Conference Paper Title*

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Abstract—**Protein classification is a fundamental task in computational biology, aiding in function prediction and disease analysis. This study uniquely investigates the complementary role of secondary structure information alongside primary sequences in protein classification using machine learning and deep learning models. A comprehensive dataset of 203,591 protein samples from the top 20 protein classes was processed using TF-IDF and embedding-based feature extraction. Seven classifiers, including Random Forest, XGBoost, SVM, Neural Networks, LSTM, Bi-LSTM, and GRU, were evaluated using hold-out cross-validation. Results indicate that primary sequence-based classification achieved the highest accuracy (92%) using Random Forest with TF-IDF, while secondary structure-based classification was comparatively less effective (86%). Combining primary and secondary sequences yielded marginal improvements, suggesting that while secondary structure contributes to feature distribution and biological interpretability, it does not significantly enhance classification accuracy. Notably, oxidoreductase, ribosome, and virus proteins exhibited consistently high classification performance across all models, with 99% accuracy in both primary and combined sequence classification. SHAP analysis highlighted that secondary structure features ('Secondary_hh' and 'Secondary_hhh') contributed to virus protein classification, emphasizing their potential for structural interpretation.**

Index Terms—**Protein sequence classification, secondary structure, machine learning, deep learning, feature extraction, Random Forest, Long Short-Term Memory, SHAP**

I. Introduction

Proteins, often referred to as the building blocks of life, are macromolecules composed of one or more long chains of amino acid residues. They play a fundamental role in various biological processes, including enzymatic reactions, cellular signaling, immune responses, and structural support [4], [6]. The accurate classification of proteins into their respective functional or structural classes is crucial for numerous biological applications, such as drug discovery, functional annotation, motif identification, and disease-related protein characterization [5], [7], [8].

Traditionally, protein classification has relied on manual curation and sequence alignment methods, which are both time-consuming and increasingly impractical given the rapid expansion of protein sequence databases and the presence of unknown or novel sequences. To address these challenges, machine learning (ML) [9]–[11] and deep learning (DL) [12]–[14] have emerged as powerful computational approaches for automated protein classification. These methods can analyze vast amounts of sequence data, capture underlying patterns, and classify proteins with high accuracy.

Most ML- and DL-based classification systems focus on primary amino acid sequences or three-dimensional structural information [16]. However, given that the primary structure serves as the foundation for higher-order folding patterns, it is crucial to evaluate its effectiveness in protein classification. At the same time, the ability of secondary structure information, such as alpha-helices and beta-sheets, to independently classify proteins warrants further investigation [15]. Integrating both primary and secondary structure information raises the question of whether this combination enhances classification performance beyond individual features. Furthermore, determining the key features that contribute the most significantly to the classification of specific protein classes can provide deeper insights into protein function and stability.

In this study, we aim to address these aspects by systematically analyzing the impact of sequence and structural information in ML/DL-based protein classification models. The primary objective is to investigate whether the integration of structural features alongside sequence-based representations improves predictive accuracy and biological interpretability. The key tasks in this work involve encoding protein sequences using TF-IDF-based n-gram representations and incorporating secondary structure elements to develop a hybrid feature set. The classification is performed using machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and Neural Networks. Our findings suggest that while primary sequence features drive high classification accuracy, secondary structure features enhance biological interpretability, particularly in distinguishing virus, ribosomal, and oxidoreductase proteins.

The rest of the paper is organized as follows. Section 2 discusses related work on protein classification, Section 3 outlines the methodology and feature extraction techniques, Section 4 presents the results and key observations, and Section 5 concludes the study with potential future directions.

II. Literature Review

Recent advancements in protein sequence classification have leveraged machine learning (ML) and deep learning (DL) techniques, focusing on both primary amino acid sequences and secondary structural information. Siddha et al. [17] explored traditional ML algorithms and DL models for classifying protein sequences, highlighting the effectiveness of natural language processing techniques in feature extraction to improve classification accuracy. Brandes et al. [18] introduced ProteinBERT, a universal deep-learning model that efficiently learns representations from amino acid sequences, demonstrating versatility across multiple protein classification tasks, even with limited labeled data. Similarly, Wang [19] compared different DL architectures, including bi-directional LSTM and convolutional models, for classifying proteins based on Protein Data Bank sequences, concluding that DL models significantly outperform classical ML approaches. Parikh et al. [20] achieved 91.6% accuracy using Decision Trees on PDB data, while Jalal et al. [21] implemented deep convolutional neural networks (CNNs) to classify protein sequences, achieving 90% accuracy. Islam et al. [22] utilized n-grams and skip-grams to enhance feature extraction, demonstrating the importance of sequence pattern recognition in classification tasks.

Beyond primary sequence classification, several studies have explored the role of secondary structure in protein classification. Yuan et al. [23] proposed an ensemble DL approach combining bidirectional temporal convolutional networks and bidirectional LSTMs to improve secondary structure prediction, thereby enhancing protein classification accuracy. Wang et al. [24] introduced DeepCNF, a deep learning method integrating convolutional neural networks with conditional random fields to predict both 3-state and 8-state secondary structures, achieving high prediction accuracy. Yu et al. [25] developed an end-to-end DL model to predict and design secondary structure content directly from amino acid sequences, providing insights into protein folding patterns and stability. Zhang et al. [26] introduced a novel DL architecture for predicting 8-state secondary structures, offering a more detailed understanding of protein folding mechanisms and contributing to more accurate protein classification.// Our research evaluates protein sequence classification using both primary and secondary structures, showing a marginal accuracy improvement in different classification models.

III. Proposed Methodology

We collected the Kaggle PDB dataset [1] for the input of the primary sequence and then used DSSP (Dictionary of Secondary Structure in Proteins) [2], [3] to extract the 8-state secondary structure sequence for each structural component. Various classifiers and feature extraction methods were employed.

A. Data Preprocessing

1) Primary Sequence Dataset: Preprocessed the collected dataset to refine it for analysis. Initially, the dataset was available in 2 separate CSV files, which were later merged based on a common attribute “StructureId”. The merged dataset comprised 471,149 amino acid sequences, which were later filtered based on the 7 valid types of macromolecule: Protein, Protein DNA, Protein DNA RNA, Protein RNA, Protein DNA DNA RNA Hybrid, Protein DNA DNA RNA Hybrid, Protein DNA RNA Hybrid and by removing unnecessary data and sequences with “X” occurrences. The obtained dataset comprised 271,170 amino acid sequences where we considered only the top 20 most common protein classes.

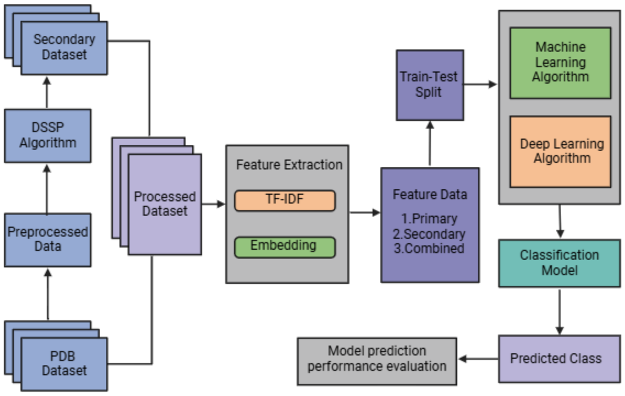

Fig. 1. Proposed Methodology.

2) Structure Sequence Dataset: DSSP (Dictionary of Secondary Structure in Proteins), an algorithm originally designed by Wolfgang Kabsch and Chris Sander to standardize secondary structure assignment based on atomic coordinates was used to obtain the structure sequences of proteins. For a given StructureId of protein, DSSP outputs an 8-state secondary structure sequence comprising: H = α-helix, B = residue in isolated β-bridge, E = extended strand, participates in β ladder, G = 3_{10}-helix, I = π-helix, P = k-helix (poly-proline II helix), T = hydrogen-bonded turn, S = bend, ”-” = None. Each protein has different chains, which may or may not have the same secondary structure sequence. For some proteins, DSSP did not give any outputs.

Thus the final dataset obtained by merging the sequence dataset and structure dataset comprised of (203591, 7) rows and columns included only the top 20 classes for model training and testing.

B. Feature Extraction

Character-level feature extraction is necessary for protein sequence classification for capturing subtle patterns and variations within amino acid sequences. TF-IDF and embedding techniques were used for machine learning and deep learning models.

1) TF-IDF: TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical technique used to evaluate the importance of a character in a document relative to a collection of documents. It combines two metrics:

- Term Frequency (TF): Measures how often a character appears in a document.
- Inverse Document Frequency (IDF): Measures how unique the character is across all documents.

The formula is:

$$TF - IDF(c, d) = TF(c, d) \times \log\left(\frac{N}{n_c}\right) \tag{1}$$

where $TF(c,d)$ is the frequency of character c in the document d, N is the total number of documents and n_c is the number of documents containing the character c.

We used the parameters `max_features` = 1000 and `ngram_range` = (1, 3) for both primary and structural sequences to extract the top 1000 most significant TF-IDF features, considering unigrams, bigrams, and trigrams. While the primary sequences successfully utilized all 1000 features, the structural sequences only provided 550 features.

2) Embedding: Character-level embedding is a representation method that converts text into numerical vectors, capturing detailed contextual and structural information about each character. In this process:

- We have used the `tokenize()` function, which assigns a specific integer value to each unique character in the dataset.
- For primary sequences, we have 24 unique characters, each assigned to a unique integer: {'L': 1, 'A': 2, 'G': 3, 'V': 4, 'E': 5, 'S': 6, 'D': 7, 'I': 8, 'K': 9, 'T': 10, 'R': 11, 'P': 12, 'N': 13, 'F': 14, 'Q': 15, 'Y': 16, 'H': 17, 'M': 18, 'W': 19, 'C': 20, 'U': 21, 'Z': 22, 'B': 23, 'O': 24 } and for secondary structure sequences, we have 9 unique characters, each assigned to a unique integer:{'H': 1, 'E': 2, '-': 3, 'T': 4, 'S': 5, 'G': 6, 'P': 7, 'B': 8, 'I': 9}
- To ensure uniform input size, all sequences are padded with zeros to maintain a fixed length of 350, which is the average sequence length.

3) Classification Models: After feature extraction, the TF-IDF and embedding features from the primary sequence, secondary structure sequence, and their combination were individually applied to machine learning models, including SVM, Random Forest, and XGBoost, as well as deep learning models such as Feedforward Neural Network, LSTM, BiLSTM, and GRU.

A classification report was generated to measure the prediction values per class. The performance of these models was estimated based on accuracy(A), precision(P), recall(R), and F_1-score(F1) as evaluation metrics.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F_1 - Score = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \tag{5}$$

where TP, FP, TN, and FN are the number of true positive, false positive, true negative, and false negative of protein classes respectively.

4) Model Interpretation: Out of the 20 protein classes considered, three classes were selected for detailed analysis based on their performance with the best-performing classifier. These classes were extracted after evaluating multiple models, and the classifier that demonstrated the highest accuracy was chosen for further interpretation.

To gain insights into the model's decision-making process, we applied SHAP (SHapley Additive exPlanations) [29] for feature interpretability. SHAP was used to analyze the contribution of individual sequence features to the classification outcomes. Both primary and secondary sequence-based features were examined to assess their respective impacts on model predictions. This step allowed us to identify which sequence elements played the most significant roles in classification and to compare the relative importance of structural information versus sequence-based information.

IV. Results and Discussion

A. Experimental Setup

The experiment employed the processed dataset, comprising 203,591 protein sequence samples post-preprocessing, focusing on the top 20 classes: HYDROLASE, TRANSFERASE, OXIDOREDUCTASE, LYASE, IMMUNE SYSTEM, TRANSCRIPTION, HYDROLASE INHIBITOR, TRANSPORT PROTEIN, SIGNALING PROTEIN, ISOMERASE, VIRAL PROTEIN, LIGASE, PROTEIN BINDING, STRUCTURAL GENOMICS, MEMBRANE PROTEIN, DNA BINDING PROTEIN, CHAPERONE, RIBOSOME, STRUCTURAL PROTEIN and VIRUS.

The evaluation involved hold-out cross-validation, with 80% allocated for model training and 20% for assessment. The experimental review compared three approaches to protein classification: primary sequence-based, secondary structure-based, and their combination. Seven classifiers were used with TF-IDF and embedding-based feature extraction methods across 20 protein classes.

B. Classification Performance Analysis

1) Primary Sequence-Based Classification: When analyzing primary sequences, the Random Forest classifier emerged as the best performer, achieving 92% accuracy with TF-IDF and a slightly lower 91% with embeddings. Deep learning models, such as LSTM, Bi-LSTM, and GRU, exhibited competitive performance, reaching 90% accuracy with embeddings, highlighting their capability in capturing sequential dependencies. However, SVM performed the worst, with an accuracy of only 60%, demonstrating its limitation in handling high-dimensional sequence data (see Table I). These results suggest that primary sequence information alone provides strong discriminatory power for protein classification, with ensemble and deep learning models yielding the best outcomes.

2) Secondary Structure-Based Classification: Compared to primary sequence analysis, classification based solely on secondary structure generally resulted in lower performance. The Random Forest classifier remained the top performer, achieving an accuracy of 86% with embeddings, while other models

TABLE I
PERFORMANCE METRICS OF DIFFERENT MODELS USING DIFFERENT SEQUENCE TYPE

Classifiers	**Feature**	**Primary Sequence**				**Structure Sequence**				**Combined Sequence**			
		P	R	F1	A	P	R	F1	A	P	R	F1	A
NN	TF-IDF	0.81	0.80	0.80	0.80	0.62	0.60	0.58	0.60	0.81	0.81	0.81	0.81
	Embedding	0.89	0.89	0.89	0.89	0.75	0.74	0.74	0.74	0.88	0.88	0.88	0.88
SVM	TF-IDF	0.61	0.60	0.59	0.60	0.44	0.42	0.37	0.42	0.65	0.65	0.64	0.65
Random Forest	TF-IDF	0.92	0.92	0.92	0.92	0.84	0.84	0.83	0.84	0.93	0.92	0.92	0.92
	Embedding	0.91	0.91	0.91	0.91	0.86	0.86	0.85	0.86	0.91	0.90	0.90	0.90
XGBoost	TF-IDF	0.84	0.83	0.83	0.83	0.71	0.70	0.69	0.70	0.85	0.84	0.84	0.84
	Embedding	0.86	0.85	0.85	0.85	0.74	0.71	0.71	0.71	0.85	0.84	0.84	0.84
LSTM	TF-IDF	0.86	0.86	0.86	0.86	0.67	0.68	0.67	0.68	0.87	0.87	0.87	0.87
	Embedding	0.90	0.90	0.90	0.90	0.83	0.83	0.83	0.83	0.91	0.91	0.91	0.91
Bi-LSTM	TF-IDF	0.86	0.86	0.86	0.86	0.66	0.66	0.65	0.66	0.86	0.86	0.86	0.86
	Embedding	0.90	0.90	0.90	0.90	0.84	0.84	0.84	0.84	0.91	0.91	0.91	0.91
GRU	TF-IDF	0.86	0.86	0.86	0.86	0.68	0.69	0.67	0.69	0.86	0.86	0.86	0.86
	Embedding	0.90	0.90	0.90	0.90	0.82	0.82	0.82	0.82	0.91	0.91	0.90	0.91

saw a notable decline, particularly SVM (accuracy: 42%). This performance gap indicates that secondary structure alone may not be sufficient for precise protein classification (see Table I). The reduced accuracy suggests that while secondary structure plays a role in protein function, it may not always be a strong standalone predictor when compared to primary sequence information.

3) Combined Sequence and Structure Classification: The integration of primary sequence and secondary structure information showed only minimal improvements over using primary sequence alone. Random Forest maintained the same performance (accuracy: 0.92), while deep learning models showed only a marginal increase (0.90 to 0.91). This suggests that secondary structure information may not contribute substantial additional discriminative power for protein classification tasks(see Table I). Embedding-based feature extraction generally performed better than TF-IDF across most models, though the improvements were modest. These results indicate that while both primary and secondary structure information can be useful for protein classification, primary sequence information alone may be sufficient for achieving optimal performance in most cases.

C. Class-Specific Performance Analysis

Across all classifiers, from 20 classes, three classes - OXIDOREDUCTASE, RIBOSOME, and VIRUS - demonstrated superior and consistent performance. Further analysis focused on these classes using a Random Forest classifier with TF-IDF feature extraction. The model achieved remarkable performance across all sequence types, achieving 99% accuracy for both primary and combined sequences and 97% for secondary sequences(see Table II). This consistent high performance across different sequence types suggests the robust discriminative power of the selected features for these three protein classes, potentially due to inherent structural or functional properties that make them easier to classify.

TABLE II
PERFORMANCE METRICS FOR THREE PARTICULAR CLASSES

Classifier	**Feature**	**Sequence Type**	**P**	**R**	**F1**	**A**
Random Forest	TF-IDF	Primary	0.99	0.99	0.99	0.99
		Secondary	0.97	0.97	0.97	0.97
		Combined	0.99	0.99	0.99	0.99

D. Feature Importance Analysis

To understand feature contributions, SHAP analysis revealed distinctive patterns across the three classes and the value of incorporating secondary structure information which is shown in Fig 2 and Fig 3. In oxidoreductase classification, while 'Primay_r' dominated in both models, secondary structure features like 'Secondary_–' emerged as strong predictors in the combined model. For ribosomes, 'Primay_r' maintained the highest impact (SHAP values up to 0.175), with secondary structure elements providing moderate but complementary contributions. In virus classification, the integration of secondary structure features, particularly 'Secondary_hh' and 'Secondary_hhh', alongside primary sequence features like 'Primay_t', suggested that structural information captures unique viral protein folding patterns. Although both models achieved identical accuracy (99.23%), the SHAP analysis demonstrates that incorporating secondary structure information leads to more distributed feature importance and potentially more robust biological interpretability, as it captures protein structural characteristics that complement sequence-based features.

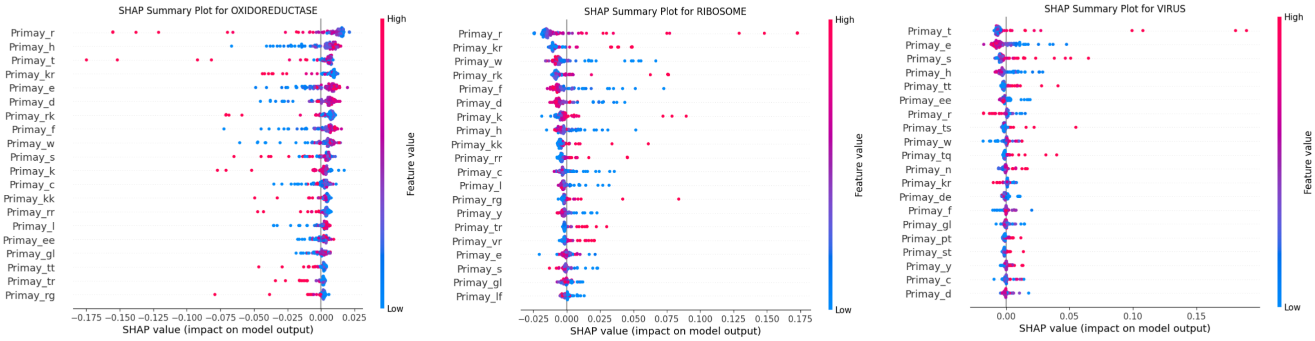

Fig. 2. SHAP Summary plots for three classes using primary sequence.

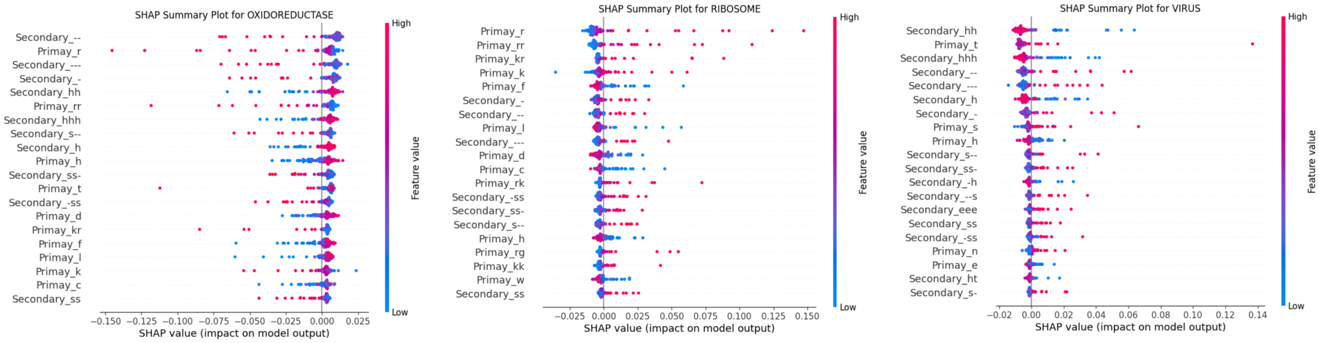

Fig. 3. SHAP Summary plots for three classes using combined sequence.

E. Comparison with Previous Studies

Comparisons with previous studies are challenging due to variations in datasets and the types of sequences used. While earlier studies typically involved the same number of classes and relied solely on primary sequences, our study focused on the top 20 classes and incorporated three different types of sequences, providing a broader perspective. Compared to existing research, our model performed well by incorporating secondary structure sequences in addition to primary sequences(see Table III).

V. Conclusion and Future Directions

This study comprehensively explored protein classification methodologies across 20 protein classes, revealing critical insights into the role of primary sequence and secondary structure information. Among the classifiers used, Random Forest consistently outperformed others, achieving 92% accuracy with TF-IDF feature extraction on primary sequences and maintaining this performance even when secondary structure data was incorporated. At the same time, secondary structure-based classification alone showed limited performance (86% accuracy for Random Forest). Notably, three protein classes - OXIDOREDUCTASE, RIBOSOME, and VIRUS - exhibited exceptional classification capabilities, reaching 99% accuracy across different sequence types. The SHAP analysis further illuminated the nuanced feature contributions, highlighting that secondary structure features like 'Secondary_hh' and 'Secondary_hhh' provide valuable biological interpretability by capturing unique structural patterns, particularly in virus proteins. These findings underscore the complexity of protein classification and the potential for advanced computational approaches in understanding protein characteristics.

For future work, expanding the dataset to include additional protein families and refining secondary structure representations could enhance classification performance. Integrating tertiary structure data or employing advanced deep learning architectures like transformer models may further improve accuracy and interpretability. Additionally, leveraging mutation databases to explore the biological implications of sequence-structure variations could provide deeper insights into disease mechanisms and protein function prediction.

References

[1] Structural protein sequences.https://www.kaggle.com/shahir/protein-data-set. [Online;accessed 20-November-2024].

[2] Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend A series of PDB related databases for everyday needs. Nuc. Acids Res. 2010; 39:D411-D419.

[3] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22:2577-2637.

[4] Whitford, David. Proteins: structure and function. John Wiley & Sons, 2013.

[5] Valavanis IK, Spyrou GM, Nikita KS (2010) A comparative study of multi-classification methods for protein fold recognition. J Bioinform Comput Biol 1: 332–346

[6] Lodish H., Berk A., Kaiser C.A., Krieger M. Molecular Cell Biology. W.H. Freeman and Company; New York, NY, USA: 2004.

TABLE III
COMPARISON WITH PREVIOUS WORK ON PROTEIN SEQUENCE CLASSIFICATION

Ref.	**Dataset**	**Sequence Type**	**Feature Extraction**	**Classifiers**	**Accuracy (%)**
[20]	[1] (Top 10 classes)	Primary	—	DT, RF, ExtraTree	DT: 91%
[27]	[1] (Top 10 classes)	Primary	One-hot with embedding	CNN	CNN: 90%
[28]	[1] (Top 10 classes)	Primary	CountVectorizer	NB, RF	RF: 91%, NB: 86%
[17]	[1] (Top 20 classes)	Primary	TF-IDF, Embedding	DT, CNN, RF, LSTM	DT: 78.7% CNN: 75% RF: 77% LSTM: 51%
[12]	[1] (Top 20 classes)	Primary	TF-IDF, BLOSUM, Integer Encoding, Word embedding, One-hot, NLF, Count vectorizer	SVM, NB, NN, Logistic Regression, CNN, ProtCNN	SVM with Count vectorizer: 92% CNN with Integer encoding: 90%
Proposed	Processed PDB-DSSP dataset (Top 20 classes)	Primary, Secondary, Combined	TF-IDF, Embedding	SVM, RF, NN, LSTM, Bi-LSTM, GRU	RF with TF-IDF: 92% for both primary and combined sequences

[7] Kuksa, Pavel, Pai-Hsi Huang, and Vladimir Pavlovic. ”A fast, large-scale learning method for protein sequence classification.” 8th Int. Workshop on Data Mining in Bioinformatics. 2008.

[8] Iqbal, M. J., Faye, I., Samir, B. B., & Said, A. M. (2013). Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics. The Scientific World Journal, 2014(1), 173869.

[9] Cao, J., & Xiong, L. (2013). Protein Sequence Classification with Improved Extreme Learning Machine Algorithms. BioMed Research International, 2014(1), 103054.

[10] Kaur, K., Patil, N. (2019). A Novel Technique of Feature Selection with ReliefF and CFS for Protein Sequence Classification. In: Sa, P., Bakshi, S., Hatzilygeroudis, I., Sahoo, M. (eds) Recent Findings in Intelligent Computing Techniques . Advances in Intelligent Systems and Computing, vol 707. Springer, Singapore.

[11] Gupta, C. P., Bihari, A., & Tripathi, S. (2019). Human Protein Sequence Classification using Machine Learning and Statistical Classification Techniques. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 2, pp. 3591–3599). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP.

[12] Farzana Tasnim, Sultana Umme Habiba, Tanjim Mahmud, Lutfun Nahar, Mohammad Shahadat Hossain, Karl Andersson, Protein Sequence Classification Through Deep Learning and Encoding Strategies, Procedia Computer Science, Volume 238,2024, Pages 876-881, ISSN 1877-0509,

[13] Lilhore, U.K., Simiaya, S., Alhussein, M. et al. Optimizing protein sequence classification: integrating deep learning models with Bayesian optimization for enhanced biological analysis. BMC Med Inform Decis Mak 24, 236 (2024).

[14] Mahmud, Tanjim & Barua, Anik & Islam, Dilshad & Hossain, Mohammad & Chakma, Rishita & Barua, Koushick & Monju, Mahabuba & Andersson, Karl. (2023). Ensemble Deep Learning Approach for ECG-Based Cardiac Disease Detection: Signal and Image Analysis. 70-74.

[15] Edwards, Hannah & Deane, Charlotte. (2015). Structural Bridges through Fold Space. PLoS computational biology.

[16] Gordeev, A. B., Kargatov, A. M., & Efimov, A. V. (2010). PCBOST: Protein classification based on structural trees. Biochemical and Biophysical Research Communications, 397(3), 470-471. .

[17] Siddha,S.S.,2020.Protein sequence classification using machine learning and deep learning.

[18] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022, Pages 2102–2110,

[19] Wang, Aaron. ”Deep learning methods for protein family classification on PDB sequencing data.” arXiv preprint arXiv:2207.06678 (2022).

[20] Parikh, Yash, and Eman Abdelfattah. ”Machine learning models to predict multiclass protein classifications.” 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE, 2019.

[21] Jalal, Shayan Ihsan, Jiling Zhong, and Suman Kumar. ”Protein secondary structure prediction using multi-input convolutional neural network.” 2019 SoutheastCon. IEEE, 2019.

[22] Islam, SM Ashiqul, et al. ”Protein classification using modified n-grams and skip-grams.” Bioinformatics 34.9 (2018): 1481-1487.

[23] Yuan L, Ma Y and Liu Y (2023) Ensemble deep learning models for protein secondary structure prediction using bidirectional temporal convolution and bidirectional long short-term memory. Front. Bioeng. Biotechnol. 11:1051268.

[24] Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports, 6(1), 1-11.

[25] Yu, H., Chen, W., Chiang, H., Guo, K., Moldes, Z. M., Kaplan, D. L., & Buehler, M. J. (2022). End-to-End Deep Learning Model to Predict and Design Secondary Structure Content of Structural Proteins. ACS Biomaterials Science & Engineering, 8(3), 1156.

[26] Zhang, B., Li, J. & Lü, Q. Prediction of 8-state protein secondary structures by a novel deep learning architecture. BMC Bioinformatics 19, 293 (2018).

[27] Shinde, A. and D'Silva, M. (2019). Protein Sequence Classification using Natural Language Processing, International Journal of Engineering Development and Research, pp. 169-175

[28] Sekhar, S.R.M., Siddesh, G.M., Raj, M. et al. Protein class prediction based on Count Vectorizer and long short term memory. Int. j. inf. tecnol. 13, 341–348 (2021).

[29] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2019). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56-67.