Araştırma Makalesi / Research Article

Classifying Protein Sequences Using Convolutional Neural Network

Bihter DAŞ[1], Suat TORAMAN[2*]

[1]*Firat University, Department of Software Engineering, Elazığ*
[2]*Firat University, Department of Informatics, Elazığ*
(ORCID: 0000-0002-2498-3297) (ORCID: 0000-0002-7568-4131)

Abstract
One of the major challenges in bioinformatics is the classification and identification of protein structure and function. Large amounts of Ribonucleic Acid (RNA) data cannot be managed using traditional laboratory methods. For this, proteins should be separated according to their structure and families. Therefore, proteins need to be classified to define their biological families and functions. In traditional machine learning approaches, various feature extraction algorithms are used to classify proteins. In manual feature extraction, the selected features directly affect performance. Therefore, in the proposed method of this study, protein sequences were digitized by the amino acid composition technique. The digitized protein sequences were converted to spectrograms, and automatic feature extraction was performed using two-dimensional Convolutional Neural Network (CNN) models (VGG19, ResNet). The extracted features were classified with Support Vector Machine (SVM) and k-Nearest Neighbors (k-NN). As a result, the accuracy of 95.03% was achieved in the classification of protein sequences using ResNet.

Keywords: Protein classification, bioinformatics, convolutional neural network, machine learning.

Evrişimsel Sinir Ağlarını Kullanarak Protein Dizilimlerinin Sınıflandırılması

Öz
Biyoinformatikteki en büyük zorluklardan biri, protein yapısının ve fonksiyonunun öngörülmesi ve sınıflandırılmasıdır. Çok miktarda Ribonükleik Asit (RNA) verisi geleneksel laboratuvar yolu kullanılarak yönetilemez. Bunun için proteinler yapılarına ve ailelerine göre ayrılmalıdır. Bu nedenle proteinlerin biyolojik ailelerini ve fonksiyonlarını tanımlamak için sınıflandırılması gerekmektedir. Geleneksel makine öğrenme yaklaşımlarında, proteinler sınıflandırılırken çeşitli özellik çıkarım algoritmaları kullanılmaktadır. Elle özellik çıkarımında, seçilen özellikler, başarımı doğrudan etkilemektedir. Bu nedenle, bu çalışmada önerilen yaklaşımda ise protein sekanslarını amino asit bileşimi yöntemi ile sayısallaştırılmıştır. Sayısallaştırılan protein dizilimleri spektrograma dönüştürülmüş ve iki boyutlu Evrişimsel Sinir Ağı (ESA) modelleri (VGG19, ResNet) kullanılarak otomatik özellik çıkarımı gerçekleştirilmiştir. Çıkarılan özellikler Destek Vektör Makineleri (DVM) ve k-En Yakın Komşuluk (k-NN) ile sınıflandırılmıştır. Sonuç olarak, ResNet kullanılarak gerçekleştirilen protein sekanslarının sınıflandırma işleminde %95.03'lük bir doğruluğa ulaşılmıştır.

Anahtar kelimeler: Protein sınıflama, biyoinformatik, evrişimsel sinir ağları, makine öğrenmesi.

1. Introduction

Proteins that consist of a long chain of amino acids are important parts of biological processes. Proteins perform important functions, move molecules from one place to another, and increase Deoxyribonucleic Acid (DNA). The proteins have been classified into protein families and superfamilies. Recently, large scale experiments and projects cause a large increase in the number of biological data such as protein and DNA. Many new and unknown proteins are available. It is difficult to recognize a large number of protein families by traditional methods and to classify the new protein molecule [1]. A rapid growth in the number of protein sequences has increased efforts to find appropriate and reliable methods to analyze

*Corresponding author: storaman@firat.edu.tr
Received: 21.12.2019, Accepted: 09.04.2020

protein sequence data [2, 3]. Machine learning algorithms such as Naive Bayes, Decision Trees, Support Vector Machines, and Artificial Neural Networks are traditional methods that are used for the classification of protein molecules [4-6]. The most important thing in the problem of protein classification is feature selection. The protein properties are based on the protein surface, amino acid sequence, or functions. The used methods are aimed to classify the structural proteins. Therefore, it will be easier to understand the evolutionary and functional relationships between proteins. Various computational methods have been proposed to estimate structural classes from sequence information using some statistical methods [3, 7, 8]. Figure 1 shows four structures of the proteins.

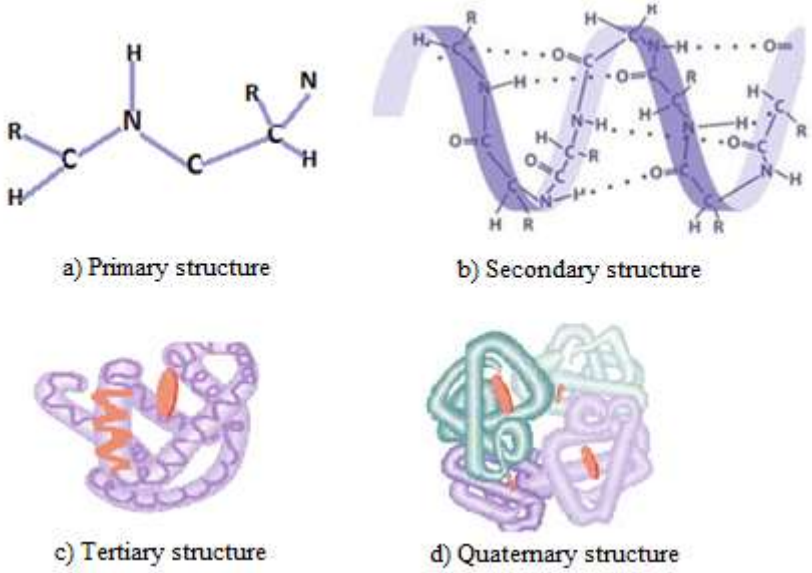

Figure 1. Protein Primary, Secondary, Tertiary and Quaternary Structure formation

Many researchers proposed some approaches to deal with the protein sequence classification problem [9]. Tsuda et al. [10] proposed an efficient method based on multiple protein networks to classify protein sequences. For selecting important features, it assigned weights to multiple networks. Huang, et al. [11] presented a new method for protein classification using the hydropathy blocks that occur in protein sequences. They generated a feature vector for each protein sequence using the frequency of the hydropathy blocks [10, 11]. In the study of Maqsood Hayat et al. [12] proteins were used for classification, and features were extracted using two different techniques. These feature extraction techniques were the PAAC method and Discrete Wavelet Analysis. They also used k-nearest neighbor, neural networks, support vector machine, random forest methods to classify proteins [13]. Mansoori et al. [14] proposed a classifier based on fuzzy-rule for protein classification and a method that is alignment-free. They used a different method to determine the best features for training. When their classifier was compared with a classifier based on a non-fuzzy rule, their classifier had better results [13]. Lacey et al. [15] proposed two methods aligned against the protein sequence using raw sequence data. They proposed two methods. One of them was the Hidden Markov Model. It was used to directly align protein sequences. The other one was the Random Forest model. It was used to extract metadata from protein sequences, divided data into user classes, and distinguished between two protein groups based on properties such as amino acid groups and Physico-chemical properties [13]. Iqbal et al. [16] proposed a feature selection method based on the statistical metric. In the proposed method, they used the n-gram method to extract features and then applied a statistical metric approach to discard insignificant features from the protein sequences. Their results showed that the feature selection method based on statistical metrics performed well [13].

In this study, we carried out a classification that separates supervised and semi-supervised protein sequences without hand-designed feature extraction and selection. We digitized protein sequences by amino acid composition technique and trained them as 2-dimensional spectrogram images in a 2D-CNN model, then classified them using Support Vector Machine and k-NN. For the classification of protein sequences, we achieved an average accuracy of 95.03% for k-NN and an average accuracy of 92.52% for SVM. The paper is organized as follows. In the material and method

section, the data set, convolutional neural networks were mentioned. In the experimental results section, the findings are discussed. In the conclusion, the general contributions of the study are presented.

2. Material and Method

2.1. Dataset

Protein sequences were taken from the NCBI dataset. The dataset, which belongs to the supervised dataset, categorizes three proteins such as superfamilies globin, trypsin, and ras [13]. Table 1 shows the dataset.

Table 1. The dataset

Superfamily Name	Number of Protein Sequence
Globin	395
Ras	337
Trypsin	254

2.2. Amino acid composition

We digitized protein sequences with amino acid composition [17] that counts the frequency of number of amino acids occur in protein sequences. *N* represents 20 amino acids such as *A, N, R, E, C, D, H, I, G, K, L, Q, M, P, F, W, S, T, V, Y* in Equation 1.

$$P_{lm} = \frac{count_l(m)}{\sum_{m=1}^{20} count_l(m)} \quad (1)$$

Here, in P_{lm} *m* represent the number of times particular amino acid, and *l* shows protein sequence. This method digitized the protein sequence using the frequency of each amino acid.

2.3. Creating spectrogram images

Each protein sequence needs to be represented by a fixed number of features. We digitized RNA sequences using the amino acid composition technique. As shown in Table 1, the shortest sequence is trypsin. Globin and ras were taken in 254 lengths because of adjusting the data to be equal length. The digitized data were partitioned by the 50-unit and 80-unit sliding window method. The sliding window is moved forward one unit above on the data. While 174 sub-signal sets were obtained with an 80-unit sliding window, 204 signal sets were obtained with a 50-unit sliding window. Spectrogram images were then generated using these signal clusters. When creating spectrogram images, the window width was determined as 8ms (Hamming windowing), the overlap value was 4ms, and the number of Fourier transforms was 512. Spectrogram images were generated using the Viridis color map. Figure 2 shows the spectrogram images of the globin, trypsin, and ras protein sequences after protein sequence preprocessing.

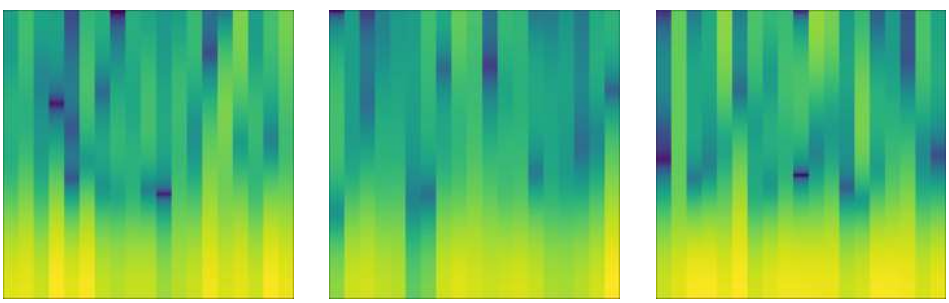

Figure 2. Spectrogram images of globin, trypsin, and ras protein sequences, respectively

2.4. Convolutional neural network

As a result of the rapid development in the Graphics Processing Unit (GPU) technology, significant improvements have been implied in the field of computer vision. Starting with LeNet and then continuing with AlexNet, CNN model design is one of the most widely used deep learning methods today. CNN is used effectively in many areas such as signal/picture classification, object recognition, and tracking. In this study, VGG19 and ResNet models were used for the classification of protein families. Various pre-trained CNN models are available, such as VGG19 and ResNet. The advantage of pre-trained models is that they can effectively extract features in a small data set. This ability is provided to the pre-trained CNN models through training using millions of data. It is not always possible to find large data sets to train a CNN model designed from the beginning. Therefore, a classification process using the weights of pre-trained CNN models gives much better results. This process to be used is called transfer learning. Thus, a small data set in a different field can be classified successfully [18].

2.4.1. VGG19

After LeNet and AlexNet, the model developed by the Visual Geometry Group (VGG) of Oxford University is VGG16. This model consists of 13 convolution layers and 3 fully connected layers, whereas VGG19 has 16 convolution layers and 3 fully connected layers. VGG19 has five maximum pooling layers in 2x2 size. In the last layer, there is a softmax to classify incoming input data [18–20].

2.4.2. ResNet

In parallel with the developments in GPU technology, deep learning architectures have also progressed. With the developing hardware features, the number of layers of network architectures increased. Increasing the number of layers does not mean that the learning will increase. On the contrary, the growth of network architectures causes different problems such as training difficulties. This also causes a loss of input and vanishing gradients [21]. To solve this problem, instead of mapping using the nonlinear *F(x)* function in classical CNN models, ResNet makes a shortcut connection from an input *(x)* to output. As a result, it performs a more effective training process. Figure 3 shows ResNet connection example [22, 23].

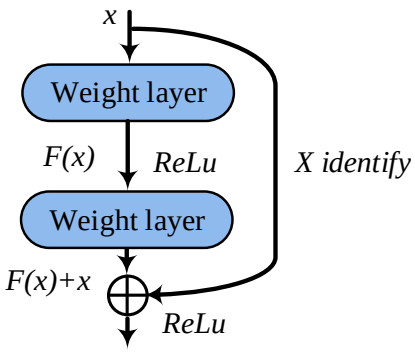

Figure 3. ResNet connection example

2.5. Performance evaluation

Features extracted with ResNet and VGG19 are classified with SVM and k-NN [24, 25]. To evaluate the performance of the proposed method, the *k*-fold cross-validation method was used. *k* value was chosen as 5. As shown in Figure 4, 80% of the data set was used for training, and 20% for testing. After the accuracy of each fold was found, the accuracy of the method was calculated by using the average of 5 folds. The flow diagram of the study is shown in Figure 5.

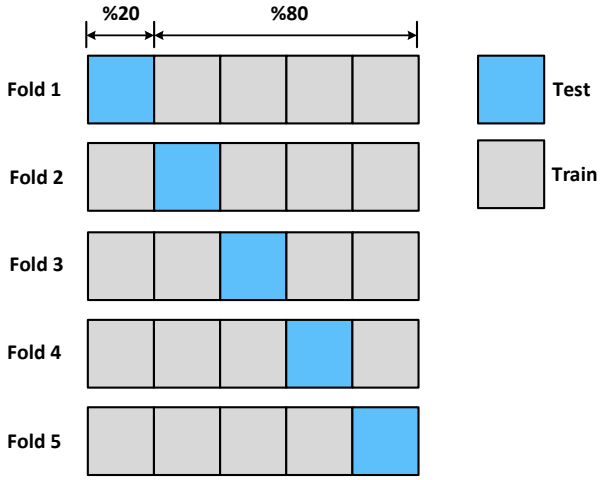

Figure 4. The graphical representation of training and test data

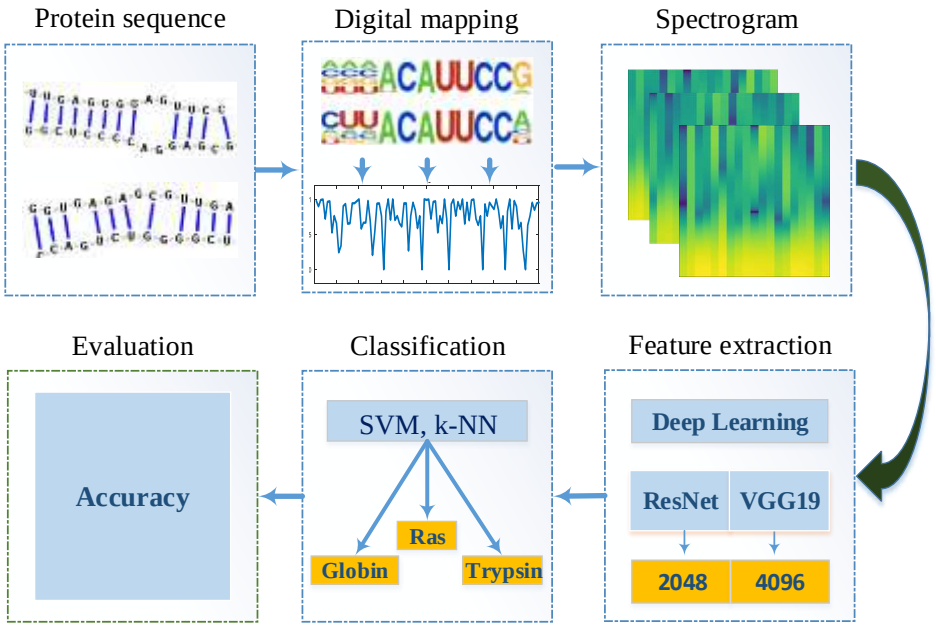

Figure 5. The flow diagram of the proposed method

3. Results and Discussion

254 base-length protein sequences (globin, trypsin, ras) were used in the study. Protein sequences were digitized by the amino acid composition technique. The digitized sequences were partitioned using the 50-unit and 80-unit sliding window method. Each signal segment was converted into a spectrogram image. Spectrogram images are 875x656 pixels in size. These images have been resized to 224x224 pixels for VGG19 and ResNet models. Using the VGG19 and ResNet deep learning models, the characteristics of each spectrogram image were extracted. 4096 vectors with VGG19 and 2048 vectors with ResNet were obtained. Subsequently, the extracted feature vectors were classified with SVM and k-NN. 5-fold cross-validation method was used for the classification more objective. The results of the classification procedures are shown in Table 2.

Table 2. The accuracy values after feature extraction using VGG19 and ResNet

CNN Models	SVM	k-NN
VGG19 (50 units)	84.17 ± 25.69	72.58 ± 10.10
VGG19 (80 units)	87.96 ± 21.76	89.51 ± 11.24
ResNet (50 units)	85.67 ± 16.95	82.02 ± 8.00
ResNet (80 units)	**92.52 ± 22.36**	**95.03 ± 6.69**

As shown in Table 2, the best classification results were obtained with the ResNet model. ResNet has a deeper network structure than VGG19. Thanks to this architectural structure, it was able to extract more effective features. Figure 6 gives the classification accuracy of both models.

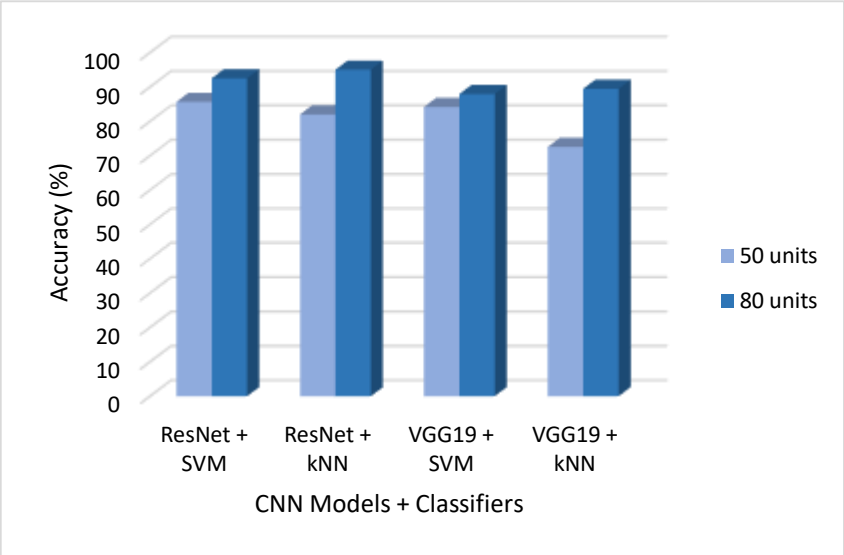

Figure 6. The classification results for 50-unit and 80-unit window width

Increasing the size of the sliding window positively affected the classification accuracy. When both VGG19 and ResNet results are examined, it is seen that the 80-unit length sliding window gives better results. When SVM and k-NN classifiers are compared, it is seen that k-NN has achieved a better performance in the 80-unit sliding window. Also, the best results were obtained with k-NN when all classification results were examined. In Figure 7, the SVM and k-NN classification accuracy of the 80-unit sliding window of both CNN models is given as a comparison with the box graph.

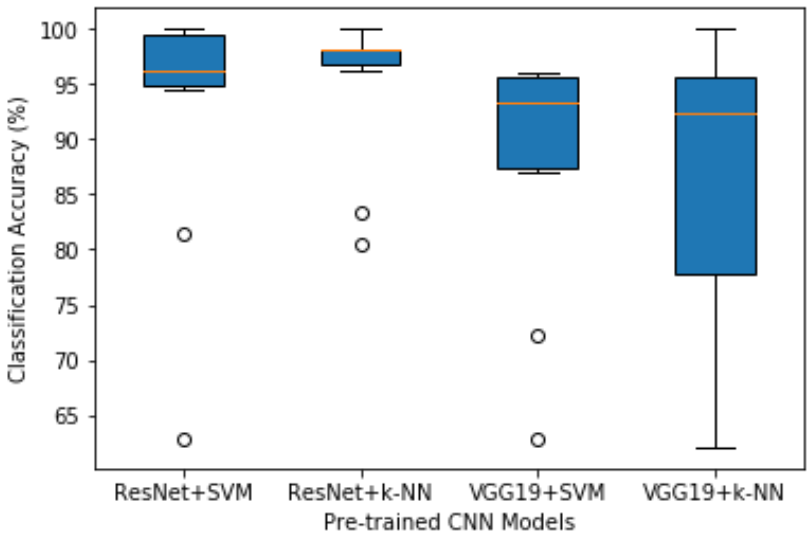

Figure 7. The box graph representation of classification results

Besides, confusion matrices are given in Figure 8 to see which globin, trypsin, and ras protein sequences are correctly recognized by the method. In Figure 8, the classification accuracy that was obtained using the 80-unit sliding window method is given. While k-NN correctly recognized 168 trypsin and 166 ras data of 174 trypsin and ras data, SVM correctly classified 161 trypsin and 159 ras data. In the classification of globin data, while SVM correctly classified 163 globins, k-NN classified 162 data correctly.

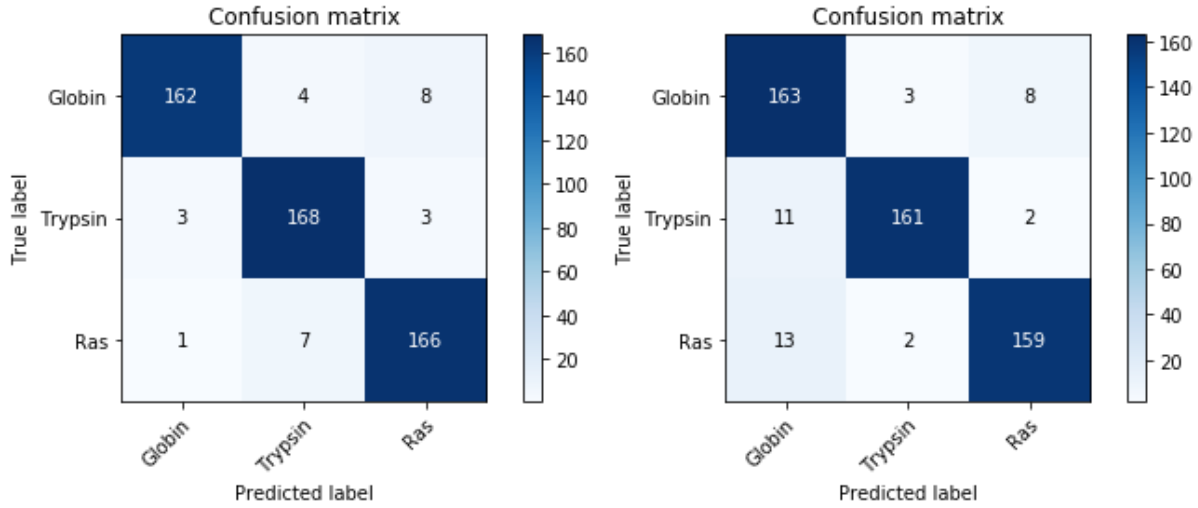

Figure 8. Confusion matrices of k-NN (left) and SVM (right) classification results for feature extraction of the spectrogram images using 80-unit windowing in ResNet

In the classification, the highest accuracy was achieved by the Radial Basis Function (RBF) kernel function of SVM, and the *C* parameter was examined in the range [10^{-3} … 10^{+3}]. For the k-NN classifier, the *k* parameter was examined in the range [1 … 5]. The best classification result was obtained at $k = 1$. Table 3 shows the performance evaluations of different studies for protein classification.

Table 3. Classification results of various protein sequences

Authors	Methods	Dataset	Results (Accuracy)
Lina Yang et al. [2]	Hybrid wavelet fractal method	ND5 protein	More accurate than the existing methods such as Su's model, Zhang's model, etc.
Bharti Chaturvedi et al [13]	Support vector machine	Supervised, semi-supervised protein	The accuracy with 93.20% for semi-supervised
Charalambos Chrysostomou et al.[26]	SVM in combination with Signal processing amino acid indices	BIOV880101 BIOV880102	88.01% with BIOV880101 85.17% with BIOV880102
Igbal M.J. et al. [27]	An optimized tree-classification algorithm	Protein sequences of a specific superfamily	97-98% accuracy
KM Showhat Zamil et al. [28]	Multi-scale local descriptor (MLD)	Helicobacter pylori	55.70% with Decision tree 65.74% with Random Forest 66.67% with Bootstrap aggregation
The proposed method	**ResNet + k-NN**	**Three protein superfamilies globin, trypsin and ras**	**95.03%**

For the experimental applications, a computer with Intel Core i5-8400 CPU and 8 GB RAM was used. The spectrogram data was obtained in MATLAB. Other operations were performed using the Python Keras library.

4. Conclusion

In this study, a novel approach for the classification of RNA sequences is applied to convert the data into spectrogram images by digitizing. Three different protein sequences are classified. In practice, the manual property extraction method was not used as in the traditional methods. Instead, the spectrogram images of protein sequences were automatically extracted by CNN models. Thus, a more effective classification was carried out. It needs to prove that the validity and reliability of the proposed method using larger data sets. In future studies, it is planned to classify protein sequences with larger data sets and different CNN models and methods.

Authors' Contributions

Two authors contributed equally to this work.

Statement of Conflicts of Interest

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The authors declare that this study complies with Research and Publication Ethics.

References

[1] Satpute B., Yadav R. 2018. Machine Intelligence Techniques for Protein Classification. 3rd International Conference for Convergence in Technology (I2CT). IEEE, pp 1–4, 6-8 April, Pune, India.

[2] Yang L., Yan Tang Y., Yang L., Luo H. 2015. A Fractal Dimension and Wavelet Transform Based Method for Protein Sequence Similarity Analysis. IEEE/ACM Trans Comput Biol Bioinforma, 12 (2): 348-359.

[3] Charuvaka A., Rangwala H. 2014. Classifying Protein Sequences Using Regularized Multi-Task Learning. IEEE/ACM Trans Comput Biol Bioinforma, 11 (6): 1087-1098.

[4] Wang D., Huang G.-B. 2005. Protein sequence classification using extreme learning machine. Proceedings of the International Joint Conference on Neural Networks, 31 July-4 Aug, Montreal,Que., Canada.

[5] Bandyopadhyay S. 2005. An efficient technique for superfamily classification of amino acid sequences: Feature extraction, fuzzy clustering and prototype selection. Fuzzy Sets Syst., 152 (1): 5-16.

[6] Ma P.C.H., Chan K.C.C. 2008. UPSEC: An Algorithm for Classifying Unaligned Protein Sequences into Functional Families. J Comput Biol, 15 (4):431-443.

[7] Jaakkola T., Diekhans M., Haussler D. 2000. A Discriminative Framework for Detecting Remote Protein Homologies. J Comput Biol, 7 (1-2): 95-114.

[8] Saigo H., Vert J.-P., Ueda N., Akutsu T. 2004. Protein homology detection using string alignment kernels. Bioinformatics, 20 (11):1682-1689.

[9] Bharill N., Tiwari A., Rawat A. 2005. A Novel Technique of Feature Extraction with Dual Similarity Measures for Protein Sequence Classification. Procedia Comput Sci, 48: 795-801.

[10] Tsuda K., Shin H.J., Schölkopf B. 2005. Fast protein classification with multiple networks. Bioinformatics, 21 (2): 59-65.

[11] Huang D.S., Zhao X.M., Huang G.-B., Cheung Y.M. 2005. Classifying protein sequences using hydropathy blocks. Pattern Recognit., 39 (12): 2293-2300.

[12] Hayat M., Khan A. 2010. Membrane protein prediction using wavelet decomposition and pseudo amino acid based feature extraction. 6th International Conference on Emerging Technologies (ICET). IEEE, pp 1–6,18-19 Oct, Islamabad, Pakistan.

[13] Chaturvedi B., Patil N. 2015. A novel semi-supervised approach for protein sequence classification. Souvenir of the IEEE International Advance Computing Conference, IACC 2015, 12-13 June, Banglore, India.

[14] Mansoori E.G., Zolghadri M.J., Katebi S.D. 2009. Protein superfamily classification using fuzzy rule-based classifier. IEEE Trans Nanobioscience, 8 (1): 92-99.

[15] Lacey A., Deng J., Xie X. 2014. Protein classification using Hidden Markov models and randomised decision trees. 7th International Conference on Biomedical Engineering and Informatics. IEEE, pp 659–664. 14-16 Oct, Dalian, China.

[16] Iqbal M.J., Faye I., Samir B.B., Md Said A. 2014. Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics. Sci World J; 2014: 1-12.

[17] Yang W.-Y., Lu B.-L., Yang Y. 2006. A Comparative Study on Feature Extraction from Protein Sequences for Subcellular Localization Prediction. IEEE Symposium on Computational

Intelligence and Bioinformatics and Computational Biology. IEEE, pp 1–8, 28-29 Sept, Toronto, Ont., Canada.

[18] Gopalakrishnan K., Khaitan S.K., Choudhary A., Agrawal A. 2017. Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection. Constr Build Mater,157: 322-330.

[19] Ullah I., Hussain M., Qazi E.-H., Aboalsamh H. 2018. An automated system for epilepsy detection using EEG brain signals based on deep learning approach. Expert Syst Appl., 107: 61-71.

[20] Documentation K. 2019. Keras. https://keras.io/. (Accessed: 01.10.2019).

[21] He K., Zhang X., Ren S., Sun J. 2016. Deep residual learning for image recognition. IEEE on Computer Vision and Pattern Recognition. pp: 770-778. 27-30 June. LasVegas, NV, USA.

[22] Zagoruyko S., Komodakis N. 2017. Wide residual networks. arXiv 2017;1–15.

[23] Reddy N., Rattani A., Derakhshani R. 2018. Comparison of Deep Learning Models for Biometric-based Mobile User Authentication. IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), 22-25 Oct, Redondo Beach, CA, USA.

[24] Arslan Tuncer S., Akılotu B., Toraman S. 2019. A deep learning-based decision support system for diagnosis of OSAS using PTT signals. Med Hypotheses, 127: 15-22.

[25] Toraman S., Girgin M., Üstündağ B., Türkoğlu İ. 2019. Classification of the likelihood of colon cancer with machine learning techniquesusing FTIR signals obtained from plasma. Turkish J Electr Eng Comput Sci., 27: 1765-1779.

[26] Chrysostomou C., Seker H. 2016. Structural classification of protein sequences based on signal processing and support vector machines. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp 3088-3091, 16-20 Aug, Orlando, FL, USA.

[27] Iqbal M.J., Faye I., Said A.M. 2015. Belhaouari Samir B. Optimized tree-classification algorithm for classification of protein sequences. International Symposium on Mathematical Sciences and Computing Research (iSMSC). IEEE, pp 110–115, 19-20 May, Ipon, Malaysia.

[28] Shawkat Z.M., Rahman J. 2018. Prediction of Protein-Protein Interaction from Amino Acid Sequence Using Ensemble Classifier. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). 8-9 Feb, Rajshahi, Bangladesh.