of 94.45%, a sensitivity of 96.87%, a specificity of 92.03%, and an MCC of 0.912.

Conclusions

The evaluated PSSM-Sumo demonstrated high reliability in identifying sumoylated sites with superior accuracy. PSSM-Sumo outperformed previous models in this domain, leveraging an optimized deep learning algorithm as an advanced feature extraction technique. The model also proved effective across various statistical measurements, including MCC, accuracy, sensitivity, and specificity. The AUC, on average, showed significantly better results compared to other models, achieving a prediction rate of 98.71%. The model's accuracy was verified through rigorous tenfold cross-validation, indicating strong generalizability in real-world applications. Furthermore, a comparison of PSSM-Sumo with widely used machine learning algorithms such as KNN, RF, and SVM revealed several distinctive strengths, leading to much more accurate predictions of sumoylation sites.

The PSSM-Sumo model also suggested a positive impact on drug discovery and disease diagnosis by accurately identifying sumoylation sites. Future work should address the model's applicability to other biological characteristics and its practical implementation in personalized medicine or specific drug therapies. We also plan to employ parallel programming methods, distributing data across multiple processing nodes using big data analytics platforms, which will significantly enhance the model's scalability and efficiency in handling larger datasets [43, 44].

Acknowledgements
This work was supported by Research Supporting Project Number (RSPD2024R585), King Saud University, Riyadh, Saudi Arabia

Author contributions
All authors contributed equally.

Funding
This research is not funded.

Availability of data and materials
The datasets used and/or analyzed during the current study are available on Github link https://github.com/salman-khan-mrd/Sumo.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have is no competing interests.

Received: 12 July 2024 Accepted: 27 August 2024
Published online: 30 August 2024

References

1. Kessler BM, Edelmann MJ. PTMs in conversation: activity and function of deubiquitinating enzymes regulated via post-translational modifications. Cell Biochem Biophys. 2011;60:21–38.
2. Huber SC, Hardin SC. Numerous posttranslational modifications provide opportunities for the intricate regulation of metabolic enzymes at multiple levels. Curr Opin Plant Biol. 2004;7:318–22.
3. Jensen ON. Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol. 2006;7:391–403.
4. Bao W, Yang B. Protein acetylation sites with complex-valued polynomial model. Front Comput Sci. 2024;18:183904.
5. Bao W, Liu Y, Chen B. Oral_voting_transfer: classification of oral microorganisms' function proteins with voting transfer model. Front Microbiol. 2024;14:1277121.
6. Bao W, Gu Y, Chen B, Yu H. Golgi_DF: Golgi proteins classification with deep forest. Front Neurosci. 2023;17:1197824.
7. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and quantitative assessment of the ubiquitin-modified proteome. Mol Cell. 2011;44:325–40.
8. Drazic A, Myklebust LM, Ree R, Arnesen T. The world of protein acetylation. Biochim Biophys Acta Proteins Proteom. 2016;1864:1372–401.
9. Guo M, Huang BX. Integration of phosphoproteomic, chemical, and biological strategies for the functional analysis of targeted protein phosphorylation. Proteomics. 2013;13:424–37.
10. Verdin E, Ott M. 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. Nat Rev Mol Cell Biol. 2015;16:258–64.
11. Warden SM, Richardson C, O'Donnell J Jr, Stapleton D, Kemp BE, Witters LA. Post-translational modifications of the β-1 subunit of AMP-activated protein kinase affect enzyme activity and cellular localization. Biochem J. 2001;354:275.
12. Lee H, Iqbal N, Chang W, Lee S-Y. A calibration method for eye-gaze estimation systems based on 3D geometrical optics. IEEE Sens J. 2013;13:3219–25.
13. OuYang B, Xie S, Berardi MJ, Zhao X, Dev J, Yu W, et al. Unusual architecture of the p7 channel from hepatitis C virus. Nature. 2013;498:521–5.
14. Oxenoid K, Dong Y, Cao C, Cui T, Sancak Y, Markhard AL, et al. Architecture of the mitochondrial calcium uniporter. Nature. 2016;533:269–73.
15. 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;09:67–91.

16. Bettermann K, Benesch M, Weis S, Haybaeck J. SUMOylation in carcinogenesis. Cancer Lett. 2012;316:113–25.
17. Xue Y, Zhou F, Fu C, Xu Y, Yao X. SUMOsp: a web server for sumoylation site prediction. Nucleic Acids Res. 2006;34 Web Server:W254–7.
18. Ren J, Gao X, Jin C, Zhu M, Wang X, Shaw A, et al. Systematic study of protein sumoylation: development of a site-specific predictor of SUMOsp 2.0. Proteomics. 2009;9:3409–12.
19. Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, et al. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. Nucleic Acids Res. 2014;42:W325–30.
20. Xu Y, Ding Y-X, Deng N-Y, Liu L-M. Prediction of sumoylation sites in proteins using linear discriminant analysis. Gene. 2016;576:99–104.
21. Xu J, He Y, Qiang B, Yuan J, Peng X, Pan X-M. A novel method for high accuracy sumoylation site prediction from protein sequences. BMC Bioinform. 2008;9:8.
22. Chen Y-Z, Chen Z, Gong Y-A, Ying G. SUMOhydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties. PLoS ONE. 2012;7:e39195.
23. Jia J, Zhang L, Liu Z, Xiao X, Chou K-C. pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics. 2016;32:3133–41.
24. Sharma A, Lysenko A, López Y, Dehzangi A, Sharma R, Reddy H, et al. HseSUMO: sumoylation site prediction using half-sphere exposures of amino acids residues. BMC Genom. 2019;19:982.
25. Khan S, Khan M, Iqbal N, Dilshad N, Almufareh MF, Alsubaie N. Enhancing sumoylation site prediction: a deep neural network with discriminative features. Life. 2023;13:2153.
26. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
27. Chou K-C, Shen H-B. REVIEW: recent advances in developing web-servers for predicting protein attributes. Nat Sci. 2009;01:63–92.
28. Chou K-CC. Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol. 2011;273:236–47.
29. Liu Z, Wang Y, Gao T, Pan Z, Cheng H, Yang Q, et al. CPLM: a database of protein lysine modifications. Nucleic Acids Res. 2014;42:D531–6.
30. Kaur P, Gosain A. Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise. In: Advances in intelligent systems and computing; 2018, pp. 23–30.
31. Yen S-J, Lee Y-S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In: Intelligent control and automation. Springer, Berlin; 2006, pp. 731–740.
32. Chou K-C. Pseudo amino acid composition and its applications in bioinformatics. Proteom Syst Biol Curr Proteom. 2009;6:262–74.
33. Waris M, Ahmad K, Kabir M, Hayat M. Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix. Neurocomputing. 2016;199:154–62.
34. Khan S, Khan M, Iqbal N, Amiruddin Abd Rahman M, Khalis Abdul Karim M. Deep-piRNA: bi-layered prediction model for PIWI-interacting RNA using discriminative features. Comput Mater Contin. 2022;72:2243–58.
35. Khan S, Khan M, Iqbal N, Khan SA, Chou K-C. Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC. Chemom Intell Lab Syst. 2020;203:104056.
36. Khan S, Uddin I, Khan M, Iqbal N, Alshanbari HM, Ahmad B, et al. Sequence based model using deep neural network and hybrid features for identification of 5-hydroxymethylcytosine modification. Sci Rep. 2024;14:9116.
37. Wu Y, Tan H, Qin L, Ran B, Jiang Z. A hybrid deep learning based traffic flow prediction method and its understanding. Transp Res Part C Emerg Technol. 2018;90:166–80.
38. Khan S, Khan M, Iqbal N, Hussain T, Khan SA, Chou K-C. A two-level computation model based on deep learning algorithm for identification of piRNA and their functions via Chou's 5-steps rule. Int J Pept Res Ther. 2020;26:795–809.
39. Al-Jumaili MHA, Siddique F, Abul Qais F, Hashem HE, Chtita S, Rani A, et al. Analysis and prediction pathways of natural products and their cytotoxicity against HeLa cell line protein using docking, molecular dynamics and ADMET. J Biomol Struct Dyn. 2023. https://doi.org/10.1080/07391102.2021.2011785.
40. Voisin T, Rouet-Benzineb P, Reuter N, Laburthe M. Orexins and their receptors: structural aspects and role in peripheral tissues. Cell Mol Life Sci. 2003;60:72–87.
41. Baratloo A, Hosseini M, Negida A, El Ashal G. Part 1: simple definition and calculation of accuracy, sensitivity and specificity. Emergency (Tehran, Iran). 2015;3:48–9.
42. Khan S, Naeem M, Qiyas M. Deep intelligent predictive model for the identification of diabetes. AIMS Math. 2023;8:16446–62.
43. Khan S, Khan M, Iqbal N, Li M, Khan DM. Spark-based parallel deep neural network model for classification of large scale RNAs into piRNAs and non-piRNAs. IEEE Access. 2020;8:136978–91.
44. Khan S, Khan MA, Khan M, Iqbal N, AlQahtani SA, Al-Rakhami MS, et al. Optimized feature learning for anti-inflammatory peptide prediction using parallel distributed computing. Appl Sci. 2023;13:7059.

Publisher's Note