Hindawi
Advances in Fuzzy Systems
Volume 2018, Article ID 5125103, 10 pages
https://doi.org/10.1155/2018/5125103

Research Article

Prediction of Ubiquitination Sites Using UbiNets

Sarthak Yadav,[1] Manoj Gupta,[2] and Ankur Singh Bist[1]

[1]*Department of Computer Science and Engineering, KIET, Ghaziabad, India*
[2]*Department of Computer Science and Engineering, SMVDU, Katra, Jammu and Kashmir, India*

Correspondence should be addressed to Ankur Singh Bist; ankur1990bist@gmail.com

Received 27 November 2017; Revised 4 February 2018; Accepted 11 February 2018; Published 20 March 2018

Academic Editor: Mehmet Onder Efe

Ubiquitination controls the activity of various proteins and belongs to posttranslational modification. Various machine learning techniques are taken for prediction of ubiquitination sites in protein sequences. The paper proposes a new MLP architecture, named UbiNets, which is based on Densely Connected Convolutional Neural Networks (DenseNet). Computational machine learning techniques, such as Random Forest Classifier, Gradient Boosting Machines, and Multilayer Perceptrons (MLP), are taken for analysis. The main target of this paper is to explore the significance of deep learning techniques for the prediction of ubiquitination sites in protein sequences. Furthermore, the results obtained show that the newly proposed model provides significant accuracy. Satisfactory experimental results show the efficiency of proposed method for the prediction of ubiquitination sites in protein sequences. Further, it has been recommended that this method can be used to sort out real time problems in concerned domain.

1. Introduction

The discovery of ubiquitin in early 1970s leads to the wining of Nobel Prize in 2004 by Ciechanover et al. [1]. Ubiquitin mediated proteolysis is an important process in regulating protein functions. In this process an unwanted protein would be tagged by an enzyme system with the molecules of residue protein ubiquitin; it is a protein of only 76 amino acids. Ubiquitination role in regulating protein is quite significant and has various biological functions, like quality control, Immune response, signal transduction, DNA repair, metabolism, receptor modulation, Celi cycle, and transcription regulation [2, 3]. Due to vital regulatory role of ubiquitination in protein function and implication in many diseases, a lot of experiments have been performed to identify ubiquitination sites, such as mass spectrometry analysis [4–8], enzymatic approach coupled with the interaction of parallel protein and microarray protein [9], and combinations of multidimensional liquid chromatography and tandem mass spectrometry [10].

Various machine learning techniques have been used for the prediction of ubiquitination site. Deep learning is the emerging form of machine learning method that has been accepted widely in industry as well as in research. The main focus of this paper falls on feed forward artificial neural network of deep nature that is convolutional neural network. After understanding the basic structure of convolution neural network, modified dense convolution network is taken for prediction of ubiquitination sites. Dense convolution neural network joins each layer to every layer in a feed forward manner. The role of deep learning techniques for ubiquitination site prediction is not explored too much in past literature. Now the research of deep network in itself is growing with very fast pace. Our major concern is to explore the significance of deep learning techniques for ubiquitination site prediction.

The contributions of this study includes (1) development of an efficient technique for detecting ubiquitination sites; (2) performance evaluation of different learning models; (3) investigating the impact of proposed learning model for ubiquitination site prediction on different datasets. The organization of paper is as follows: Section 2 gives a complete description on the techniques used by various researchers to predict ubiquitination sites. Section 3 discusses the proposed technique used to detect ubiquitination site. Section 4 discusses the performance analysis of the proposed algorithm

with comparative analysis. Conclusion and future work are given in the final section. The section below discusses the type of prediction techniques used in past literature.

2. Survey

Ubiquitination site prediction using physicochemical property (PCP) is one of the important emerging areas investigated by biology scientist and researchers. Various techniques have been proposed for the prediction of ubiquitination sites [11, 12, 15, 23]. Bayesian networks come under probability based classification method. In the field of biomedical Bayesian networks have been used to sort out critical classification problems [20, 24]. SVM is widely used for classification tasks. The popularity of SVM attracted the attention of researchers in the field of bioinformatics. Past literature [13, 14, 17] shows the effectiveness of SVM in the field of biomedical problems. Regression analysis is one of the conventional techniques in the field of machine learning. Past literature [16, 20, 25] shows the effectiveness of same technique for biomedical problem. Jia et al. [17] used composition of k-spaced amino acid pairs (CKSAAP) to encode samples. Category based feature weighting scheme is used for providing weights to amino acids. Support vector machine is used for classification of CKSAAP dataset. Recent research work has introduced two novel prediction tools for the detection of ubiquitination sites on large-scale proteome data [14, 26]. The UbiProber [14], which combines key position and amino acid residue features, was prepared to expect both normal and species-specific ubiquitination sites. Cross-validation testing has exposed that UbiProber gives better results over existing tools in predicting species-specific ubiquitination sites.

Cai and Jiang [20] used various machine learning algorithms like Bayesian network, naive Bayes, feature selection naive Bayes, model averaged naive Bayes, and efficient Bayesian multivariate classifier. Regression methods are taken for study includes support vector machines, logistic regression, and least absolute shrinkage and selection operator. Fivefold cross validation is used; results showed that efficient Bayesian multivariate classifier obtained higher accuracy score as compared to other algorithms taken for analysis. Ghosh and Dass [18] were able to predict ubiquitination site through homology modelling of NF-kB. At the same time noncanonical pathway network modelling is used. This study predicted the crucial cofactors in the alternate pathway of NF-kB activation. Nguyen et al. [21] used support vector machine to design classification model of ubiquitination site prediction and fivefold cross validation is used. In addition a motif identification tool is taken to find out the motifs of ubiquitination sites. 78.50 accuracy score was obtained with the help of proposed method. Radivojac et al. [12] predicted 141 new ubiquitination sites using mixture of chromatography, mass spectrometry, and mutant yeast strains. Random forest is used as a prediction model for ubiquitination sites. Proposed method obtained accuracy score of 0.72 and area under curve at 80%.

Saeed et al. [19] designed ubipredictor tool for identifying ubiquitinated lysine in protein sequence of various dataset like mouse, human, and yeast using linear discriminant analysis. Tung and Ho [11] established an ubiquitination dataset containing 157 ubiquitination sites and 3676 putative nonubiquitination sites taken from 105 proteins. Support vector machine, K-nearest neighbour, and naive Bayes are used to design prediction model for ubiquitination site identification. In addition authors proposed informative physicochemical property mining algorithm (IPMA). Experimental results showed that IPMA improved accuracy score from 72.19% to 84.44%. Wang et al. [23] used an evolutionary screening algorithm to predict human ubiquitination sites. Results achieved 92% testing accuracy score and the value of Matthews' correlation coefficient is found to be 0.48. Experimental analysis performed by various researchers demonstrates the effectiveness of machine learning techniques for ubiquitination site prediction.

3. Description of the Proposed Approach

The dominance of deep neural networks (DNNs) for machine learning has increased a lot in the past few years due to significant advances in both computer hardware and software. Now it has become possible to train truly deep neural networks for real time datasets.

But merely increasing the depth of the neural network is not enough, as knowledge content about input vanishes after travelling through deep layers. This is one of the emerging problems that need to be resolved for proper execution of deep network. Recent research has geared towards resolving this issue and has given rise to quite a large number of different techniques and architectures. Among the latest of these techniques are Densely Connected Convolution Networks (DenseNet). Densely Connected CNNs ensure maximum knowledge content travel between layers in a deep network.

In the experiments, python is used as programming language to design PCP matrix datasets from different of data sequence, as well as for implementing the multilayer perceptron model (MLP), UbiNet, and random forest for ubiquitination site prediction. General design logic for all the MLP models is comprised of 2 hidden layers with fixed number of neurons in each layer. Each hidden layer is followed by a dropout layer with 0.3 dropout rate. The number of hidden units per layer is given by the first two digits of the name of the model. The last 3-4 characters of the MLP models describe the activation function used. Thus 50x2_mlp_elu is a 2-hidden layer MLP with 50 hidden units per layer and ELU activation function.

We propose UbiNets, inspired by DenseNets. UbiNets are comprised of dense (fully connected layers only) as compared to DenseNets, which are comprised of chained 2D convolution operations. In UbiNet, every layer's output is given as input to every other layer in a block, each block is comprised of 4 layers, and the outputs merge through concatenation. After a block has ended, a dense layer followed by a dropout layer connects the outputs of the previous block to the next UbiNet. This block is henceforth called the

"UbiNet Block." Version 1 (ub_v1) is comprised of single such block. "ub_deeper" has 2 such blocks, while "ub_deeperx2" has 3 such blocks, respectively. Dense_res_v2 has a simple residual network based design and is really shallow. All models were trained with stochastic gradient descent on 10 different subsets of training and validation data. The test data stays the same. For 10 model AVG AUC result, we get predictions on the test set for each of the 10 models and average them. Figure 1 demonstrates a network that incorporated residual skip-connection based design which was originally proposed in [27]. Layer "input_1" is the input layer which represents the input 544-dimensional input tensor x. x is then duplicated and the two independent copies of x flow to the layers "dense_1" and "dense_3," respectively. The output of these layers, are then added in the form of a residual skip connection to produce a single 50 dimensional tensor, followed by "batchnormalization_1" and "dropout_1" layers. The output from these layers thus traverses the usual paths as depicted in Figure 1, where duplicate tensors are passed down parallel paths. "merge_3" layer merges the inputs passed to it in an element-wise sum, representing a skip connection. This helps regulate back propagation of gradients despite the presence of multiple layers, thus facilitating faster convergence and the ability to train truly deep networks (as seen in ResNets).

Figure 2 demonstrates a "50x2_mlp_elu" model. This follows the "*mxn*_mlp_*xyz*" naming convention, where m stands for number of hidden units per hidden layer, n represents the number of such hidden layers in the network, and xyz represents the nonlinearity applied at each hidden layer. Therefore, "50x2_mlp_elu" is a feed forward neural network that has 2 hidden layers with 50 hidden units each and ELU activation functions. "Dense_input_1" denotes the 544-dimensional input. This layer is followed by a 50-hidden layer fully connected layer "dense_1" with ELU nonlinearity, followed up by a dropout layer with 0.3 dropout rate, followed by another dense-elu-dropout combination. The final fully connected layer has a single hidden unit with sigmoid nonlinearity and it outputs the probability of an input sample belonging to a protein sequence that contains an ubiquitination site.

Following the same convention as 50x2_mlp_elu, 50x2_mlp_relu is a 2 hidden layer, 50 hidden nodes per layer feed forward network with ReLU activation applied at each hidden layer, 100x2_mlp_elu is a feed forward network with 2 hidden layers, each having 100 hidden units and ELU activation, and 100x2_mlp_relu is one with ReLU activation.

Figure 3 shows the "ub_deeper_x2" network architecture. "input_1" represents the input layer which feeds the network a matrix of size "p x 544" where p is the batch size. The tensor then flows into "dense_1" fully connected layer, having ReLU activation. Here, four copies of the output of "dense_1" layer are made: x_1, x_2, x_3, and x_4, which are fed to the layers "dense_2," "merge_1," "merge_2," and "merge_3," respectively. Every layer with suffix "merge_" concatenates all the outputs of all the dense layers before it within the same block. Therefore, "merge_1" concatenates the outputs from "dense_1" (i.e., x_2) and "dense_2" layers to produce a 100-dimensional tensor which is then fed to layer "dense_3." "merge_2" concatenates the outputs from "dense_1," "dense_2," and "dense_3" to form a single large tensor of 150 dimensions, the output fed to "dense_4". And the last merge layer of the first UbiNet Block, "merge_3," concatenates the outputs of all dense layers prior to it, namely, "dense_1," "dense_2," "dense_3," and "dense_4" to produce a 200-dimensional output. This finishes the first UbiNet Block of ub_deeper_x2 network. As evident from Figure 3, similar blocks are repeated multiple times over the network graph (3 times to be precise). Each UbiNet Block is succeeded by a fully connected layer, which in turn is succeeded by a dropout layer. It is worth noting that all but the last fully connected layer of the network has ReLU activation, and the dropout ratio is fixed at for all layers within a network, ranging between 0.3 and 0.4. The final fully connected layer has a single hidden node with Sigmoid Activations, the output of which predicts the probability that an input sample has ubiquitination site. The design is highly inspired by the approach proposed in [27], designed to imitate the benefits of the said approach, namely, the ability to train really deep models, alleviating the vanishing gradient problem and improving feature propagation (each layer in a block has direct access to outputs of the previous layer) as well as reducing the number of parameters. The main differences between our approach and DenseNets are as follows:

(1) Our approach focuses on fully connected layers instead of convolutional layers.

(2) Instead of continually connecting all the layers of the network in a DenseNet like fashion, we resort to limiting the skip connections to a particular block. This limits the "growth" of the feature maps to a single block and as soon as the next block starts, the number of hidden nodes is reset.

The reason behind the second difference as stated above is that fully connected layers are more prone to overfitting as compared to convolutional layers, which already possess a much sparser feature map as compared to fully connected layers. Therefore, in order to prevent overfitting the feature maps of the dense layers in the proposed approach grow only within a block. This results in fewer hyperparameters, reduced overfitting, and faster training/testing times. Table 2 gives brief description about classification models used for the experiment.

4. Dataset

A protein is a biological entity that contains sequence of amino acid residues. Protein sequence generally contains 20 distinct amino acids (AAs). Lysine is one of the crucial amino acid for ubiquitination. Data generation process includes the following:

(a) Extraction of sequence segment.

(b) Generation of AA-PCP matrix.

(c) Segment PCP prediction matrix generation.

Standard datasets are taken from past literature. The details of datasets are given below.

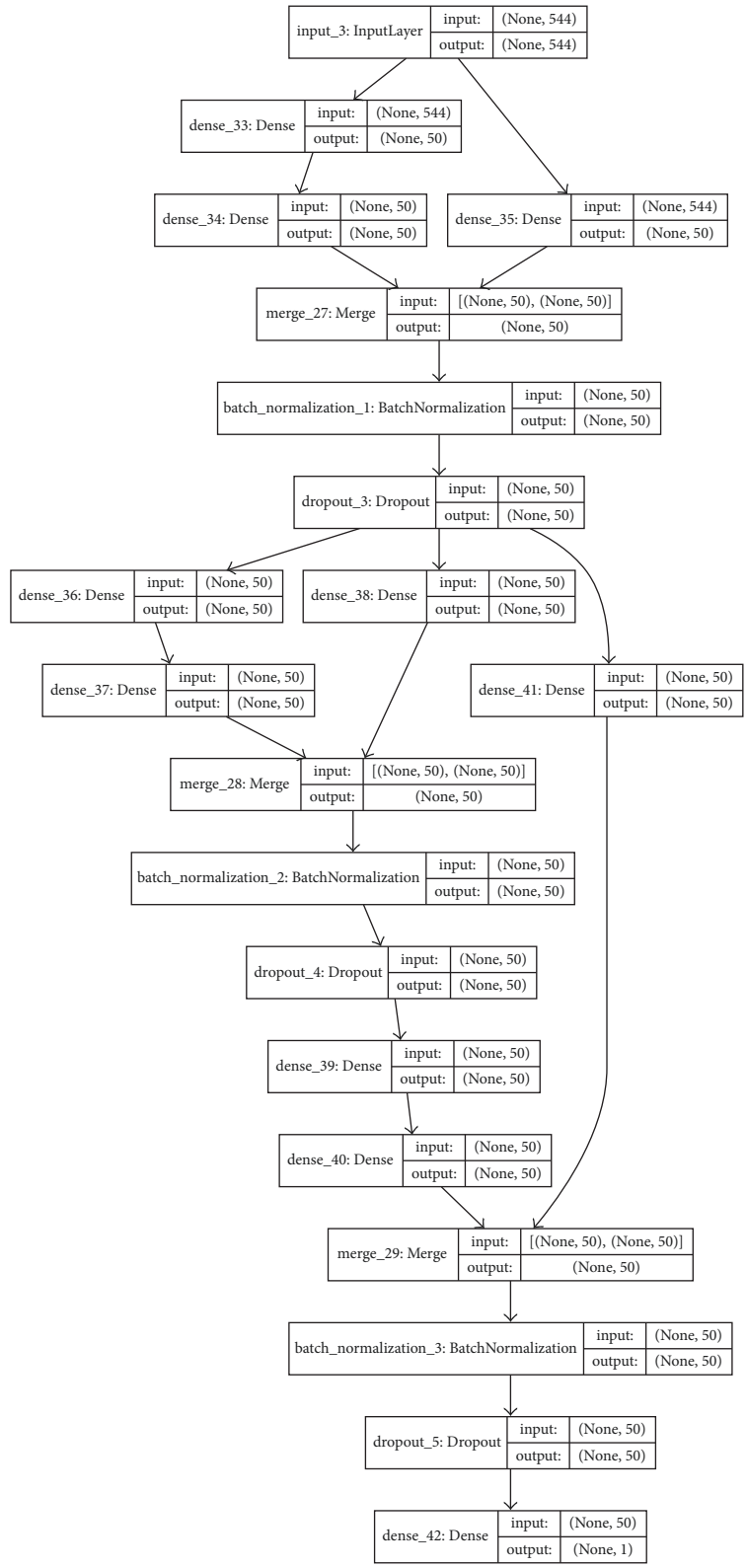

Figure 1: Sample residual skip-connection based network. The "merge_3" layer is the best node of the network for demonstrating the skip connections as adopted from [22]. "merge_3" layer performs an element-wise sum operation on the outputs of "dense_8" and "dense_9", thus effectively allowing easier gradient backpropagation during training.

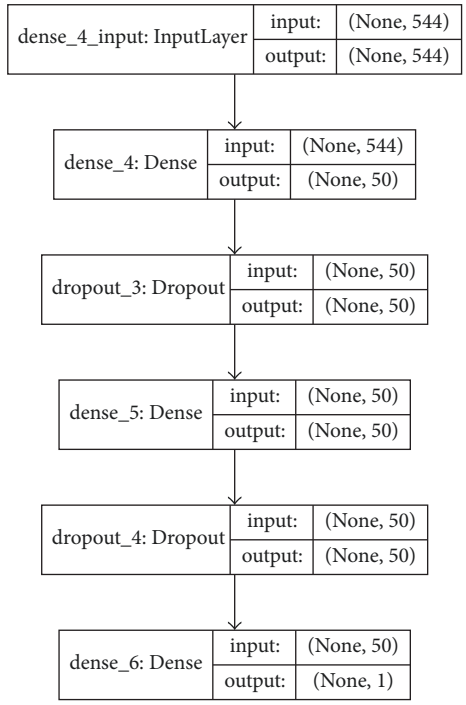

Figure 2: 50x2_mlp_elu model. 2 Hidden layer MLP with dropout layer following each hidden layer.

Dataset 1 is taken from [14] (Independent Set), http://protein.cau.edu.cn/cksaap_ubsite/download/DatasetForhCKSAAP_UbSite.rar, which has half-half ubiquitination and nonubiquitination K sites, and datasets 2 is taken from [13] DOI: 10.1371/journal.pone.0022930.s001 which contains 263 ubiquitination and 4345 nonubiquitination sites. Dataset 3 [12] is taken from http://www.ubpred.org/UbPred_DataSets.zip which contains 131 ubiquitination positive segments and 3520 nonubiquitination ones. PCP prediction matrix is generated from sequence segment matrix and amino acid PCP matrix.

5. Experimental Results

Initially in this section, the experimental results for different models for ubiquitination site prediction are given. Table 3 gives the 10-run average AUC, 10-run average accuracy and 10-run model averaging AUC scores generated by different prediction algorithms. From the table we can see that most of the techniques obtained 10-run average AUC between 0.6831 and 0.7000 for the six segment PCP datasets. This reflects that the segment PCP data has important information that can be utilized for prediction of ubiquitination site.

Experimental analysis demonstrates two important aspects. First aspect is the effectiveness of different machine learning techniques for prediction of ubiquitination site. Second important aspect is the significance of deep learning methods prediction of ubiquitination site. From Table 3 we can find that ub_deeper, 50x2_mlp_elu (Figure 2), 100x2_mlp_elu, 100x2_mlp_relu, random forest (100 estimators), random forest (200 estimators), and Gradient Boosting Machines (gbm) with 100 estimators performed comparable to each other for the six segment PCP datasets, whereas ub_v1, ub_deeper_x2, Dense_residual_v2, and 50x2_mlp_relu showed comparable performance with each other for the six segment PCP datasets. In a word, ub_v1 and ub_deeper_x2 perform better than other machine learning techniques like random forest taken for ubiquitination prediction. Experiment results for different PCP datasets using machine learning techniques are shown in survey section. PCP is very crucial identity of protein sequences for identification of ubiquitination site. PCP averaging process has been taken for the prediction of ubiquitination site. Selection of optimal protein segment length, different PCP summarization techniques and their impacts on prediction, and evolving machine learning techniques for ubiquitination site prediction are some important questions that we will try to answer in our future research.

Experimental results show the effectiveness of 11 machine learning techniques for ubiquitination site prediction using PCP data. Note that Dense_residual_v2 performed slightly better than all other machine learning methods. Figures 4, 5, and 6 explain the variation of accuracy values under different machine learning techniques.

Same experiment is performed for dataset 2. Experimental results are given in Table 4. 11 machine learning techniques applied deep learning methods prediction of ubiquitination site. From Table 2 we can find that ub_deeper, 50x2_mlp_elu, 100x2_mlp_elu, 100x2_mlp_relu, random forest (100 estimators), random forest (200 estimators), and Gradient Boosting Machines (gbm) with 100 estimators performed comparable to each other for the six segment PCP datasets, whereas ub_v1, ub_deeper_x2, and Dense_residual_v2 showed comparable performance with each other for the six segment PCP datasets. In a word, ub_v1, ub_deeper_x2 and Dense_residual_v2 perform better than other machine learning techniques taken for ubiquitination prediction. Among these three best methods Dense_residual_v2 generated best accuracy.

Same experiment is performed for dataset 3. From Table 5 we can find that ub_deeper, 50x2_mlp_elu, 100x2_mlp_elu, random forest (100 estimators), random forest (200 estimators), and Gradient Boosting Machines (gbm) with 100 estimators performed comparable to each other for the six segment PCP datasets, whereas ub_v1, ub_deeper_x2, Dense_residual_v2, and 100x2_mlp_relu showed comparable performance with each other for the six segment PCP datasets. In a word, Dense_residual_v2, 100x2_mlp_relu, ub_v1, and ub_deeper_x2 perform better than other machine learning techniques taken for ubiquitination prediction. Among these three best methods Dense_residual_v2 generated best accuracy.

Thus, Tables 1, 2, and 3 demonstrate the results achieved by the various predictive techniques used for experimentation on dataset 1, dataset 2, and dataset 3, respectively. ub_v1, ub_deeper, and ub_deeperx2 continue to perform better than the majority of the models, though Dense_residual_v2 outperforms all models for dataset 2 and dataset 3. This demonstrates the effectiveness of the proposed techniques. Figures 7, 8, and 9 represent ROC curve for dataset 1,

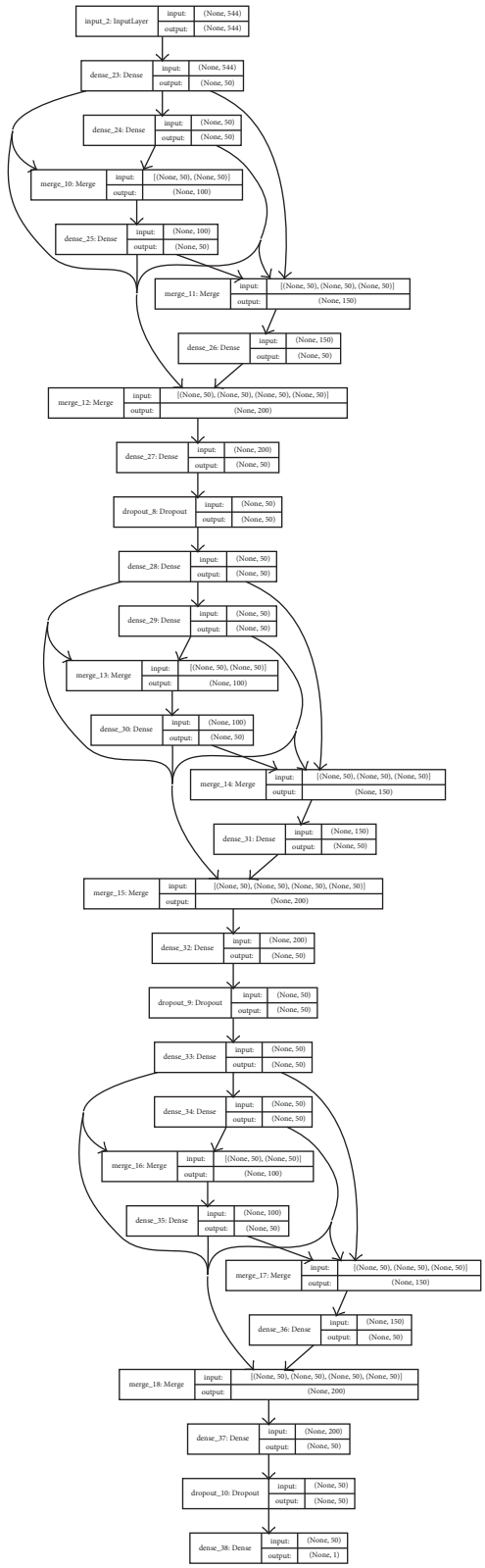

Figure 3: The "ub_deeper_x2" model. This model clearly demonstrates the underlying ideology of how the nodes of the network interact with each other. Zooming in to "merge_8" one can clearly see distinct output paths travelling into the layer. The "merge_8" layer concatenates the outputs of "dense_11," "dense_12," and "dense_13" to form a single large tensor, thus demonstrating how the layers of a block are all directly connected to each other instead of following a more "sequential" flow of execution.

Table 1: Related work in chronological order.

S. number	Year	Title	Technique and results
(1)	2008	Computational identification of ubiquitylation sites from protein sequences [11]	Authors used svm, knn, and naive Bayes for analysis and obtained 84.44% accuracy
(2)	2010	Identification, analysis, and prediction of protein ubiquitination sites [12]	Authors used random forest predictor as classification model and obtained 72% accuracy
(3)	2011	Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs [13]	Authors used SVM as classification model and obtained accuracy of 73.40%
(4)	2013	hCKSAAP_UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties [14]	Authors used SVM as classification model based on the composition of k-spaced amino acid pairs and obtained accuracy of 75.7%
(5)	2014	RUBI: rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance [15]	Authors proposed Rapid UBIquitination (RUBI), a sequence-based ubiquitination predictor, and obtained 86.8% accuracy
(6)	2014	Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor [16]	Authors proposed RA-T prediction model and obtained 44% improvement across multiple machine learning algorithm
(7)	2016	Prediction of ubiquitination sites with feature weighting scheme and naive Bayes vectorizer [17]	Category based feature weighting scheme is used and prediction model. Proposed technique performed better than SVM
(8)	2016	ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives [18]	Authors used evolutionary screening algorithm and obtained testing accuracy 92% and Matthews' correlation 0.48
(9)	2016	Noncanonical pathway network modelling and ubiquitination site prediction through homology modelling of NF-κB [19]	Authors used loop_model and asses_dope functions and enhanced understanding of cofactors involved and ubiquitination sites employed during the activation process
(10)	2016	Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences [20]	Authors used various techniques like SVM and naive Bayes for predictionm and obtained AUC value greater than or equal to 0.6
(11)	2017	A new scheme to characterize and identify protein ubiquitination sites [21]	Authors used SVM as prediction model and obtained 68.70% average accuracy

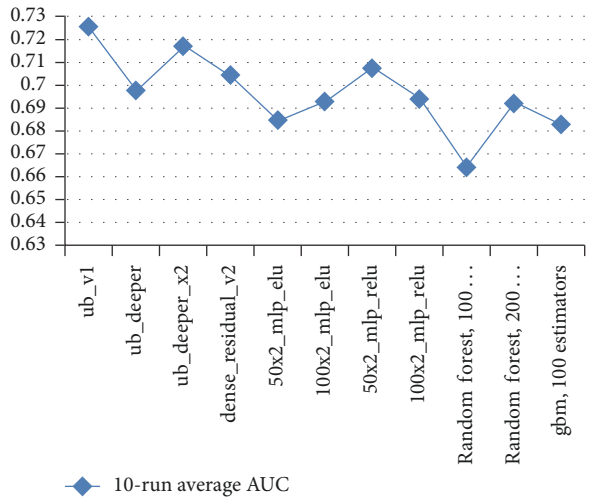

Figure 4: 10-run average AUC.

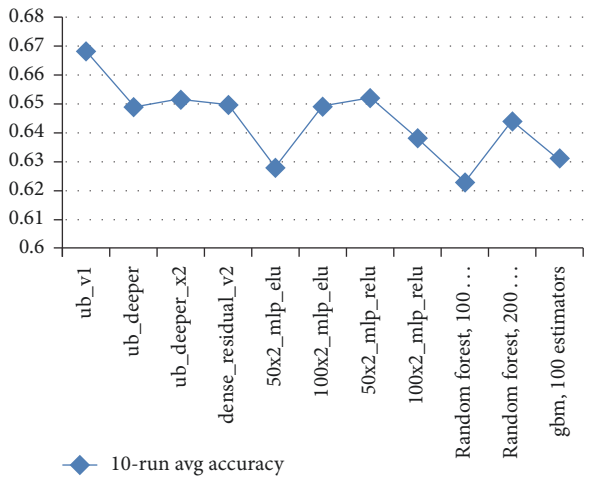

Figure 5: 10-run average accuracy.

dataset 2, and dataset 3, respectively. Dense_residual_v2 generated best accuracy for dataset 1, dataset 2, and dataset 3.

6. Conclusion

Prediction models based on machine learning are employed for the prediction of ubiquitination site depending on

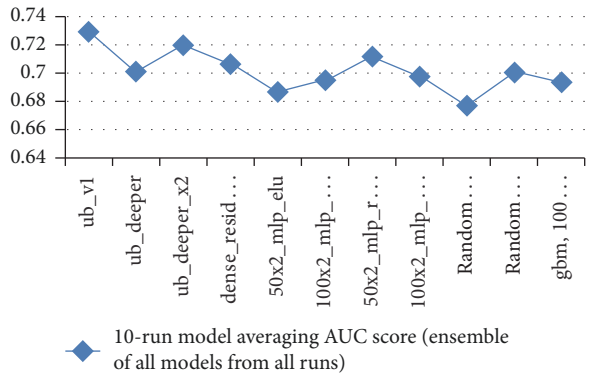

Figure 6: AUC scores of various models (averaged over 10 runs).

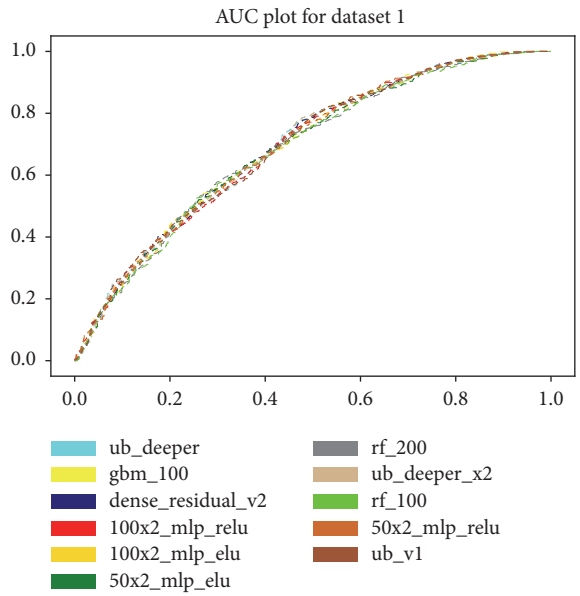

Figure 7: ROC curve for dataset 1.

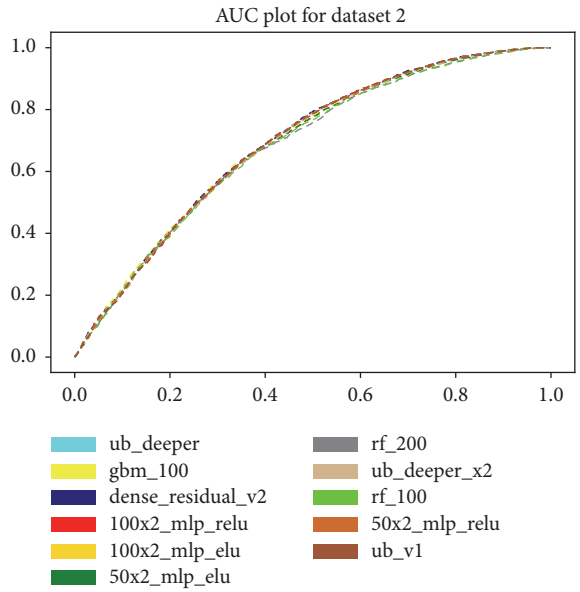

Figure 8: ROC curve for dataset 2.

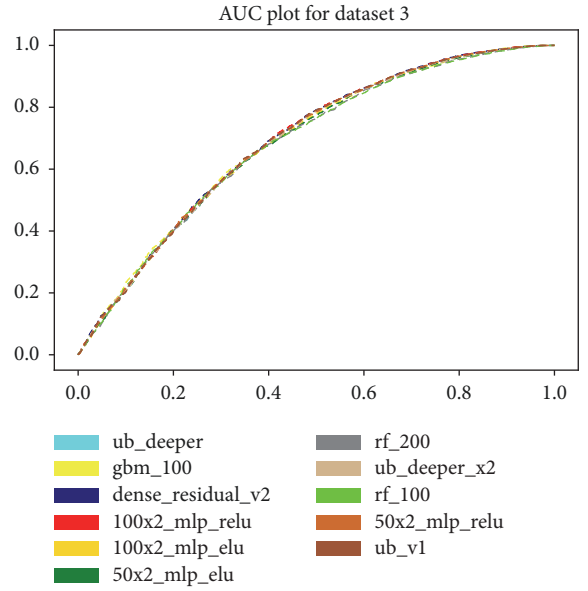

Figure 9: ROC curve for dataset 3.

Table 2: Classification models and their description.

Models	Model description
ub_v1	Is comprised of a single UbiNet Block
ub_deeper	A deeper network, is comprised of 2 UbiNet Blocks
ub_deeper_x2	The deepest UbiNet architecture, is comprised of 3 UbiNet Blocks
Dense_residual_v2	A shallow residual network inspired design
50x2_mlp_elu	2 hidden layer feed forward network with ELU activation and 50 nodes per hidden layer
100x2_mlp_elu	2 hidden layer feed forward network with ELU activation and 100 nodes per hidden layer
50x2_mlp_relu	2 hidden layer feed forward network with ReLU activation and 50 nodes per hidden layer
100x2_mlp_relu	2 hidden layer feed forward network with ReLU activation and 100 nodes per hidden layer
Random forest, 100 estimators	Random forest of 100 metaestimators
Random forest, 200 estimators	Random forest of 200 metaestimators
gbm, 100 estimators	Gradient Boosting Machine is comprised of 100 metaestimators

physicochemical properties of amino acids and proteins. Features from protein sequence segment are computed depending on PCP values from amino acid index database by taking average of physicochemical properties values of all amino acids on various segments. Eleven machine learning techniques are taken for analysis.Comparative results show

Table 3: Classifier performance on dataset 1.

Models	10-run average AUC	10-run average accuracy	10-run model averaging AUC score
ub_v1	0.68927	0.63013	0.69230
ub_deeper	0.68915	0.63255	0.69194
ub_deeper_x2	0.68981	0.63299	0.69200
Dense_residual_v2	0.69142	0.63717	0.69349
50x2_mlp_elu	0.68679	0.63607	0.68818
100x2_mlp_elu	0.68643	0.63226	0.68772
50x2_mlp_relu	0.68871	0.63072	0.69175
100x2_mlp_relu	0.68670	0.63065	0.69017
Random forest, 100 estimators	0.67386	0.62727	0.68425
Random forest, 200 estimators	0.68009	0.62918	0.68678
gbm, 100 estimators	0.68339	0.63233	0.69438

Table 4: Classifier performance on dataset 2.

Models	10-run average AUC	10-run average accuracy	10-run model averaging AUC score
ub_v1	0.69030	0.64450	0.69130
ub_deeper	0.68940	0.64410	0.69030
ub_deeper_x2	0.68810	0.64150	0.68888
Dense_residual_v2	0.69230	0.64430	0.69340
50x2_mlp_elu	0.68660	0.63830	0.68700
100x2_mlp_elu	0.68640	0.63850	0.68700
50x2_mlp_relu	0.68960	0.64360	0.69070
100x2_mlp_relu	0.68940	0.64360	0.69070
Random forest, 100 estimators	0.67425	0.63075	0.68312
Random forest, 200 estimators	0.67773	0.63185	0.68336
gbm, 100 estimators	0.68818	0.63922	0.69269

Table 5: Classifier performance on dataset 3.

Models	10-run average AUC	10-run average accuracy	10-run model averaging AUC score
ub_v1	0.68980	0.64370	0.69120
ub_deeper	0.68880	0.64320	0.68970
ub_deeper_x2	0.68820	0.64250	0.68910
Dense_residual_v2	0.69170	0.64480	0.69270
50x2_mlp_elu	0.68540	0.63970	0.68620
100x2_mlp_elu	0.68650	0.63940	0.68730
50x2_mlp_relu	0.68970	0.63410	0.69080
100x2_mlp_relu	0.69000	0.64480	0.69140
Random forest, 100 estimators	0.67498	0.63050	0.68415
Random forest, 200 estimators	0.67788	0.63362	0.68335
gbm, 100 estimators	0.68790	0.63978	0.69213

that Dense_residual_v2 tends to perform better than other techniques taken for analysis. Use of various other popular and emerging deep learning methods and their impact on other complex datasets in concerned domain is left for future work.

Conflicts of Interest

There are no conflicts of interest.

References

[1] R. J. Mayer, "The nobel prize for chemistry," *European Pharmaceutical Review*, 2004.

[2] P. Ebner, G. A. Versteeg, and F. Ikeda, "Ubiquitin enzymes in the regulation of immune responses," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 52, no. 4, pp. 425–460, 2017.

[3] F. L. Fontes, D. M. L. Pinheiro, A. H. S. D. Oliveira, R. K. D. M. Oliveira, T. B. P. Lajus, and L. F. Agnez-Lima, "Role of DNA

repair in host immune response and inflammation," *Mutation Research—Reviews in Mutation Research*, vol. 763, pp. 246–257, 2015.

[4] W. Zachariae, A. Shevchenko, P. D. Andrews et al., "Mass spectrometric analysis of the anaphase-promoting complex from yeast: Identification of a subunit related to cullins," *Science*, vol. 279, no. 5354, pp. 1216–1219, 1998.

[5] H. Nishikawa, S. Ooka, K. Sato et al., "Mass spectrometric and mutational analyses reveal lys-6-linked polyubiquitin chains catalyzed by BRCA1-BARD1 ubiquitin ligase," *The Journal of Biological Chemistry*, vol. 279, no. 6, pp. 3916–3924, 2004.

[6] J.-S. Lee, U.-S. Hong, T. H. Lee, K. Y. Sungjoo, and J.-B. Yoon, "Mass spectrometric analysis of tumor necrosis factor receptor-associated factor 1 ubiquitination mediated by cellular inhibitor of apoptosis 2," *Proteomics*, vol. 4, no. 11, pp. 3376–3382, 2004.

[7] D. S. Kirkpatrick, C. Denison, and S. P. Gygi, "Weighing in on ubiquitin: the expanding role of mass-spectrometry-based proteomics," *Nature Cell Biology*, vol. 7, no. 8, pp. 750–757, 2005.

[8] S. A. Wagner, P. Beli, B. T. Weinert et al., "A proteome-wide, quantitative survey of in vivo ubiquitylation sites reveals widespread regulatory roles.," *Molecular & cellular proteomics : MCP*, vol. 10, no. 10, p. M111.013284, 2011.

[9] R. Gupta, B. Kus, C. Fladd et al., "Ubiquitination screen using protein microarrays for comprehensive identification of Rsp5 substrates in yeast," *Molecular Systems Biology*, vol. 3, article 116, 2007.

[10] J. Peng, D. Schwartz, J. E. Elias et al., "A proteomics approach to understanding protein ubiquitination," *Nature Biotechnology*, vol. 21, no. 8, pp. 921–926, 2003.

[11] C. W. Tung and S. Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC bioinformatics*, vol. 9, no. 1, p. 310, 2008.

[12] P. Radivojac, V. Vacic, C. Haynes et al., "Identification, analysis, and prediction of protein ubiquitination sites," *Proteins: Structure, Function, and Bioinformatics*, vol. 78, no. 2, pp. 365–380, 2010.

[13] Z. Chen, Y. Z. Chen, X. F. Wang, C. Wang, R. X. Yan, and Z. Zhang, "Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs," *PLoS ONE*, vol. 6, no. 7, Article ID e16351, 2011.

[14] Z. Chen, Y. Zhou, J. Song, and Z. Zhang, "HCKSAAP_UbSite: Improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1834, no. 8, pp. 1461–1467, 2013.

[15] I. Walsh, T. Di Domenico, and S. C. E. Tosatto, "RUBI: Rapid proteomic-scale prediction of lysine ubiquitination and factors influencing predictor performance," *Amino Acids*, vol. 46, no. 4, pp. 853–862, 2014.

[16] C. J. Bendell, S. Liu, T. Aumentado-Armstrong et al., "Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor," *BMC Bioinformatics*, vol. 15, no. 1, article 82, 2014.

[17] L. Jia, T. Sun, F. Yang, H. Sun, and B. Zhang, "Prediction of ubiquitination sites with feature-weighting scheme and naive bayes vectorizer," *Journal of Computational and Theoretical Nanoscience*, vol. 13, no. 1, pp. 286–293, 2016.

[18] S. Ghosh and J. F. P. Dass, "Non-canonical pathway network modelling and ubiquitination site prediction through homology modelling of NF-κB," *Gene*, vol. 581, no. 1, pp. 48–56, 2016.

[19] M. Saeed, W. Ajmal, A. Masood, M. R. Riaz, and M. N. Akhtar, "Ubipredictor: a new tool for species-specific prediction of ubiquitination sites using linear discriminant analysis," *Current Bioinformatics*, vol. 11, no. 2, pp. 269–276, 2016.

[20] B. Cai and X. Jiang, "Computational methods for ubiquitination site prediction using physicochemical properties of protein sequences," *BMC Bioinformatics*, vol. 17, no. 1, article 116, 2016.

[21] V.-N. Nguyen, K.-Y. Huang, C.-H. Huang, K. R. Lai, and T.-Y. Lee, "A New Scheme to Characterize and Identify Protein Ubiquitination Sites," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 2, pp. 393–403, 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*, pp. 770–778, July 2016.

[23] J. R. Wang, W. L. Huang, M. J. Tsai, K. T. Hsu, H. L. Huang, and S. Y. Ho, "ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives," *Bioinformatics*, vol. 33, no. 5, pp. 661–668, 2016.

[24] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2002.

[25] A. Jammalamadaka, S. Banerjee, B. S. Manjunath, and K. S. Kosik, "Statistical analysis of dendritic spine distributions in rat hippocampalcultures," *BMC Bioinformatics*, vol. 14, no. 1, p. 287, 2013.

[26] Z. Chen, Y. Zhou, J. Song, and Z. Zhang, "HCKSAAP-UbSite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties," *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics*, vol. 1834, no. 8, pp. 1461–1467, 2013.

[27] G. Huang, Z. Liu, V. D. Maaten, and L. K. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.