Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 238 (2024) 876–881

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Symposium on Frontiers in Ambient and Mobile Systems(FAMS)
April 23-25, 2024, Hasselt, Belgium

Protein Sequence Classification Through Deep Learning and Encoding Strategies

Farzana Tasnim[a], Sultana Umme Habiba[b], Tanjim Mahmud[c,*], Lutfun Nahar[d], Mohammad Shahadat Hossain[e], Karl Andersson[f]

[a]*International Islamic University Chittagong, Bangladesh*
[b]*Bangladesh University of Engineering & Technology, Bangladesh*
[c]*Rangamati Science and Technology University, Rangamati 4500, Bangladesh*
[d]*International Islamic University Chittagong, Bangladesh*
[e]*University of Chittagong, Bangladesh*
[f]*Lulea University of Technology, Skelleftea, Sweden*

Abstract

Protein sequence classification is vital for understanding protein functionalities, aiding in the inference of novel protein functions. Machine learning and deep learning algorithms have revolutionized this field, offering insights into specific protein classes and functions. This study employs Natural Language Processing (NLP) techniques, including Integer and Blosum encoding, for efficient classification. SVM with count vectorizer achieves the highest accuracy of 92%, while Integer encoding with CNN surpasses NLP embedding techniques by 4%. The goal is to develop an automated system for predicting protein functionality based on sequence classification, contributing to advancements in proteomics and computational biology.

Keywords: Protein sequence; Natural Language Processing; Deep learning ; Machine Learning ;Feature Extraction; Encoding Technique; Protein Sequence Classification

1. Introduction

Proteins play pivotal roles in biological processes, impacting cell functions and offering insights into personalized medicines and agricultural enhancements [24, 31, 27, 26]. Manual efforts in pharmacological protein sequence classification prove time-consuming [27], necessitating efficient encoding techniques for sequence-function relationships [40, 38, 34]. Natural language processing (NLP) methods[18], applied in structural analysis, enhance bioin-

* Corresponding author. Tel.: +8801818752331.
E-mail address: tanjim_cse@yahoo.com

1877-0509
10.1016/j.procs.2024.06.106

formatics algorithms [37, 11, 28, 42]. Protein sequence classification elucidates superfamily connections and aids sequence matching for functional insights [39]. However, challenges persist in handling vast sequence databases and unknown sequences. Thus, automated classification systems leveraging machine learning[19, 13, 14] and deep learning algorithms[5, 16, 15] become indispensable [28, 42]. This research emphasizes encoding techniques' superiority over NLP feature extraction methods, fostering accurate protein sequence classification .

The key contributions of this study are as follows:

1. Development of an automated system for protein sequence classification.
2. Application of various machine learning and deep learning classification algorithms.
3. Utilization of natural language processing (NLP) feature extraction techniques and encoding methods.
4. Compare encoding techniques and NLP methods, highlighting encoding superiority.

This research reviews prior work in Section 2, outlines the classification methodology in Section 3, and presents detailed results analysis and comparisons with other works in the subsequent sections.

2. Literature Review

Recent research in protein sequence classification has addressed challenges posed by complex biological data. UniProt [2] offers comprehensive protein sequence information, aiding classification tasks. Li M et al. [12] focused on G-Protein Coupled Receptor Protein Sequences, while Parikh et al. [29] achieved 91.6% accuracy using Decision Trees on PDB data. Naveenkumar et al. [25] utilized natural language processing for classification, and Jalal et al. [8] achieved 90% accuracy with deep convolutional neural networks. Saidi et al. [32] employed substitution matrices, while Islam et al. [7] utilized n-grams and skip-grams. Siddha et al. [35] addressed imbalanced data using random undersampling, achieving 78.71% accuracy. Sekhar et al. [33] attained 91% accuracy with naïve Bayes using count vectorizer. Our research introduces space-efficient integer encoding and Blosum encoding, demonstrating superior performance over traditional embedding techniques in deep learning.

3. Proposed Methodology

We have processed the PDB[1] dataset from Kaggle, collected through the Research Collaboratory for Structural Bioinformatics. Various classifiers and feature extraction methods based on natural language processing have been employed. Figure 1 demonstrates the proposed methodology of this research.

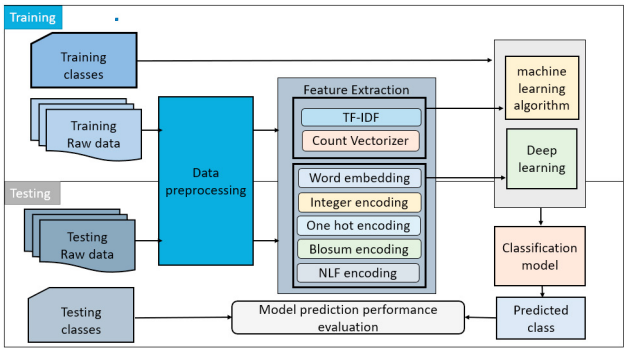

Fig. 1. Proposed Methodology

3.1. Dataset Preprocessing

The protein sequence classification dataset underwent preprocessing to refine it for analysis. Initially comprising 467,305 amino acid sequences, it was filtered to retain 344,956 protein sequences after removing unnecessary data

and ”X” occurrences. The final dataset of (226,693, 2) rows and columns included only the top 20 classes for model training and testing. Tokenization methods treated individual amino acids as character-level tokens, with subword segmentation emerging as a potential approach for motif identification [41, 3].

3.2. Feature Extraction

In protein sequence classification, character-level feature extraction is indispensable for capturing nuanced patterns and variations within amino acid sequences. TF−IDF and count vectorizer used for traditional ML. Deep learning uses embedding, integer encoding, one-hot encoding, blosum, and NLF.

3.2.1. TF-IDF

Character-level TF-IDF (Term Frequency-Inverse Document Frequency) is a method that assigns weights to characters based on their frequency in a document and uniqueness across a collection[20]. The formula is:

$$\text{TF-IDF}(c,d) = \text{TF}(c,d) \times \log\left(\frac{N}{n_c}\right)$$

Here, $\text{TF}(c,d)$ is the character's frequency in document d, N is the total documents, and n_c is the documents containing the character.

3.2.2. Count Vectorizer

Character-level count vectorization represents text by counting the occurrences of individual characters, creating a matrix with each unique character as a column and its frequency in the text. Input Text: ”Hello”. Character-level Count Vectorization: ’h’: 1, ’e’: 1, ’l’: 2, ’o’: 1

3.2.3. Embedding

Character-level embedding transforms text into numerical vectors, capturing information at a finer character level. We have used the tokenize() function to convert characters to numeric values. Our dataset has 25 unique characters, each assigned a unique integer I’: 0, ’Q’: 1, ’F’: 2, ’H’: 3, ’D’: 4, ’P’: 5, ’U’: 6, ’W’: 7, ’T’: 8, ’S’: 9, ’A’: 10, ’Y’: 11, ’L’: 12, ’K’: 13, ’N’: 14, ’E’: 15, ’X’: 16, ’C’: 17, ’O’: 18, ’V’: 19, ’R’: 20, ’B’: 21, ’M’: 22, ’G’: 23, ’Z’: 24. We have padded each sequence with zeroes to maintain a fixed length of 1000.

3.2.4. Integer Encoding

The updated embedding technique transforms categories into numbers. 21 amino acids: 4 rare, 1 unknown labeled as ‘X’. (B,Z,U,O,X) marked as 0 label. Remaining 20 amino acids:’A’: 1, ’C’: 2, ’D’: 3, ’E’: 4, ’F’: 5, ’G’: 6, ’H’: 7, ’I’: 8, ’K’: 9, ’L’: 10, ’M’: 11, ’N’: 12, ’P’: 13, ’Q’: 14, ’R’: 15, ’S’: 16, ’T’: 17, ’V’: 18, ’W’: 19, ’Y’: 20Zero padding is used to achieve the required sequence length. Outperforms embedding techniques.

3.2.5. One Hot with embedding technique

Vectorize sequences with embedding encoding, then pad with 0 to maintain a length of 350 for all sequences. 25 unique numbers represent 25 different amino acids in the embedding technique. ’L’: 1, ’A’: 2, ’G’: 3, ’V’: 4, ’E’: 5, ’S’: 6, ’D’: 7, ’T’: 8, ’I’: 9, ’K’: 10, ’R’: 11, ’P’: 12, ’N’: 13, ’F’: 14, ’Q’: 15, ’Y’: 16, ’H’: 17, ’M’: 18, ’W’: 19, ’C’: 20, ’X’: 21, ’Z’: 22, ’U’: 23, ’B’: 24, ’O’: 25. Each sequence is represented by a (350,25) size vector.

3.2.6. One Hot with Integer encoding

Machine learning faces challenges with categorical attributes, leading to the need for effective encoding techniques. In cases without an ordinal relation, integer encoding may fall short, prompting the use of one-hot encoding for improved performance[4].

3.2.7. BLOSUM Encoding

BLOSUM (Block Substitution Matrix) represents protein sequence alignment by highlighting conserved regions with high alignment scores. It involves calculating odd log scores for pairs of the 20 amino acids within these areas. The transformation of (350, 21) matrices into BLOSUM matrices is achieved using the BLOSUM62 algorithm.

3.2.8. NLF Encoding

NLF generates (350, 21) matrices for each text, leveraging amino acid physiochemical properties and employing the Fisher transform for enhanced feature representation. This technique captures characteristics more effectively than the original amino acid sequences.

3.3. Machine Learning and Deep Learning Classifiers

The study explores diverse classifiers and encoding techniques for protein sequence classification, aiming to automate predictive models for protein function prediction [43, 10, 30, 6]. Techniques include traditional methods like TF-IDF and count vectorizer, alongside natural language processing techniques. Various encoding algorithms are employed, such as Word Embedding, Integer encoding, One hot representation, BLOSUM, and NLF. Models like Convolutional Neural Network and ProtCNN are evaluated for performance on test datasets, advancing protein and genome research.

4. Experimental Results and Analysis

The experiment employed the PDB dataset, comprising 226,693 protein sequence samples post-preprocessing. Addressing imbalances, random undersampling was utilized, focusing on the top 20 classes: HYDROLASE, TRANSFERASE, OXIDOREDUCTASE, LYASE, IMMUNE SYSTEM, TRANSCRIPTION, TRANSPORT PROTEIN, SIGNALING PROTEIN, ISOMERASE, VIRAL PROTEIN, LIGASE, PROTEIN BINDING, DNA, STRUCTURAL GENOMICS, MEMBRANE PROTEIN, TRANSFERASE INHIBITOR, DNA BINDING PROTEIN, RIBOSOME, METAL BINDING PROTEIN.

Evaluation involved hold-out cross-validation, allocating 70% for model training and 30% for assessment. The SVM achieved the highest accuracy at 92% with count vectorization among all machine learning algorithms(See Table 1). With the exception of Neural Network, these classifiers demonstrated superior performance with count vectorization compared to the TF-IDF feature extraction technique.

Table 1. Performance Using Machine Learning Algorithms

Classifier Name	Feature Extraction	Accuracy	Precision	Recall	F1-Score
Naive Bayes	TF−IDF	0.76	0.79	0.76	0.75
	Count Vectorizer	0.83	0.84	0.83	0.83
Logistic Regression	TF−IDF	0.84	0.85	0.84	0.84
	Count Vectorizer	0.91	0.91	0.91	0.91
SVM	TF−IDF	0.91	0.91	0.91	0.91
	Count Vectorizer	**0.92**	**0.92**	**0.92**	**0.92**
Neural Network	TF−IDF	0.91	0.91	0.91	0.91
	Count Vectorizer	0.86	0.88	0.86	0.87

The CNN model, using sequential layers like embedding, 1D convolution, maxpooling, flatten, and dense with 'relu' activation, achieved 90% accuracy with integer encoding(See Table 2). BLOSUM encoding also proved effective, highlighting the importance of encoding methods for protein sequence classification. The study emphasizes the synergy between natural language processing and diverse encoding techniques. Strong correlation between validation and training accuracy demonstrates the model's ability to generalize predictions for unknown sequences. Our best result, 9% accuracy, was with an SVM classifier using count vectorization. Considering memory constraints, sequence lengths were limited to 350. Comparisons with previous studies are challenging due to dataset variations. While previous comparisons typically involved the same number of classes, our study considered the top 20 classes, unlike Sekhar et al. (2021), which focused on the top 10 classes(See Table 3)[32, 33].

5. Conclusion and Future Work

This paper explores the use of natural language processing techniques to enhance classifier performance in protein sequence classification, aiming to develop an automated system for predicting unknown protein sequences. SVM

Table 2. Performance Using Deep Learning Algorithms

Classifier Name	Feature Extraction	Accuracy	Precision	Recall	F1-Score
CNN	Embedding	0.86	0.86	0.86	0.86
	Integer encoding	**0.9**	**0.9**	**0.9**	**0.9**
	One hot with Embedding	0.86	0.86	0.86	0.86
	One hot with Integer Encoding	0.87	0.87	0.87	0.87
	Blosum	**0.89**	**0.89**	**0.89**	**0.89**
	NLF	0.75	0.75	0.75	0.75
Prot	Embedding	0.82	0.82	0.82	0.82
	Integer encoding	0.85	0.85	0.85	0.85
	One ho with Integer Encoding	0.73	0.74	0.73	0.73
	Blosum	0.73	0.73	0.73	0.73
	NLF	0.73	0.73	0.73	0.73

Table 3. Comparison among the previous works and proposed methodology

Ref	Dataset	Feature Extraction	Algorithm	Accuracy(%)
[29]	PDB(Top 10 classes)		DT, Extra Tree, RF	DT- 91%
[33]	PDB(Top 10 classes)	Count Vectorizer	NB, RF	RF-91% and NB-86%%
[32]	PDB(Top 20 classes)	NLP(Tf-Idf, embedding)	DT, CNN,RF,LSTM	DT-78.7% CNN-75% RF-77%,LSTM-51%
Proposed method	**PDB(Top 20 classes)**	**Tf-idf, count vectorizer, Integer Encoding, Word embedding, One hot, BLOSUM, NLF**	**SVM, NB, NN, Logistic Regression, CNN, ProtCNN**	**SVM with count vectorizer 92% CNN with Integer encoding 90%**

achieves a satisfactory 92% accuracy using count vectorization, while CNN with various encoding techniques also shows strong performance. Future work will explore additional encoding techniques like OETMAP and expand the number of protein sequence classes. Despite better results with machine learning algorithms, performance can be enhanced by incorporating longer sequences, optimization methods[21, 9, 23, 36, 17, 22] and overcoming computational limitations. This research presents opportunities for advancements in medical science and computational biology.

References

[1] , . Structural protein sequences. https://www.kaggle.com/shahir/protein-data-set. [Online; accessed 30-June-2021].

[2] , . Uniprot. https://en.wikipedia.org/wiki/UniProt. Accessed: 2022-06-28.

[3] Asgari, E., McHardy, A.C., Mofrad, M.R., 2019. Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (dimotif) and sequence embedding (protvecx). Scientific reports 9, 1–16.

[4] Brownlee, J., 2017. Why one-hot encode data in machine learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. [Online; accessed 30-June-2021].

[5] Das, S., Mahmud, T., Islam, D., Begum, M., Barua, A., Tarek Aziz, M., Nur Showan, E., Dey, L., Chakma, E., et al., 2023. Deep transfer learning-based foot no-ball detection in live cricket match. Computational Intelligence and Neuroscience 2023.

[6] Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al., 2020. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv preprint arXiv:2007.06225 .

[7] Islam, S.A., Heil, B.J., Kearney, C.M., Baker, E.J., 2018. Protein classification using modified n-grams and skip-grams. Bioinformatics 34, 1481–1487.

[8] Jalal, S.I., Zhong, J., Kumar, S., 2019. Protein secondary structure prediction using multi-input convolutional neural network, in: 2019 SoutheastCon, IEEE. pp. 1–5.

[9] Karim, R., Khaliluzzaman, M., Mahmud, T., et al., 2023. An expert system for clinical risk assessment of polycystic ovary syndrome under uncertainty .

[10] Koumakis, L., 2020. Deep learning models in genomics; are we there yet? Computational and Structural Biotechnology Journal 18, 1466–1473.

[11] Lample, G., Charton, F., 2019. Deep learning for symbolic mathematics. arXiv preprint arXiv:1912.01412 .

[12] Li, J., Wu, J., Chen, K., et al., 2013. Pfp-rfsm: protein fold prediction by using random forests and sequence motifs. Journal of Biomedical Science and Engineering 6, 1161.
[13] Mahmud, T., Barua, A., Begum, M., Chakma, E., Das, S., Sharmen, N., 2023a. An improved framework for reliable cardiovascular disease prediction using hybrid ensemble learning, in: 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE. pp. 1–6.
[14] Mahmud, T., Barua, A., Islam, D., Hossain, M.S., Chakma, R., Barua, K., Monju, M., Andersson, K., 2023b. Ensemble deep learning approach for ecg-based cardiac disease detection: Signal and image analysis, in: 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), IEEE. pp. 70–74.
[15] Mahmud, T., Barua, K., Barua, A., Das, S., Basnin, N., Hossain, M.S., Andersson, K., Kaiser, M. Shamim andSharmen, N., 2023c. Exploring deep transfer learning ensemble for improved diagnosis and classification of alzheimer's disease., in: 2023 International Conference on Brain Informatics, Springer. pp. 1–12.
[16] Mahmud, T., Barua, K., Habiba, S.U., Sharmen, N., Hossain, M.S., Andersson, K., 2024. An explainable ai paradigm for alzheimer's diagnosis using deep transfer learning. Diagnostics 14. URL: https://www.mdpi.com/2075-4418/14/3/345, doi:10.3390/diagnostics14030345.
[17] Mahmud, T., Islam, D., Begum, M., Das, S., Dey, L., Barua, K., 2022. A decision concept to support house hunting. International Journal of Advanced Computer Science and Applications(IJACSA) 13, 768–774.
[18] Mahmud, T., Ptaszynski, M., Eronen, J., Masui, F., 2023d. Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. Information Processing & Management 60, 103454.
[19] Mahmud, T., Ptaszynski, M., Masui, F., 2023e. Automatic vulgar word extraction method with application to vulgar remark detection in chittagonian dialect of bangla. Applied Sciences 13, 11875.
[20] Mahmud, T., Ptaszynski, M., Masui, F., 2023f. Vulgar remarks detection in chittagonian dialect of bangla. arXiv preprint arXiv:2308.15448 .
[21] Mahmud, T., Sikder, J., 2013. Intelligent decision system for evaluation of job offers. 1st National Conferenceon Intelligent Computing and Information Technology (NCICIT), November 21.
[22] Mahmud, T., Sikder, J., Naher, S.R., 2021. Decision support system for house hunting: A case study in chittagong, in: Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2, Springer. pp. 676–688.
[23] Mahmud, T., Sikder, J., Tripura, S., 2018. Knowledge-based decision support system to select hospital location. IOSR Journal of Computer Engineering 20, 39–47.
[24] Monod, J., Changeux, J.P., Jacob, F., 1963. Allosteric proteins and cellular control systems. Journal of molecular biology 6, 306–329.
[25] Naveenkumar, K., Harun, B.R.M., Vinayakumar, R., Soman, K., 2018. Protein family classification using deep learning. bioRxiv , 414128.
[26] Ofer, D., Brandes, N., Linial, M., 2021. The language of proteins: Nlp, machine learning & protein sequences. Computational and Structural Biotechnology Journal 19, 1750–1758. URL: https://www.sciencedirect.com/science/article/pii/S2001037021000945, doi:https://doi.org/10.1016/j.csbj.2021.03.022.
[27] Pandey, A., Roy, S.S., 2022. Protein sequence classification using convolutional neural network and natural language processing", booktitle="handbook of machine learning applications for genomics , 133–144URL: https://doi.org/10.1007/978-981-16-9158-4-9, doi:10.1007/978-981-16-9158-4-9.
[28] Papanikolaou, N., Pavlopoulos, G.A., Theodosiou, T., Iliopoulos, I., 2015. Protein–protein interaction predictions using text mining methods. Methods 74, 47–53.
[29] Parikh, Y., Abdelfattah, E., 2019. Machine learning models to predict multiclass protein classifications , 0300–0304.
[30] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118, e2016239118.
[31] Saha, S., Chaki, R., 2013. A brief review of data mining application involving protein sequence classification. Advances in computing and information technology , 469–477.
[32] Saidi, R., Maddouri, M., Nguifo, E.M., 2010. Protein sequences classification by means of feature extraction with substitution matrices. BMC bioinformatics 11, 1–13.
[33] Sekhar, S.M., Siddesh, G., Raj, M., Manvi, S.S., 2021. Protein class prediction based on count vectorizer and long short term memory. International Journal of Information Technology 13, 341–348.
[34] Shannon, C.E., 1951. Prediction and entropy of printed english. Bell system technical journal 30, 50–64.
[35] Siddha, S.S., 2020. Protein sequence classification using machine learning and deep learning .
[36] Sikder, J., Mahmud, T., Banik, B., Gupta, S., . Linear programming to find the critical path using spreadsheet methodology .
[37] Solan, Z., Horn, D., Ruppin, E., Edelman, S., 2005. Unsupervised learning of natural languages. Proceedings of the National Academy of Sciences 102, 11629–11634.
[38] Strait, B.J., Dewey, T.G., 1996. The shannon information entropy of protein sequences. Biophysical journal 71, 148–155.
[39] Tao, Z., Yang, Z., Chen, B., Bao, W., Cheng, H., 2022. Protein sequence classification with letnet-5 and vgg16, in: International Conference on Intelligent Computing, Springer. pp. 687–696.
[40] Trifonov, E.N., 2009. The origin of the genetic code and of the earliest oligopeptides. Research in microbiology 160, 481–486.
[41] Wang, Y., You, Z.H., Yang, S., Li, X., Jiang, T.H., Zhou, X., 2019. A high efficient biological language model for predicting protein–protein interactions. Cells 8, 122.
[42] Yu, L., Tanwar, D.K., Penha, E.D.S., Wolf, Y.I., Koonin, E.V., Basu, M.K., 2019. Grammar of protein domain architectures. Proceedings of the National Academy of Sciences 116, 3636–3645.
[43] Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al., 2020. Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems 33, 17283–17297.