UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

An Improved Deep Neural Network Algorithm for the Prediction of Limited Proteolysis in Native Protein

Haslina Hashim[1], Lim Lip Hong[2], Azurah A Samah[3]*, Hairudin Abdul Majid[4], Zuraini Ali Shah[5], Nuraina Syaza Azman[6], Nur Sabrina Azmi[7]
Artificial Intelligence and Bioinformatics Group (AIBIG),
School of Computing,
Faculty of Engineering,
Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
Email: haslinah@utm.my[1], lip.hong@graduate.utm.my[2], *azurah@utm.my[3], hairudin@utm.my[4], aszuraini@utm.my[5], nsyaza7@graduate.utm.my[6], nsabrina36@graduate.utm.my[7]

Submitted: 16/9/2021. Revised edition: 25/10/2021. Accepted: 28/10/2021. Published online: 16/5/2022
DOI: https://doi.org/10.11113/ijic.v12n1.351

Abstract*—Protease is a proteolytic enzyme that hydrolyzes the amino acid where the cleavage only occurs at specific sites of the amino acid substrate. By discovering the nick site, the prediction on the function of proteases can be identified and enable humans to control the protein's hydrolysis by their corresponding protease. This is an important process to control as it can help to control protein replication especially viral proteins. With the rise of computational methods in this era, mainly through the successful application of deep learning in various domains, the application of this method in biological data can help to improve predictions to support the experimental methods. Conventional techniques such as mass spectrometry and two-dimensional gel electrophoresis can be supported by computational methods by preparing predictions. Thus reducing the cost of experiment and time taken to identify and predict the protein proteolysis site. This study improves the deep learning algorithm by proposing the Hybrid model of Random Forest + Deep Neural Network (Hybrid RF+DNN) to classify proteolysis or nick sites. The classification in this study is compared with other machine learning algorithms such as Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Network (DNN). The proposed method enhances the classification results in identifying the positive and negative nick sites. The RF is a feature-selector that gathers the most important feature before entering the DNN classifier. This approach reduces the data dimensionality and speeds up the execution time of the training process. The performance of the models was measured by confusion matrix, specificity, sensitivity, etc. However, the proposed method is not the best performer among the mentioned classifiers as the classifiers have obtained 0.64, 0.65, and 0.58 for Datasets A, B, and C, respectively. The proposed method may become the best performer as the parameter tuning is done more precisely, even after the feature selection by the RF algorithm. Thus, the proposed method with the enhancement appears to be an alternative to the researcher discovering the limited proteolysis or nick site.

***Keywords*—Protease Nick Sites, Random Forest (RF), Support Vector Machine (SVM), Deep Neural Network (DNN), Hybrid model of Random Forest and Deep Neural Network (Hybrid RF+DNN)**

I. INTRODUCTION

Proteases are proteolytic enzymes that hydrolyses peptide bonds at the cleavage site of their substrates. Proteases can be found in all living organisms, from microbial, plant, animal, and human (Mótyán *et al.,* 2013). Proteases undergoes proteolysis (hydrolysis of the peptide bond) only if the polypeptide chain can bind to fit the active sites of the particular proteases, or commonly addressed as nick sites. The positive nick sites are the site where the proteases will undergo proteolysis to cleave the protein into substrates and products, while negative nick sites are the opposite.

Protease can be classified into six types that differ in their catalytic mechanism and biological processes. There are six types of proteases: aspartic proteases, cysteine proteases, metalloproteases, serine proteases, threonine proteases, and glutamic proteases (Rawlings and Barrett, 1999). By

discovering the nick site, the prediction on the function of proteases can be identified and enable humans to control the limited proteolysis or hydrolysis of the protein by their corresponding protease. This is an important process to control as it can help to control protein replication especially viral proteins

Thus, the research is important in epidemiology. With the rise of computational methods in this era, deep learning is becoming more famous and applied in every field of study, including the biological area. Conventional techniques such as mass spectrometry and two-dimensional gel electrophoresis can be supported by computational methods by preparing predictions. Thus reducing the cost of experiment and time taken to identify and predict the protein proteolysis site. Song and colleagues (Song *et al*., 2012) proposed an integrated feature-based server (PROSPER), a machine learning approach built based on Support Vector Machine (SVM), to predict protease cleavage sites. The study proved that the machine learning method predicts cleavage sites of multiple proteases within a single substrate sequence through PROSPER (Song *et al*., 2012). In addition, the prediction of the lysine succinylation sites has been implemented with a Random Forest (RF) algorithm by Jia and colleagues (Jia *et al.*, 2016). RF is also being applied to predict the HIV-1 protease by several studies (Li *et al*., 2018; Sigh *et al*., 2018; Lu *et al*., 2019). DeepSig is a Deep Neural Network (DNN)based approach in peptide signal detection, and nick site prediction contributed by Savojardo and colleagues (Savojardo *et al*., 2018).

This research aims to address the application of the machine learning algorithm in classifying the positive or negative nick sites of protease in the native protein. This research also introduces an improved deep neural network method and benchmarks the performance with other machine learning algorithms such as RF, SVM, and DNN. The machine learning classifiers used are Random Forest (RF), Support Vector Machine (SVM), Deep Neural Network (DNN), and proposed improved DNN – the hybrid model of RF and DNN.

In this research, data undergo pre-processing, feature selection, and applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the protease nick sites datasets. The performance of the classifiers is evaluated using several performance metrics such as confusion matrix, specificity, sensitivity, etc. The proposed machine learning algorithm is then compared to the other famous machine learning algorithms mentioned earlier. These research experiments will be carried on Google Cloud Platform (GCP) and Google Collaboratory (Collab) with the machine learning algorithms' libraries such as Scikit-learn and Keras in python language.

II. METHODOLOGY

The overall methodology of this study is illustrated in Fig. 1. There are 4 phases in general: literature review, data pre-processing, development of improved classification algorithms, and lastly, performance evaluation. The first phase is the literature review phase on investigating the proteases domain and its nick site determination. The second phase is data pre-processing, which ensures the datasets can be fitted into a machine learning model to classify the nick site. The next phase is the implementation of the classification algorithm - an improved deep neural network. In this phase, the implementing algorithms are to classify positive or negative nick sites from the datasets. Finally, the evaluation of the models of the algorithms is carried out as the last main phase in this study. Fig. 1 below shows the phases of the methodology of this study.

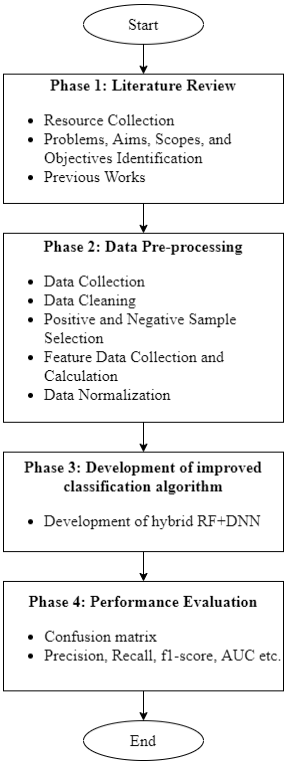

Fig. 1. Research methodology

A. Datasets

Hashim contributed the protease positive and negative nick sites' datasets from her previous study (Hashim, 2013). This study uses three datasets: Dataset A, Dataset B, and Dataset C. According to Hashim, all three datasets were gathered by different selection methods. Dataset A is collected by the top ten percent of the ranked nick sites, whereas Dataset B is gathered from the top 10 nick sites, and lastly, Dataset C is gathered from the whole random sites.

The summarization of datasets used in this study is shown in Table I. As the gathering method of datasets are different; thus the number of instances for each dataset is different. There are 10,570 instances for Dataset A, consisting of 460 positive nick sites and 10,110 negative nick sites. Meanwhile, Dataset B compromises 3,761 negative nick sites and 460 positive nick sites that sum up to have 4,221 instances. Lastly, Dataset C contains 108,848 instances, consisting of 108,388 negative nick sites and 460 positive nick sites.

TABLE I. DATASETS' INSTANCES

Dataset	Positive Nick Sites	Negative Nick Sites	Total
A	460	10,110	10,570
B	460	3,761	4,221
C	460	108,388	108,848

B. Data Preparation

Data preparation was carried out to remove outliers and noise to fit the dataset into machine learning models. The datasets retrieved in flat-file format were converted to the comma-separated delimiter (CSV) file format. Data cleaning was then executed to remove corrupted data and unnecessary columns. In these datasets, few columns are meaningless as they are not features. The columns proteases identifier, name of the nick site, and its corresponding numbers were removed from the dataset during the data preparation process. Finally, all the datasets were normalized to ensure the accuracy and efficiency of the machine learning models.

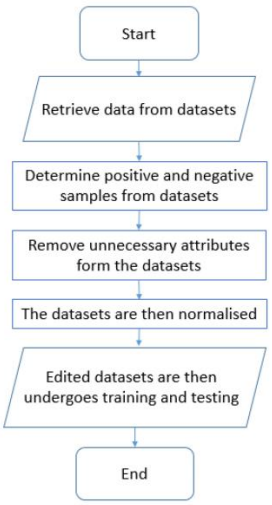

Fig. 2. Workflow of the data pre-processing

C. SMOTE Implementation

Fig. 3 depicts the classification framework for the protease nick site dataset with the implementation of SMOTE. SMOTE was implemented to solve the imbalanced datasets. The imbalanced datasets were proven to affect the efficiency of the machine learning models (Khairuddin, 2019). The undersampling and oversampling methods was not favored among researchers since many meaningful instances are removed. In contrast, the latter approach will cause the classification model to be overfitting. However, this advanced oversampling approach will create new samples for the undercount instances. The oversampled instances are not redundant but just similar. Thus, there is no data redundancy issue in the dataset after SMOTE implementation. First, the datasets were split into training and testing sets, with the ratio of 80:20, followed by SMOTE implementation on training sets. The objective of SMOTE is to evaluate the prediction from the training sets, and thus the testing sets must be original and did not need any modification. Table II below shows the summary of samples for datasets A, B, and C before and after SMOTE implementation for handling class imbalance.

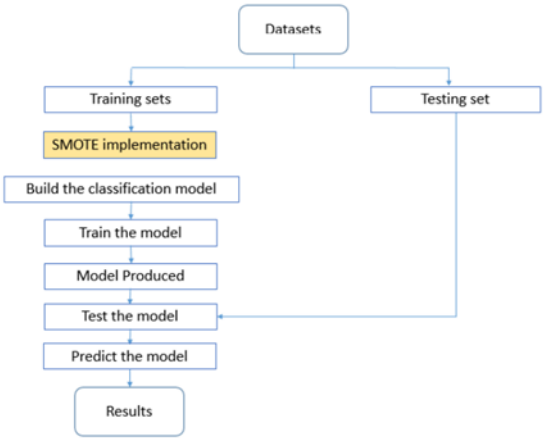

Fig. 3. Classification framework of the protease nick site datasets with SMOTE implementation.

In Table II, a noticeable change to the datasets is the counts of the instances are in the balance amount after SMOTE implementation. Then, the datasets are ready to fit into the machine learning algorithms for training and testing purposes.

TABLE II. DATASETS' INSTANCES BEFORE AND AFTER SMOTE IMPLEMENTATION

Training Dataset	Before		After	
	Positive Nick Sites	Negative Nick Sites	Positive Nick Sites	Negative Nick Sites
A	372	8,084	8084	8084
B	369	3,007	3007	3007
C	366	86,712	86,712	86,712

D. Machine Learning Models

Three machine learning classifiers methods RF, SVM, and DNN, have been applied in this study to identify positive nick sites in the native protein. An improved deep neural network proposed by the research and its performance will be compared with the previously mentioned machine learning algorithms. The following figures represent the architecture of machine learning algorithms, starting from RF (Fig. 4), SVM (Fig. 5), DNN (Fig. 6), and Hybrid RF+DNN (Fig. 7).

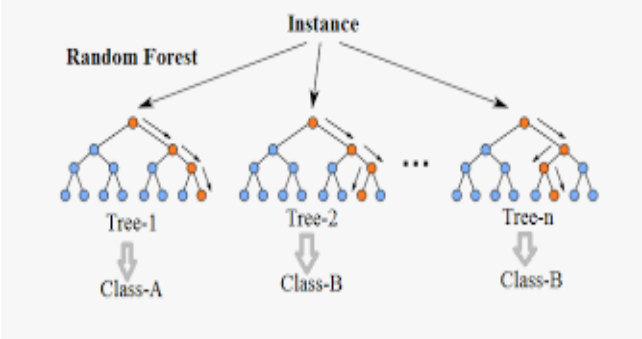

Fig. 4. Random Forest Architecture

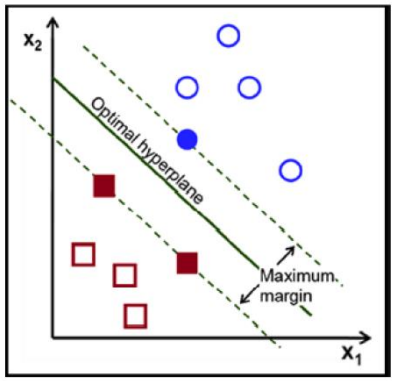

Fig. 5. Support Vector Machine Architecture

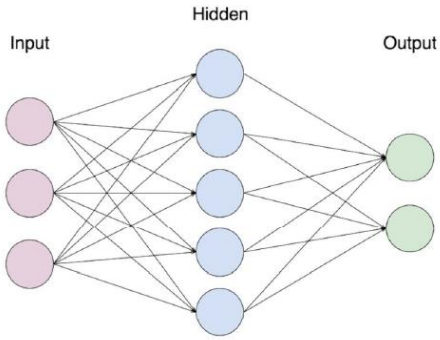

Fig. 6. Deep Neural Network Architecture

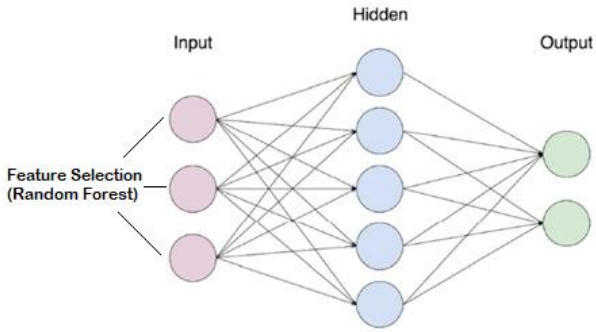

Fig. 7. The hybrid model of Random Forest and Deep Neural Network Architecture

RF is a supervised learning machine learning algorithm used in classification and regression. The RF ensemble learning approach unites groups of "weak learners" to form "strong learners", metaphor a collection of trees to form a forest. In this study, the machine learning model was constructed with the aid of the library Scikit-learn. The crucial parameters of RF are the number of trees in the forest and the number of random features considered for the best split. The Python language library's terms are represented by "n_estimators" and "max_features" respectively. The values of "n_estimators" normally applied are 16, 64, 100, 200, and 500, whereas the value of "max_features" is 5, which approximate to square roots of the 30 features in the datasets. In this study, 200 "n_estimators" and 5 "max_features" are applied to investigate the presence of positive nick sites in the native protein. The values of the parameters used are decided from the best results of the parameter tuning process.

SVM is a supervised learning algorithm used for classification and regression as its high potential in the high level of generalization. SVM works by finding the hyperplane, the decision boundaries that help classify the data points. SVM can take low-dimensional input space data and transform it into high-dimensional data with a kernel function. The Radial Basis Function (RBF) kernel is used with SVM to classify the protease substrate data in this research. Similar to RF, the machine learning model was constructed with the aid of the library Scikit-learn. There are two vital parameters to get an ideal SVM model: penalty parameter "C" and parameter "gamma". Parameter "C" controls the sample error's punishment degree; meanwhile, parameter "gamma" depicts the influence of a single training data. As there are no golden rules for both values of parameters, thus various values are applied by the previous researchers. The values of parameter "C" applied from 0.1, 1, 10, 100 to 1000, whereas the values of parameter "Gamma" used are 0.001, 0.01, 0.1, 1, 10, 100, and 1000. This study applies 10 "C" and 10 "gamma" values to identify positive nick sites in the native protein. The values of the parameters used are decided from the best results of the parameter tuning process.

DNN is another popular algorithm that consists of neural network (NN) architecture. It is also called deep learning, which is a subfield of machine learning. NN has the building block neurons that are interconnected to make up layers in the network. There are two layers in NN there are input layer and the output layer. The input layer perceives the input to feed into the network, whereas the output layer propagates the information as the final output. The NN is called "deep" when there is/are extra hidden layer(s) in between of input layer and output layer. The hidden layer consists of numerous hidden neurons that connect the input neuron and output neuron. The presence of the hidden layer enabled NN to learn more "deeply," especially for the high dimensional data.

Since the machine learning model is more complex than RF and SVM, many parameters must be tuned. However, similar to previous machine learning models, there are two crucial parameters of DNN, which are the number of layers and the number of neurons in each layer. Several studies concluded that the ideal number of hidden layers is two if compared to one, three, or more (Li *et al*., 2019). However, there are no golden rules on the number of neurons in each layer. An input layer

with 30 neurons constructs the proposed model in this study (as there are 30 features), followed by 2 hidden layers with 10 neurons and an enclosure with an output layer with a neuron.

There are numerous parameters to be considered to build an ideal model of DNN, such as activations functions, dropout, batch size, epochs, optimizer, and loss function. An activation function is a function that maps the input nodes to the output nodes with the unit's activation (Imdad *et al*., 2018). Meanwhile, the dropout value is a preventive approach to overfitting problems by randomly assigning zero to a unit out during the training process. The batch size is the number of data fed into modal in each computation, and epochs represents an iteration over the entire training data provided and calculate the weights and biases of the neural network. DNN also applied an optimizer to enhance the performance by tuning parameters and the loss function to measure the actual and the predicted value. In this study, the activations functions used are ReLu in hidden layers and sigmoid for the output layer. The dropout value was fixed at 20% with batch size 16. Epochs were set for 1000, and the Adam optimizer was employed. The values of the parameters used are decided from the best results of the parameter tuning process.

The proposed machine learning algorithm is expected to reduce time execution and enhance the accuracy of the proposed model. The algorithm is built based on the dimensionality reduction concept. Although some features will be lost throughout the process of dimensionality reduction, however in return, that training time has been saved as well as computational resources. Thus, the overall performance of the algorithms had drastically improved. Besides overcoming the curse of dimensionality, the algorithm also gets rid of the problem of overfitting. However, the proposed algorithm unlike the typical dimensionality reduction method that finds a new combination of new features, the improved algorithms keep the most important features and remove redundant features through the RF algorithm implementation. The hybrid framework initiates with the RF algorithm in finding the most important feature by ranking 30 features available through feature importance score. The threshold of feature importance score is 0.035, which means that the feature that scores above the threshold is considered necessary and labelled as a selected feature. The set features are then built up a new dataset, namely a feature-selected dataset (apply to all three datasets in use). The feature-selected datasets are then fit into the DNN model to undergo the training process. Table III illustrates the datasets' features before and after the feature selection process.

TABLE III. DATASETS' FEATURES BEFORE AND AFTER THE FEATURE SELECTION PROCESS

Dataset	Before Feature Selection	After Feature Selection
A	30	14
B	30	10
C	30	15

From Table III, the feature-selected dataset A had reduced its features from 30 to 14, whereas the feature-selected dataset B only left one-third out of 30 features which is important. For feature-selected dataset C, half of the features with a total of 30 had been filtered, left 15 important features. As the proposed algorithms had reduced the data dimensionality (number of features), and thus the training process will speed up due to fewer features being fd into the algorithm. Due to the time constraint, the researcher carried out the parameter tuning process for the machine learning algorithms only (RF, SVM, and DNN). As the improved deep neural network makes use of RF and DNN, and both algorithms had been tuned before the classifier was trained, the researcher takes the assumption that the tuned parameters are satisfied, and they are also applicable to the hybrid model.

E. Performance Measurement

In most studies, the golden rule for the performance measure matrix used is the model's accuracy. Since the datasets applied are imbalanced, it is not suitable to use accuracy as the matrix, although SMOTE oversamples it. To replace the accuracy matrix, the ideal matrices used are confusion matrices, and from the confusion matrices able to deduce other performance matrices such as specificity, recall, precision, etc. Table IV tabulated the information of the confusion matrix.

TABLE IV. CONFUSION MATRIX

Actual	**Predicted**	
	Negative Class	**Positive Class**
Negative	TN	FP
Positive	FN	TP

A confusion matrix is a matrix that describes the output and the complete model's performance by depicting the instances whether they are correctly and incorrectly labeled. As tabulated, there are four terminologies used in the confusion matrix: true negative (TN), false negative (FN), false positive (FP), and true positive (TP). The definition of the terms is described in the following:

- **TN**: The predicted instances are NO and the actual output is also NO.
- **FN**: The predicted instances are NO and the actual output is YES.
- **FP**: The predicted instances are YES and the actual output is NO.
- **TP**: The predicted instances are YES and the actual output is also YES.

Area Under Curve (AUC) represents the area under plot specificity versus sensitivity curve at different points. Specificity is the measurement of the distribution of actual negatives that are correctly identified, while sensitivity is the distribution of true positives that are correctly identified. Precision is used to express the number of actual positive results divided by the number of positive results predicted by the classifier, while recall is used to represent the number of actual

positive results divided by the number of all samples that should have been identified as positive.

F1-score is the harmonic mean between recall and precision. F0.5-score is another F-measure that emphasizes the importance of precision and lowers the importance of recall, whereas F2-score is another F-measure that emphasizes the importance of recall and lowers the importance of precision. In this research, there are few F-score (F0.5-score, F1-score, and F2-score) in use. However, there is no golden rule on which F-score is the best. The performance of each classifier determines the best F-score for each case. If the score of the classifier for precision and recall are slightly balanced, F1-score is the best F-score measure for that case. However, if the performance of the classifier is better in precision, thus F0.5-score is the best F-score measure. F2-score appears as the best F-score measure when the classifier performs well in recall score. The formula of the performance measurement mentioned are listed in the following equation below:

$$AUC = \frac{(1 + \mathrm{TP} - \mathrm{FP})}{2} \tag{1.1}$$

$$Precision = \frac{\sum_{i=1}^{m} \frac{TP_i}{TP_i + FP_i}}{m} \tag{1.2}$$

$$Specificity = \frac{\sum_{i=1}^{m} \frac{TN_i}{FP_i + TN_i}}{m} \tag{1.3}$$

$$Sensitivity\ (recall) = \frac{\sum_{i=1}^{m} \frac{TP_i}{TP_i + FN_i}}{m} \tag{1.4}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{1.5}$$

$$F0.5 - score = \frac{1.25 \times Precision \times Recall}{0.25 \times Precision + Recall} \tag{1.6}$$

$$F2 - score = \frac{5 \times Precision \times Recall}{4 \times Precision + Recall} \tag{1.7}$$

III. Results and Discussion

The results of the data preparation are obtained. Data pre-processing had removed the unnecessary columns as mentioned in the methodology. In the end, the dataset is well-prepared, leaving 30 features and a class. After SMOTE implementation, the prediction results for each machine learning classifier are tabulated in Table V. The performance of the four classifiers is evaluated to seek the most performed classifier for each dataset and the overall performance. In Table V, the F-score does not specify which type of F-score. F1-score may not be the best F-score for the classifier in the dataset. It might be the F0.5-score or F2-score is the best for the case. Thus, the F-score refers to the best F-score in that case.

TABLE V. Classification results for RF, SVM, DNN, and Hybrid RF+DNN

Classifier	Dataset	Precision (%)	Recall (%)	F-score (%)	AUC	Confusion Matrix
SVM	A	0.18	0.27	0.25	0.61	[[1920 106] [64 24]]
	B	0.86	0.35	0.67	0.67	[[749 5] [59 32]]
	C	0.04	0.88	0.47	0.63	[[21658 18] [69 25]]
RF	A	0.23	0.23	0.23	0.60	[[1959 67] [68 20]]
	B	0.68	0.42	0.60	0.70	[[736 18] [53 38]]
	C	0.04	0.88	0.18	0.90	[[19815 1861] [11 83]]
DNN	A	0.07	0.60	0.24	0.64	[[1469 557] [43 45]]
	B	0.16	0.62	0.39	0.60	[[454 300] [35 56]]
	C	0.01	0.47	0.03	0.59	[[14860 6816] [50 44]]
Hybrid RF+DNN	A	0.07	0.64	0.25	0.64	[[1323 703] [32 56]]
	B	0.19	0.64	0.43	0.65	[[499 255] [33 58]]
	C	0.01	0.53	0.03	0.58	[[13619 8057] [44 50]]

AUC is an ideal parameter in measuring the result because these parameters measure a classifier's ability to differentiate the class. This study wants to highlight the performance of the classifiers in classifying the functional and non-functional peptidase cleavage sites, which is directly proportional to the AUC scores. It means that the higher the AUC score, the better the model's performance to distinguish the positive and negative nick site.

To summarize the performance of classifier models in table V, the improved DNN classifier is expected to be the best classifier is not achieved. Based on the AUC score, the overall performance for all datasets is the best with the RF classifier. The RF classifiers have obtained 0.60, 0.67, and 0.90 for Datasets A, B, and C, respectively. For SVM, DNN and Hybrid RF+DNN, the AUC value obtained is quite low, which induced the poor classification performance. As for datasets, the best results that are produced by all classifiers is Dataset B. Lastly, the best steps to select the negative samples is by choosing the top 10 sites of the substrate.

In a nutshell, the hybrid RF+DNN algorithm is an alternative machine learning algorithm available in classifying the proteases nick sites. More efforts must be made to achieve the expectation, such as parameter tuning for the improved DNN. This is because this study assumed that the parameters had been tuned during RF and DNN classifier, and there is no

need for parameter tuning for an improved DNN classifier. However, the results obtained had rejected the assumptions. The parameter for RF must be tuned again as the purpose are feature selection, and not for classification. Next, the gathered feature-selected and original datasets have different numbers of features, and thus parameter tuning must be carried out again. The feature-selected datasets are only available fitting into classification algorithms after the parameter has been tuned.

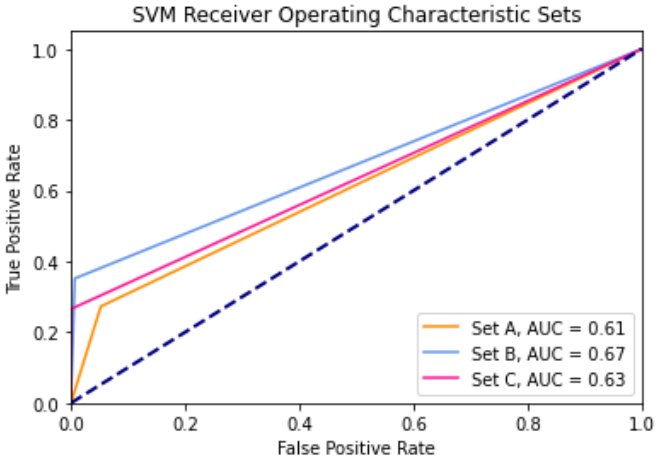

Fig. 8. ROC curve of SVM classifier in three datasets A, B, and C

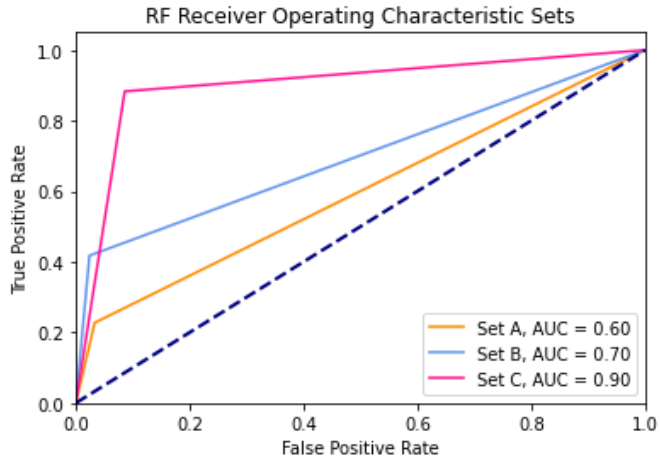

Fig. 9. ROC curve of RF classifier in three datasets A, B, and C

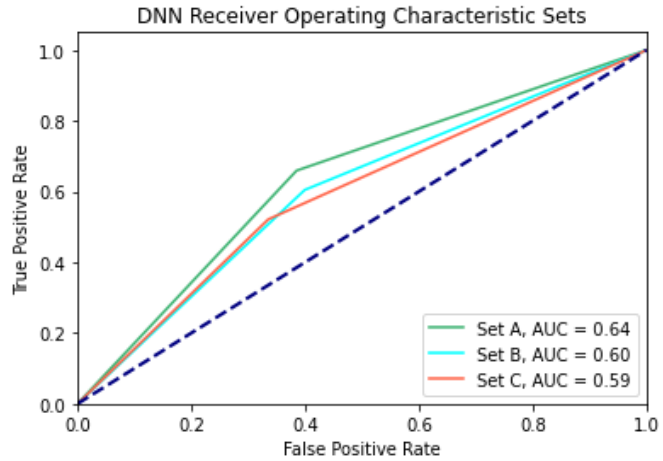

Fig. 10. ROC curve of DNN classifier in three datasets A, B, and C

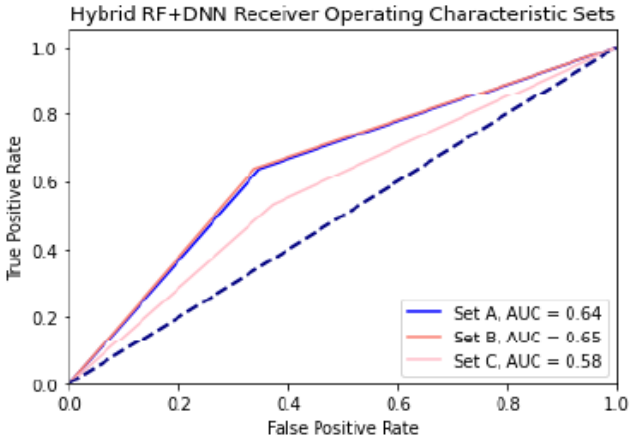

Fig. 11. ROC curve of Hybrid RF+DNN classifier in three datasets A, B, and C

In conclusion, to generalize the performance of all the classifiers with all the available datasets, RF appears as the best performer, as the algorithm scores higher AUC score in general.

IV. Conclusions

In this study, experiments have been performed to seek the ideal parameters for each machine learning classification algorithm. Hybrid RF+DNN algorithms had been implemented to classify the nick site as the outcome of the study. The study was then evaluated by comparing with the other machine learning algorithms such as SVM, RF, and DNN whether the performance is improved. The results from this study can be improved for future research. In the future, more fine-tuning of parameters can be performed to enhance the Hybrid RF+DNN performance. Instead, alternative hybrid algorithms can be introduced to improve the performance of identifying the proteases' nick sites. Besides enhanced machine learning algorithms, and alternative oversampling methods can be applied in the future rather than SMOTE techniques to handle imbalanced datasets. Finally, the latest proteases' nick sites than Hashim's version can be added to the dataset, as there are chances that new nick sites have been identified in recently.

Acknowledgments

The authors would like to express gratitude to the reviewers and editors for helpful suggestions and Universiti Teknologi Malaysia sponsoring this research by the UTM Encouragement Research (Grant Number: Q.J130000.2651.17J51). The study is also supported by the School of Computing, Faculty of Engineering, UTM.

REFERENCES

[1] Imdad, U., Ahmad, W., Asif, M., & Ishtiaq, A. (2018). Classification of Students Results Using KNN and ANN. *Proceedings - 2017 13th International Conference on Emerging Technologies, ICET2017*, *2018-January*, 1-6. https://doi.org/10.1109/ICET.2017.8281651

[2] J. Song *et al.* (2012). PROSPER: An Integrated Feature-Based Tool for Predicting Protease Substrate Cleavage Sites. *PLoS One*, 7(11).

[3] Jia, J., Liu, Z., Xiao, X., Liu, B., & Chou, K. C. (2016). pSuc-Lys: Predict Lysine Succinylation Sites in Proteins with PseAAC and Ensemble Random Forest Approach. *Journal of Theoretical Biology*, 394, 223-230.

[4] Khairuddin, N. B. B. (2019). Implementation of Synthetic Minority Oversampling Technique in Classification of Protease Substrate Cleavage Site. 300.

[5] Li, H., Lu, Y., Zheng, C., Yang, M., & Li, S. (2019). Ground Water Level Prediction for the Arid Oasis of Northwest China based on the Artificial Bee Colony Algorithm and a Back-propagation Neural Network with Double Hidden Layers. *Water (Switzerland)*, 11(4), 1-20. https://doi.org/10.3390/w11040860.

[6] Li, Y., Tian, Y., Qin, Z., & Yan, A. (2018). Classification of HIV-1 Protease Inhibitors by Machine Learning Methods. *ACS Omega*, 3(11), 15837-15849. https://doi.org/10.1021/acsomega.8b01843.

[7] Lu, X., Wang, L., & Jiang, Z. (2019). The Application of Deep Learning in the Prediction of HIV-1 Protease Cleavage Site. *2018 5th International Conference on Systems and Informatics, ICSAI 2018*, *Icsai*, 1299–1304. https://doi.org/10.1109/ICSAI.2018.8599496.

[8] Mótyán, J., Tóth, F., & Tőzsér, J. (2013). Research Applications of Proteolytic Enzymes in Molecular Biology. *Biomolecules*, *3*(4), 923-942. https://doi.org/10.3390/biom3040923.

[9] N. D. Rawlings and A. J. Barrett. (1999). MEROPS: The Peptidase Database. *Nucleic Acids Research*, 27(1), 325-331.

[10] Q. Zhang. (2010). Prediction of Limited Proteolysis from Protein Structure and Sequence. *Life Sci.*

[11] Hashim, H. (2013). Prediction of Peptidase Function via Comparative Modelling. PhD. University of Manchester.

[12] Savojardo, C., Martelli, P. L., Fariselli, P., & Casadio, R. (2018). DeepSig: Deep Learning Improves Signal Peptide Detection in Proteins. *Bioinformatics*, 34(10), 1690-1696. https://doi.org/10.1093/bioinformatics/btx818.

[13] Singh, D., Singh, P., & Sisodia, D. S. (2018). Evolutionary based Optimal Ensemble Classifiers for HIV-1 Protease Cleavage Sites Prediction. *Expert Systems with Applications*. 109, 86-99. https://doi.org/10.1016/j.eswa.2018.05.003.