Journal of Mountain Area Research Vol. 08, 2023
https://doi.org/10.53874/jmar.v8i0.167

Full length article

UBI-XGB: IDENTIFICATION OF UBIQUITIN PROTEINS USING MACHINE LEARNING MODEL

Rahu Sikander[1], Ali Ghulam[2], Ali Farman[3], Dhani Bux Talpur[4], Mir Sajjad Hussain Talpur[2], Erum Saba[2], Zulfikar Ahmed Maher[2], Saima Tunio[2]*

1. School of Computer Science and Technology, Xidian University, Xi'an 710071, China
2. Information Technology Centre, Sindh Agriculture University, Sindh, Pakistan
3. Elementary and Secondary Education, Peshawar, Khyber Pakhtunkhwa, Pakistan
4. Department of Computer Science, University of Gwaddar, Gwaddar, Balochistan

ABSTRACT

A recent line of research has focused on Ubiquitination, a pervasive and proteasome-mediated protein degradation that controls apoptosis and is crucial in the breakdown of proteins and the development of cell disorders, is a major factor. The turnover of proteins and ubiquitination are two related processes. We predict ubiquitination sites; these attributes are lastly fed into the extreme gradient boosting (XGBoost) classifier. We develop reliable predictors computational tool using experimental identification of protein ubiquitination sites is typically labor- and time-intensive. First, we encoded protein sequence features into matrix data using Dipeptide Deviation from Expected Mean (DDE) features encoding techniques. We also proposed 2nd features extraction model named dipeptide composition (DPC) model. It is vital to develop reliable predictors since experimental identification of protein ubiquitination sites is typically labor- and time-intensive. In this paper, we proposed computational method as named Ubipro-XGBoost, a multi-view feature-based technique for predicting ubiquitination sites. Recent developments in proteomic technology have sparked renewed interest in the identification of ubiquitination sites in a number of human disorders, which have been studied experimentally and clinically. When more experimentally verified ubiquitination sites appear, we developed a predictive algorithm that can locate lysine ubiquitination sites in large-scale proteome data. This paper introduces Ubipro-XGBoost, a machine learning method. Ubipro-XGBoost had an AUC (area under the Receiver Operating Characteristic curve) of 0.914% accuracy, 0.836% Sensitivity, 0.992% Specificity, and 0.839% MCC on a 5-fold cross validation based on DPC model, and 2nd 0.909% accuracy, 0.839% Sensitivity, 0.979% Specificity, and 0. 0.829% MCC on a 5-fold cross validation based on DDE model. The findings demonstrate that the suggested technique, Ubipro-XGBoost, outperforms conventional ubiquitination prediction methods and offers fresh advice for ubiquitination site identification.

KEYWORDS: Ubiquitin proteins; DDE; DPC; Ubipro-XGBoost; Machine learning

*Corresponding author: (Email: sikander@stu.xidian.edu.cn)

1. INTRODUCTION

This study used a combination of qualitative and quantitative analysis computational tool, and discovery of Ubiquitin [1] that ubiquitin is a tiny, 76-amino acid protein [2]. Protein ubiquitination is a common post-translational modification. It is a process that attaches ubiquitin, a protein, to the substrate. An increase in ubiquitin-protein levels can have a variety of effects on how a protein behaves. It can, for example, instruct the proteasome to digest proteins [3, 4]. Additionally, this process is connected to inflammation, cell change, and the immune response. [5]. A frequent post-translational modification is protein ubiquitination. It is a procedure that assigns the protein ubiquitin to the substrate. Numerous factors can change how a protein functions as a result of an increase in ubiquitin-protein levels. [6, 7]. Ubiquitination has been linked to cell change, immunological response, and inflammatory response [8]. A tiny regulatory protein called ubiquitin-protein is involved in the ubiquitination modification process and is present in practically all eukaryotic tissues. The three processes of ubiquitination are activation, binding, and connection [9]. Ubiquitination is critical to understanding protein regulation and molecular mechanisms and identifying potential ubiquitination sites is essential. It is critically needed to develop computational methods that can detect protein ubiquitination sites more quickly and precisely than traditional methods such as CHIP-CHIP analysis and mass spectrometry. The identification of protein ubiquitination sites can be done using computational approaches. A considerable amount [10] of research has focused on comprehending the mechanism of ubiquitination is the identification of ubiquitination sites. Ubiquitination is quick and reversible, though high-throughput mass spectrometry (MS) technology ubiquitin antibodies, and ubiquitin-binding proteins [11, 12], in combination with liquid chromatography and mass spectrometry [13], are examples of conventional experimental techniques. UbiProber was developed by Chen to combine sequencing information with physico-chemical parameters and amino acid composition in order to build generic models for eukaryotic proteomes and species-specific models for proteomes from a variety of different species. Physico-chemical features were added into SVM by ESA-UbiSite [14]. ESA was performed to choose the most effective negative dataset from the entire dataset, however.

The large-scale protein ubiquitination site prediction, these existing machine learning algorithms perform well on small-scale data, but there are still significant obstacles. First, the artificially designed features have a weakness. There are currently no methods that do not rely on expert knowledge for feature extraction, which results in incomplete and biased feature vectors [15, 16]. Second, there is a variety in the features. To boost accuracy, most existing prediction methods converged on a single feature and ignored the inherent heterogeneity among them. Third, there is a disparity in the number of positive and negative samples. There are only a limited number of lysine residues that can be ubiquitinated in the entire proteome, making protein ubiquitination site prediction an extremely imbalanced problem [17]. Such an imbalanced situation does not lend itself well to existing approaches for discovering probable ubiquitination sites. It's thought that deep learning, a recent trend in machine learning for massive datasets, could be the answer to these issues. To successfully analyses genomic and proteomic data, a number of deep learning networks have been used It is yet to be used in the prediction of protein ubiquitination sites by deep learning techniques. To illustrate the roles of new molecules in huge signal networks, one can use this graphic to depict ubiquitin [18]. The distinctive patterns of molecules from a certain class can be shown by a large number of interconnected proteins. [19, 20].

XGBoost ubiquitin is based on the XGBoost algorithm for protein function. Machine learning approaches such as eXtreme Gradient Boosting have been used to predict protein structures in the literature (XGBoost) [21]. Networks that use low-level features as inputs produce high-level features at the next layer. Computer vision and natural language processing both use XGBoost -based techniques. Even in biomedical data analysis, XGBoost -based methods have been found to outperform standard predictive methods used in bioinformatics and chem informatics [22] because of recent advancements in processing power [23]. Ubiquitin prediction is an area where XGBoost ubiquitin performs exceptionally well. Other machine learning classifiers like deep neural network (DNN) AdaBoost (ABC) and Random Forest (RF) classifiers are also compared to this model's prediction performance. In order to find the optimum feature extraction approach, we also use feature extraction protocols that have been successful in tackling diverse biological challenges. We believe that DDE and DPC are the most effective approaches for extracting features from a dataset.

Table 1. Collected Data as ubiquitin and non-ubiquitin sequences.

	Original data	**Similarity <30%**	**Cross-validation**
Ubiquitin proteins	550	375	375
Non-Ubiquitin proteins	650	450	450
Total	1200	825	825

An approach based on the previously mentioned machine learning classifiers is also on the table 1.

2. MATERIALS AND METHODS

2.1 Datasets

Machine learning models can be simplified by employing a quantitative approach that includes the usage of a dataset. The UniPortKB and NCBI-databases are where we get our information. Eight hundred and twenty-five different protein sequences have been obtained, with the majority 375 being ubiquitin positive and the remainder 450 non-ubiquitin positive. This is a class of proteins used to model subcellular distributions [24]. We've gotten the info from the database above. The obtained datasets are preprocessed based on the protein–pathway and protein–non-pathway interactions. Data was stored in CSV format and the parameters of our suggested model were established. The sequence of an ubiquitination-precise protein were proposed as a positive test sample. Training datasets were imbalanced by a random selection of positive and negative samples [25]. An online database containing proteins from a variety of organisms was used in this study but only human-related proteins that were specifically implicated in human pathways were investigated more than 450 non-ubiquitin proteins were received as part of the CD-HIT [26], step for similarity measures. This preprocessing method resulted in the finalization of 775 proteins by removing redundant information. There was a reduction in redundancy, and 775 proteins were received, including 375 ubiquitin proteins and 450 non-ubiquitin proteins.

2.2 Features Extraction for Ubiquitin Protein Association

The process of turning protein sequence information into numerical data, known as feature extraction, is crucial to the classification effort. In order to extract the information from protein sequences, sequence-based features, physicochemical property-based features, and evolutionary-derived features are chosen in this study. First, we encoded protein sequence features into matrix data using Dipeptide Deviation from Expected Mean (DDE) features encoding techniques. Second, we encoded features extraction model named dipeptide composition (DPC) model. A two-dimensional sparse matrices of size 20x20 was obtained and reduced to a one-dimensional vector. With this method of random projection, an effective measurement matrix was used to generate a small functionality set. Because of this, a new method of extracting compressive sensing functionality has been developed. The XGBoost, DDE, and DPC feature profiles were studied, and an essential approach for classifying pathway-specific proteins was devised. Data gathering, feature extraction, CNN development, and model evaluation are all part of the system. Figure 1 depicts our system's flowchart and provides the following explanation of its specifics. In order to detect and classify proteins that are peculiar to human pathways, a new technique was created.

2.3 Features Encoded By DDE

We distinguish between a cell's ubiquitination and non- ubiquitination, feature extraction based on amino-acid combination is studied in relation to the (DDE). The primary formula used to determine a protein sequence's dipeptide combination (DC). We encoded physicochemical, evolutionary functions from the Ubiquitin datasets. It was shown that the DDE features profiles vector were more effective than other characteristic representations in boosting the specific linear proteins linked with pathogen protection. According to earlier studies, dipeptide frequency variations were measured using dipeptide composition features in this study. The theoretical mean (Tm), theoretical variance (Tv), and dipeptide composition (DDE) were used to build the DDE feature vector (Cc). It is determined as follows: the three parameters and DDE, and DC an indicator for (Cc) is supplied by DC (i).

$$D_{c(\mathrm{i})} = \frac{n_i}{N} \tag{1}$$

It was possible to extract 400 dipeptide attributes (20 ordinary amino acids 20×20 dipeptide properties), although not all of them followed one another in any particular order. Dipeptide I and N are also not found in $L\text{-}1$ (i.e., potential quantity in P). The theoretical mean ($T_{M\,(i)}$).

$$T_{M\,(\mathrm{i})} = \frac{C_{i1}}{C_N} \times \frac{C_{i1}}{C_N} \tag{2}$$

C_{i1} is the number of codons for the first amino acid, and Ci2 is the number of codons for the second amino acid, both for the dipeptide.

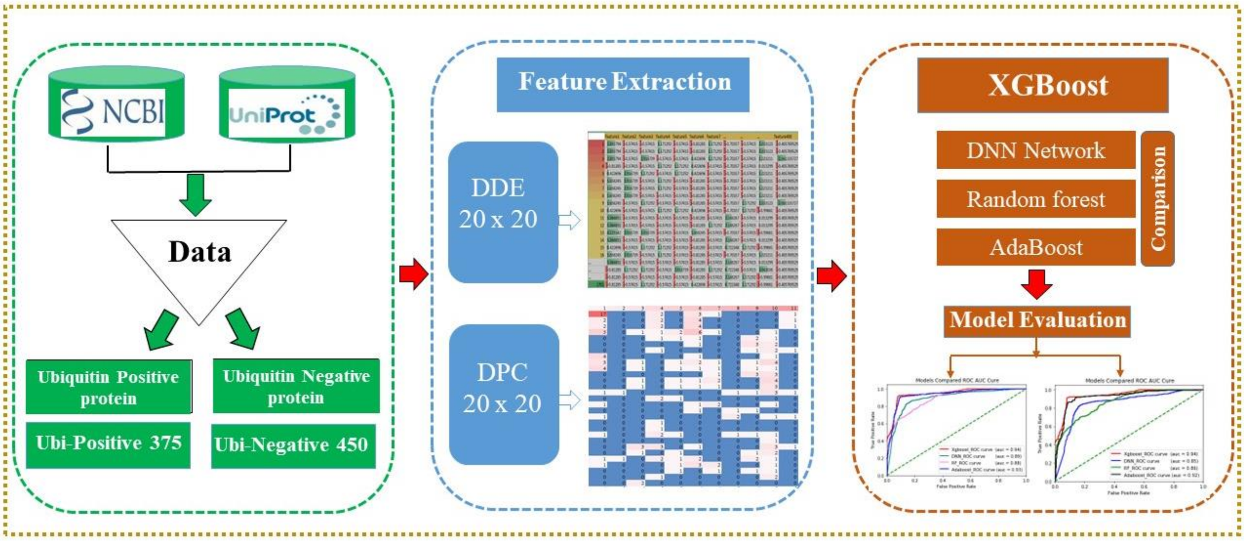

Figure 1. The XGBoost ubiquitin protein framework model has been proposed.

Except for the three codons, the total number of codons is C_N. In order to avoid having to recalculate the features of $T_{M(i)}$, only features with a length of 400 dipeptides were used. The theoretical variance of $T_{V(i)}$ is provided by dipeptide i.

$$T_{\mathrm{V(i)}} = \frac{T_{M\,(\mathrm{i})}(1-T_{M\,(\mathrm{i})})}{N} \quad (3)$$

This equation gives the theoretical average of the number j, or *TM(j)* (2). Again, and peptide P has the same number of L-1-dipeptides as before. Finally, *DDE(i)* is identified as

$$DDE_{(\mathrm{t})} = \frac{D_{c(\mathrm{i})-\mathrm{T}_{m(\mathrm{i})}}}{\sqrt{T_{V\,(\mathrm{i})}}} \quad (4)$$

We calculate the 400-dimensional features vector was used to calculate DDE for each of the 400 dipeptide features.

$$DDE_p = (DDE_{(i)} \dots .\, DDE_p), \text{where } \ i = 1,2, \dots .4) \quad (5)$$

2.4 Features encoded with DPC

The dipeptide composition is present in the first two successive residues (DPC). Sequences are limited to 400 characters. For the most part, this sequence representation provides information on the amino acid composition and local order. The DPC feature extraction procedure was performed on this model in order to extract the best foundational features. When an amino acid occurs twice in a row in a protein sequence, it is referred to as a double-prefix codon (DPC). To give an example, in the series there are dipeptide frequencies for MALMAC (two), ALLM (one), AC (one), as well as CC (one). The total number of feature elements was 400 dipeptides. In order to standardize the DPC features, we divided the frequencies by (N-1) [27]. The frequency of two adjacent amino acids in a dipeptide captures new information about the amino acid makeup. Because of this, the dipeptide composition is ideal in situations requiring localized information, such as homologic information.

$$f_j = \frac{\#of \ \text{diseptide}_j}{N-1} \times 100 \quad (6)$$

2.5 Proposed model

We build a novel machine learning model for protein association prediction by using the

XGBoost Ubiquitin Protein Sequence. A two features extractions technique is implemented in order to remove the unnecessary functionalities from the model before the model is constructed. Ubipro-XGBoost ubiquitin is then compared to two features encoding models, and the results are used as inputs to three machine learning classifiers. We can also develop hybrid features by combining various feature space combinations. For this purpose, 10-fold cross-validation tests are also carried out. As shown Figure 1 illustrates our proposed method framework.

This section proposed a unique machine learning technique and feature extraction model for predicting ubiquitination sites. As shown in Figure 1 shows the suggested method Ubipro-XGBoost framework model. First step in the green box we collected from the mentioned databases, and then removed similarity redundancy, finalized the ubiquitin positive proteins datasets and ubiquitin negative proteins datasets. Second step in the blue box we extracted features by DDE and DPC model and then feature normalization. Third step in the brown box we proposed XGBoost algorithm for the classification on the basis of 10-fold cross-validation. To evaluate the classification model's ability to predict outcomes, by using 10-fold class validation technique. We than our proposed XGBoost algorithm performance with other three machine learning classifiers. According to simulation results, the proposed strategy performs reasonably well when compared to some cutting-edge techniques [28, 29]. An ensemble algorithm known as extreme gradient boosting (XGBoost) has recently been shown to produce more accurate energy models than artificial neural networks and degree-day ordinary least square regression by Chakraborty and Elzarka [30] [31].

3. PERFORMANCE EVALUATION METHODS

The training dataset is used to tune the parameters of the models using a tenfold cross-validation approach, and the independent set is used to test the model [32]. The underlying models have been evaluated using efficiency metrics such as sensitivity (sn), specificity (Sp), accuracy (ACC), and Mathew's correlation coefficient (MCC). In this study, true positive (TP), false positive (FP), false negative (FN), and true negative (TN) are the four units in the confusion matrix derived following prediction (TN). Sensitivity, specificity, precision, accuracy, F-score, and Matthew's correlation coefficient (MCC) were some of the metrics used to evaluate the overall prediction performance of different categorization models. Previous research have utilized them, with a greater value suggesting better performance (Jing and Dong, 2017). The following are some examples of performance metrics.

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

4. RESULTS AND DISCUSSION

We used positive samples (375 sequences) and negative samples (375 sequences) benchmark dataset. In the first phase, we had to compare various matrices in order to get the best DDE and DPC matrix for our model. In the end, we found that the DPC matrix (20x20) was the best one for solving the imbalance data. Next, we set into the XGBoost algorithm with all 400 retrieved feature sets. There are many ways in which experimentation might be developed. According to our two DPC models, we then employed the DDE model.

4.1 Ubiquitin and Non- Ubiquitin Sequence for the AAC

The number of amino acids in ubiquitin and non-ubiquitin sequences was calculated in order to determine their composition. The 20 amino acids that contribute significantly to two datasets. There are few notable exceptions to the general rule that there are no significant differences between the two categories of data. The highest concentrations of C and P amino acids can be found throughout proteins. So, the finding of ubiquitin proteins in these amino acids is crucial. In light of the various properties of these amino acids, our model is able to Accurately predict ubiquitin proteins as shown Table 2.

Table 2. Metric Performance obtain by XGBoost.

ML-Classifier	ACC	Precision	Sensitivity	Specificity	MCC	F1
RF -DPC	0.779%	0.769%	0.783%	0.775%	0.570%	0.758%
DNN -DPC	0.841%	0.838%	0.831%	0.852%	0.705%	0.798%
AdaBoost -DPC	0.901%	0.941%	0.852%	0.950%	0.821%	0.861%

4.2 Ubiquitin between XGBoost and Shallow Machine Learning with a Comparable Efficiency

According to this finding, multiple machine learning algorithms were tested in order to identify proteins derived from Ubiquitin. We employee four machine learning classifiers were used in our study (e.g., AdaBoost [33], Random-Forest, and DNN). Our XGBoost[34] was compared to the DNN Deep Neural Network's implementation of perceptions, and the results were compared to our XGBoost. Table 3 and Figure 2 shows that we used the optimum parameters in all of our trials so that we could compare each classifier to the others. We found that our XGBoost performed better than other standard machine learning approaches in the same experiment framework. Our Ubipro-XGBoost, in particular, created algorithms based on a distinct dataset.

Table 3. Performance Comparison ML classifiers by DDE model

	ACC	Prc	Sens	Spec	Mcc	F1
RF-DDE	0.752	0.759	0.767	0.738	0.519	0.742
DNN-DDE	0.827	0.870	0.805	0.849	0.688	0.779
AdaBoost-DDE	0.878	0.874	0.844	0.912	0.767	0.832

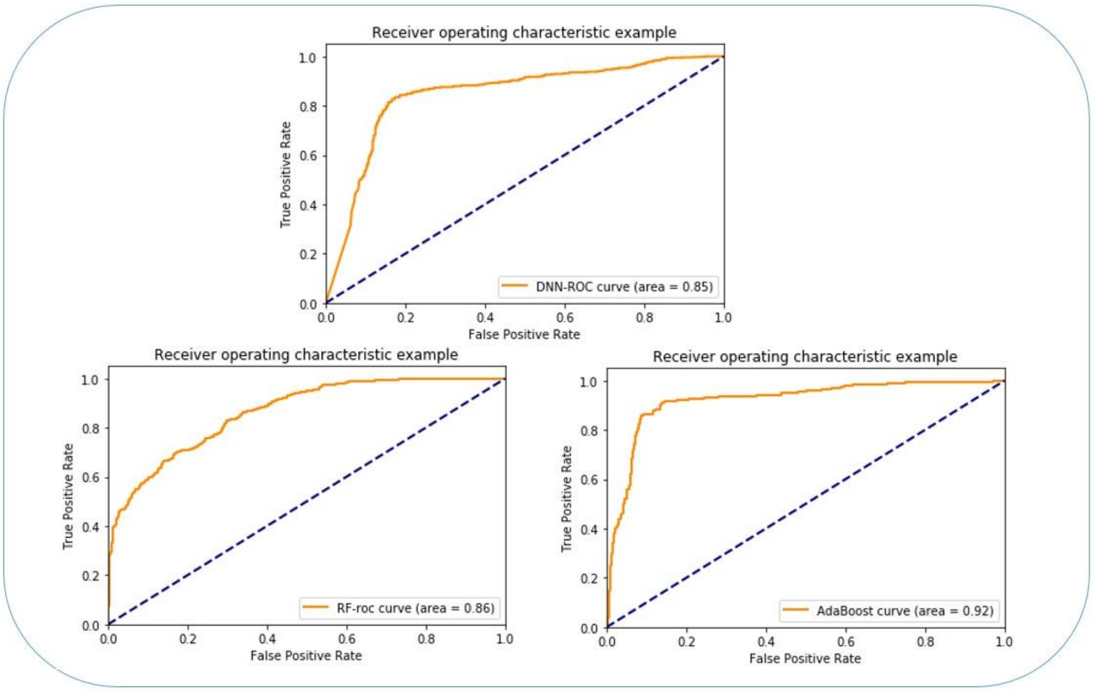

Figure. 2. Proposed model compare with other classifiers

Table 4. Performance Comparison other three ML classifiers by DPC model

Machine Learning Classifier	ACC	Precision	Sensitivity	Specificity	MCC	F1
RF -DPC	0.779%	0.769%	0.783%	0.775%	0.570%	0.758%
DNN -DPC	0.841%	0.838%	0.831%	0.852%	0.705%	0.798%
AdaBoost -DPC	0.901%	0.941%	0.852%	0.950%	0.821%	0.861%

Table.4 shows how the XGBoost hybrid features can be used to demonstrate the classifier's predictive power. The Random Forest classification, on the other hand, performed very well in this mixed-feature comparison. This model classification was more accurate than XGBoost, which predicted Random Forest model classifications with 93.53% accuracy using XGBoost data. Table 4 and Figure 3 shows the results of a comparison with three MLCs.

4.3 ROC (Auc) Comparative Performance by DDE and DPC

The results analysis consists of prior investigations into the binary classification issue that we used in our study. Our results were discovered to be accurate and consistent with the majority of machine learning classification algorithms. Researchers also employ other metrics in the ROC curve plot and the ROC (AUC), such as the algorithm's accuracy or the confusion matrix.

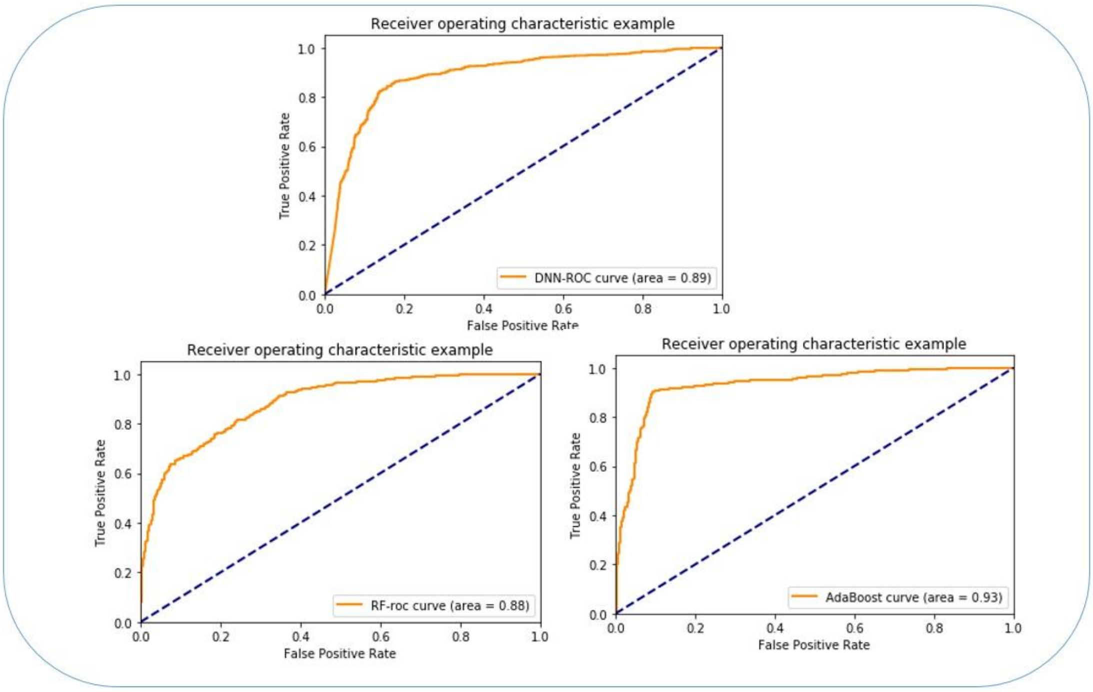

Figure. 3. Proposed model compare with other classifiers

Figures 4 show the results of classifying the Ubipro-XGBoost output using the ROC AUC curve. The Ubipro-XGBoost Ubiquitin multilink ROC auc curve is shown. It appears that our Ubipro-XGBoost model perform well even with multi-classification, however more data were needed to investigate this discovery in more depth. There were no over fitting issues with our suggested cross-validation Ubipro-XGBoost model, which had an accuracy rate of 0.914% percent.

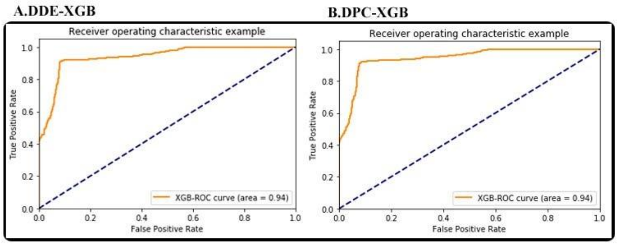

Figure 4. ROC (AUC) with DDE and ROC-Auc with DPC Model

ROC and ACU scores of DPC model cross validation datasets were found to be 0.94% percent, while RCO-AUC scores of with DDE model datasets were found to be 0.94% percent.

4.4 ROC (Auc) Score Comparison with other 3 three Classifiers by Using DDE and DPC

As can be seen, Ubipro-XGBoost performs better than the alternatives. We calculated ROC (AUC) score comparison for several machine learning approaches as shown in figure 5. As can be observed each methods prediction rate is considerably higher than random prediction. Additionally, the XGBoost classifier performs better than the others. Ubipro-XGBoost ubiquitin identification as shown Figure 5 distinct ubiquitin datasets are represented by different ROC–AUC curves score. DDE model achieved performance such as AdaBoost ROC (AUC) generate 0.92%, RF ROC (AUC) generate 0.86%, DNN ROC (AUC) generate 0.85%, and our proposed model XGBoost ROC (AUC) generate 0.94%, which is better than other classifiers. DPC model achieved performance such as AdaBoost ROC (AUC) generate 0.93%, RF ROC (AUC) generate 0.88%, DNN ROC (AUC) generate 0.89%, and our proposed model XGBoost ROC (AUC) generate 0.94%, which is better than other classifiers.

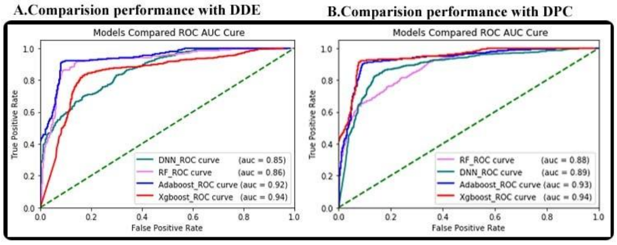

Figure. 5. ROC curves of the comparison performance with DDE and DPC methods.

5. DISCUSSION

The Ubipro-XGBoost predictor is trained on the most comprehensive database of protein Ubiquitin modifications. Using a machine learning classification model, an XGBoost is used to predict ubiquitination. First, ubiquitination is predicted using the machine learning classification models. The best result for the XGBoost classification model accuracy, 0.836%. Then a DPC precision score 0.892% was achieved in the XGBoost model and DDE precision score 0.881% was achieved in the XGBoost model, and the accuracy score was achieved with the XGBoost model, which indicates that our test overall XGBoost classification is initially and then secondly the AdaBoost classification model according to our experimental tests. DNN Analysis, the third-best classifier with ROC (AUC) on the DPC [35] and DDE model [36] was achieved in addition to the highest analyses as shown in Figure 5. Ubiquitin proteins [37]. Ubipro-XGBoost had an AUC (area under the Receiver Operating Characteristic curve) of 0.914% accuracy, 0.836% Sensitivity, 0.992% Specificity, and 0.839% MCC on a 5-fold cross validation based on DPC model, and 2nd 0.909% accuracy, 0.839% Sensitivity, 0.979% Specificity, and 0. 0.829% MCC on a 5-fold cross validation based on DDE model.

CONCLUSION

The XGBoost algorithm was used to produce Ubipro-XGBoost, a predictor for the correct identification of Ubiquitin proteins. As compared to earlier predictors, we have attained state-of-the-art performance on the benchmark dataset. It is possible to infer three main conclusions. To begin, the XGBoost algorithm consistently and accurately predicts Ubiquitin levels when compared to other algorithms. To further enhance model performance, the DDE and DPC feature selection method was used to optimize feature vectors, which extracted the most significant features from a huge number of candidates features and increased the model's accuracy. This is a significant advantage over other sequence-based Ubiquitin predictors, which are limited in their ability to provide relevant explanations for samples provided using the SHAP technique. DPC features contributed to the final prediction direction, which is explained here. Also explained is the importance of paying attention to some specific identities, as well as a range of other traits.

The end results demonstrated that Ubipro-XGBoost obtained a satisfactory and promising performance, which is steady and credible. There are still unknowns about

Ubiquitin, such as how many of them there are and what they do. This limits the accuracy of the model. In addition, it is necessary to investigate some possible connections among the features. Ubiquitin and non-Ubiquitin will be separated in the future by finding and extracting as many features as possible from a vast amount of data.

DECLARATIONS

Funding: No funding was received for this study.

Conflicts of interest/Competing interests: The authors declare no any conflict of interest/competing interests.

Data availability: Not applicable.

Code availability: Not applicable.

Authors' contributions: Rahu Sikander, Ali Ghulam and Farman Ali jointly contributed to the design of the study. Rahu Sikander conceptualized the review and finalized the manuscript. Ali Ghulam and Dhani Bux Talpur wrote the initial manuscript. Farman Ali helped to draft the manuscript. Ashfaq Ahmed revised the manuscript and Rahu Sikander polished the expression of English. All of the authors have read and approved the final manuscript.

REFERENCES

[1] Goldstein G, Scheid M, Hammerling U, Schlesinger DH, Niall HD, Boyse EA. Isolation of a polypeptide that has lymphocyte-differentiating properties and is probably represented universally in living cells. Proc Natl Acad Sci U S A.72(1)(1975)11–5

[2] Wilkinson KD. The discovery of ubiquitin-dependent proteolysis. Proc Natl Acad Sci U S A. 2005; 102(43):15280–2.

[3] Pickart CM, Eddins MJ. Ubiquitin: structures, functions, mechanisms. Biochim Biophys Acta. 2004; 1695(1–3):55–72.

[4] Welchman RL, Gordon C, Mayer RJ. Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol.6 (8)(2005)599–609.

[5] Peng JM, Schwartz D, Elias JE, Thoreen CC, Cheng DM, Marsischky G, et al. A proteomics approach to understanding protein ubiquitination. Nat Biotechnol.21(8) (2003)921–6

[6] Herrmann J, Lerman LO, Lerman A. Ubiquitin and ubiquitin-like proteins in protein regulation. Circ Res.;100(9)(2007)1276–91.

[7] Welchman R, Gordon C, Mayer RJ. Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol.6 (8)(2005)599–609.

[8] Schwartz AL, Ciechanover A. The ubiquitin-proteasome pathway and pathogenesis of human diseases. Annu Rev Med.50 (1999) 57–74.

[9] Zhong J, Shaik S, Wan L, Tron AE, Wang Z, Sun L, Anushka H, Wei W.SCF beta-TRCP targets MTSS1 for ubiquitination-mediated destruction to regulate cancer cell proliferation and migration. Oncotarget. 4(12) (2013) 2339–53

[10] B. Yu, Z. Yu, C. Chen, A. Ma, B. Liu, B. Tian, Q. Ma, DNNAce: prediction of prokaryote lysine acetylation sites through deep neural networks with multi-information fusion, Chemomet. Intell. Lab. 200 (2020) 103999.

[11] G. Xu, J.S. Paige, S. R Jaffrey, Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling, Nat. Biotechnol. 28 (2010) 868–873.

[12] W. Kim, E.J. Bennett, E.L. Huttlin, A. Guo, J. Li, A. Possemato, M.E. Sowa, R. Rad, J. Rush, M.J. Comb, J.W. Harper, S.P. Gygi, Systematic and quantitative assessment of the ubiquitin-modified proteome, Mol. Cell. 44 (2011) 325–340.

[13] P. Radivojac, V. Vacic, C. Haynes, R.R. Cocklin, A. Mohan, J.W. Heyen, M. G. Goebl, L.M. Iakoucheva, Identification, analysis, and prediction of protein ubiquitination sites, Proteins 78 (2010) 365–380.

[14] Huang CH, Su MG, Kao HJ, Jhong JH, Weng SL, Lee TY. UbiSite:incorporating two-layered machine learning method with substrate motifsto predict ubiquitin-conjugation site on lysines. BMC Syst Biol.10 (Suppl 1)(2016)6.

[15] Nguyen VN, Huang KY, Huang CH, Lai KR, Lee TY. A new scheme tocharacterize and identify protein ubiquitination sites. IEEE/ACM Trans Comput Biol Bioinform.14 (2) (2017)393–403.

[16] Qiu WR, Xiao X, Lin WZ, Chou KC. iUbiq-Lys: prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a gray system model. J Biomol Struct Dyn.33(8) (2015)1731–42.

[17] Chen X, Qiu JD, Shi SP, Suo SB, Huang SY, Liang RP. Incorporating key position and amino acid residue features to identify general and speciesspecific ubiquitin conjugation sites. Bioinformatics.29(13) (2013)1614–22.

[18] Wang JR, Huang WL, Tsai MJ, Hsu KT, Huang HL, Ho SY. ESA-UbiSite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives. Bioinformatics. 33(5)(2017)661–8

[19] Yuan Y, Xun G, Jia K, Zhang A, Acm: a multi-view deep learning method for epileptic seizure detection using short-time Fourier transform; 2017.

[20] Yuan Y, Xun G, Jia K, Zhang A. A Novel Wavelet-based Model for EEG Epileptic Seizure Detection using Multi-context Learning. In: Hu XH, Shyu CR, Bromberg Y, Gao J, Gong Y, Korkin D, Yoo I, Zheng JH, editors. 2017 Ieee International Conference on Bioinformatics and Biomedicine; (2017).p. 694 –9.

[21] SAnchez, R. O. B. E. R. T. O., & Sali, A. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. Proceedings of the National Academy of Sciences, 95(23), (1998) 13597-13602.

[22] Husnjak, K., & Dikic, I.Ubiquitin-binding proteins: decoders of ubiquitin-mediated cellular functions. Annual review of biochemistry, 81, (2012) 291-322.

[23] Agrahari, A. K., Bose, P., Jaiswal, M. K., Rajkhowa, S., Singh, A. S., Hotha, S. ... & Tiwari, V. K. Cu (I)-catalyzed click chemistry in glycoscience and their diverse applications. Chemical Reviews, 121(13),(2021) 7638-7956.

[24] Wang, M., Cui, X., Li, S., Yang, X., Ma, A., Zhang, Y., & Yu, B. DeepMal: Accurate prediction of protein malonylation sites by deep neural networks. Chemometrics and Intelligent Laboratory Systems, 207,(2020) 104175.

[25] Liu, Y., Jin, S., Song, L., Han, Y., & Yu, B. Prediction of protein ubiquitination sites via multi-view features based on eXtreme gradient boosting classifier. Journal of Molecular Graphics and Modelling, (2021) 107962.

[26] Alsanousi WA, Ahmed NY, Hamid EM, Elbashir MK, Musa MEM, Wang J, et al.A novel deep learning-assisted hybrid network for plasmodium falciparum parasite mitochondrial proteins classification. PLoS ONE 17(10): e0275195. https://doi.org/10.1371/journal.pone.0275195.(2022)

[27] Min, S., Lee, B. & Yoon, S.Brief. Bioinform. 18, (2016) 851–869 .

[28] Kandaswamy,K.K.,Pugalenthi,.,Kalies,K.U.,Hartmann,E.,Martinetz,T.,2013

[29] Saravanan, V. & Gautham, N. Harnessing computational biology for exact linear B-cell

epitope prediction: A novel amino acid composition-based feature descriptor. OMICS 19, (2015) 648–658 .

[30] V. Saravanan and N. Gautham, ‘‘Harnessing computational biology for exact linear B-Cell epitope prediction: A novel amino acid composition based feature descriptor,’’ OMICS, A J. Integrative Biol., vol. 19, no. 10, pp. (2015) 648–658,doi: 10.1089/omi.2015.0095.

[31] V. Saravanan and N. Gautham, ‘‘BCIgEPRED—A dual-layer approach for predicting linear IgE epitopes,’’ Mol. Biol., vol. 52, no. 2, (2018) pp. 285–293,doi: 10.1134/S0026893318020127.

[32] L. Zou, C. Nan, and F. Hu, ‘‘Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles,’’ Bioinformatics, vol. 29, no. 24, (2013) pp. 3135–3142,doi: 10.1093/bioinformatics/btt554

[33] Ghulam, A., Sikander, R., Ali, F., Swati, Z. N. K., Unar, A., & Talpur, D. B. (2022). Accurate prediction of immunoglobulin proteins using machine learning model. Informatics in Medicine Unlocked, 29, (2022) 100885.

[34] Sikander, R., Ghulam, A. & Ali, F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Sci Rep 12, 5505 (2022).

[35] Ghualm, Ali, et al. "Identification of Pathway-Specific Protein Domain by Incorporating Hyperparameter Optimization Based on 2D Convolutional Neural Network." IEEE Access 8 (2020) 180140-180155.

[36] Ghulam, A., M. Memon, M. Hyder, Z. A. Maher, A. Unar, Z. N. K. Swati, D. B. Talpur, R. Sikander, I. Ullah, and A. Farman. "Identification of Novel Protein Sequencing SARS CoV-2 Coronavirus Using Machine Learning." Bioscience Research (2021) 47-58.

[37] Sikander, R., Arif, M., Ghulam, A., Worachartcheewan, A., Thafar, M. A., & Habib, S. Identification of the ubiquitin–proteasome pathway domain by hyperparameter optimization based on a 2D convolutional neural network. Frontiers in Genetics, 13(2022).

Received: 19 Sep. 2022. Revised/Accepted: 10 Nov. 2022.