ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

PROTEIN FAMILY CLASSIFICATION USING DEEP LEARNING TECHNIQUES

[1]MUTYALA ANAND VARDHAN, [2]Y SRINIVAS RAJU

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Assistant Professor**,** Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

Protein family classification is a fundamental task in bioinformatics, essential for understanding protein functions, structures, and evolutionary relationships. With the rapid growth of biological data, traditional sequence alignment methods have become less efficient and time-consuming. This project proposes a deep learning-based approach for protein family classification using advanced neural network architectures such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The system utilizes protein sequence data obtained from the SideChainNet dataset, which includes detailed information such as sequences, coordinates, and structural angles. Due to the large size of the dataset, preprocessing and feature extraction are performed using GPU-based platforms like Google Colab, and the processed data is stored in NumPy format for efficient reuse. The dataset is split into training and testing sets in an 80:20 ratio to evaluate model performance. Each deep learning model is trained on the extracted features to learn sequential patterns within protein sequences. The models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

Experimental results show that the RNN model achieves the highest accuracy of 95%, followed by LSTM with 94% and GRU with 93%. The system also provides a user interface for dataset processing, model training, and family classification of new protein sequences. This approach offers an efficient and scalable solution for protein classification, contributing to advancements in computational biology and drug discovery.

Keywords : Protein Classification, Deep Learning, RNN, LSTM, GRU, Bioinformatics, SideChainNet, Sequence Analysis, Neural Networks, Computational Biology

I. INTRODUCTION

Proteins are fundamental biological molecules that play a crucial role in various cellular processes, including metabolism, signaling, and structural support. Understanding protein functions and their relationships is essential in

fields such as bioinformatics, drug discovery, and disease diagnosis. One of the key tasks in bioinformatics is protein family classification, which involves grouping proteins based on their structural and functional similarities. Traditional methods such as sequence alignment and manual analysis are time-consuming and often fail to handle the rapidly growing volume of protein data. These limitations have led to the adoption of computational approaches for efficient and accurate protein classification.

With the advancement of deep learning techniques, new opportunities have emerged for analyzing biological sequences. Deep learning models such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) are particularly well-suited for sequence-based data due to their ability to capture temporal dependencies and patterns. These models can automatically learn complex relationships within protein sequences without the need for manual feature engineering. By leveraging these techniques, it is possible to significantly improve classification accuracy and reduce computational complexity compared to traditional methods.

In this project, a deep learning-based framework is proposed for protein family classification using the SideChainNet dataset. The system involves preprocessing protein sequences, extracting relevant features, and training multiple deep learning models. The performance of these models is evaluated using metrics such as accuracy, precision, recall, and F1-score. The system also provides a user-friendly interface for dataset processing, model training, and classification of new protein sequences. This approach offers a scalable and efficient solution for protein classification, contributing to advancements in bioinformatics and computational biology.

II SURVEY OF RESEARCH

1. Traditional Protein Classification Methods

Early approaches to protein family classification relied heavily on sequence alignment techniques such as BLAST and Hidden Markov Models (HMM). These methods compare protein sequences to identify similarities and evolutionary relationships. While effective for small datasets, they become computationally expensive and time-consuming when dealing with large-scale biological data. Additionally, these methods often struggle to capture complex patterns in sequences, especially when similarities are not obvious. Research has shown that traditional techniques may fail to classify proteins accurately when sequences are highly diverse. These limitations have motivated researchers to explore more advanced computational techniques such as machine learning and deep

learning for improved classification performance.

2. Machine Learning in Protein Classification

Machine learning techniques have been widely used to improve protein classification accuracy. Algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forest have been applied to classify protein sequences based on extracted features. These methods require manual feature engineering, where domain knowledge is used to select relevant attributes from protein sequences. Studies indicate that machine learning models can outperform traditional alignment methods in terms of speed and scalability. However, their performance is highly dependent on the quality of extracted features. This challenge has led to the adoption of deep learning methods, which can automatically learn features from raw data.

3. Deep Learning for Sequence Analysis

Deep learning has revolutionized the field of bioinformatics by enabling automatic feature extraction and improved prediction accuracy. Models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been successfully applied to protein sequence analysis. RNN-based models are particularly effective for sequential data as they can capture temporal dependencies within sequences. Research shows that deep learning models outperform traditional machine learning approaches by learning complex patterns directly from raw sequences. This project utilizes RNN-based architectures to enhance protein family classification performance.

4. LSTM and GRU Models in Bioinformatics

Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are advanced variants of RNN designed to overcome issues such as vanishing gradients. These models are capable of learning long-term dependencies in sequence data, making them highly suitable for protein sequence analysis. Studies have demonstrated that LSTM and GRU models achieve higher accuracy in classification tasks compared to standard RNNs. GRU, being computationally efficient, is often preferred in scenarios with limited resources, while LSTM provides better performance for complex datasets. This project compares the performance of LSTM, GRU, and RNN to identify the most effective model for protein classification.

5. Dataset and Feature Extraction Techniques

The availability of large-scale datasets such as SideChainNet has significantly contributed to advancements in protein classification research. These datasets provide detailed information including sequences, structural coordinates, and angles, enabling more comprehensive

analysis. Feature extraction techniques such as encoding sequences into numerical representations and normalizing data are essential for model training. Research highlights that proper preprocessing and feature extraction significantly impact model performance. In this project, features are extracted using GPU-based processing and stored in NumPy format for efficient reuse, ensuring faster training and improved accuracy.

6. Evaluation Metrics in Protein Classification

Evaluating the performance of classification models is critical in determining their effectiveness. Common metrics used in protein classification include accuracy, precision, recall, and F1-score. These metrics provide insights into how well the model predicts different protein families. Confusion matrices and graphical representations are also used to analyze model performance and identify misclassifications. Research emphasizes the importance of using multiple evaluation metrics to ensure reliable results. In this project, all deep learning models are evaluated using these standard metrics, and comparative analysis is performed to determine the best-performing algorithm.

III. WORKING METHODOLOGY

The proposed system begins with data collection and preprocessing using the SideChainNet dataset, which contains protein sequences along with structural information such as coordinates and angles. Due to the large size and complexity of the dataset, preprocessing is performed using GPU-based platforms like Google Colab to efficiently extract features. The extracted features are then converted into numerical format and stored as NumPy arrays for faster access and reuse. Data normalization is applied to ensure consistency across all input features. The dataset is then divided into training and testing sets in an 80:20 ratio, where the training data is used to build the model and the testing data is used to evaluate its performance.

In the next phase, deep learning models including Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) are implemented. These models are trained using the processed protein sequences to learn patterns and relationships within the data. Each model is optimized using appropriate parameters such as learning rate, batch size, and number of epochs. During training, the models adjust their weights to minimize prediction error. After training, the models are tested on unseen data to evaluate their performance. Metrics such as accuracy, precision, recall, and F1-score are calculated to compare the effectiveness of each algorithm.

Finally, the system performs protein family classification using the best-performing model. Users can upload new protein sequences

through the interface, and the trained model predicts the corresponding protein family. The system also provides visualizations such as confusion matrices and performance graphs to help analyze the results. By combining efficient preprocessing, advanced deep learning models, and user-friendly interaction, the proposed methodology ensures accurate and scalable protein classification suitable for real-world bioinformatics applications.

IV RESULTS EXPLANATIONS

In propose work we are employing deep learning algorithms such as GRU, RNN and LSTM to predict family from protein sequences. To train above models we have used Protein Sequences dataset from SideChainNet which can be downloaded and processes using below code

```
def load_and_thin_casp12_30(thin_factor=10):
    """Load CASP 12-30 data and thin it by selecting every nth protein."""
    # Load SidechainNet data (this will download if not already present)
    data = scn.load(casp_version="12", thinning=30, with_pytorch="dataloaders")  # Initial thinning to reduce memory

    # Get the training data
    train_data = data["train"]

    return train_data
```

Above block of code will load CASP protein sequences dataset using 12 version

Above block of code used to extract sequences, coordinates and angles from CASP sequences batch dataset. All extracted sequences can be input to DL algorithms to train a model and this trained model can be applied on test sequences to calculate prediction accuracy.

We have extracted all features and then saved those features in Dataset folder as NUMPY format. We cannot load and process above dataset in NORMAL CPU system so we need to load and extract features from Google COLAB GPU and then can save those features in NUMPY and reuse those features for training with DL algorithms in normal CPU.

Each DL algorithm performance is evaluated in terms of accuracy, precision, recall and FSCORE.

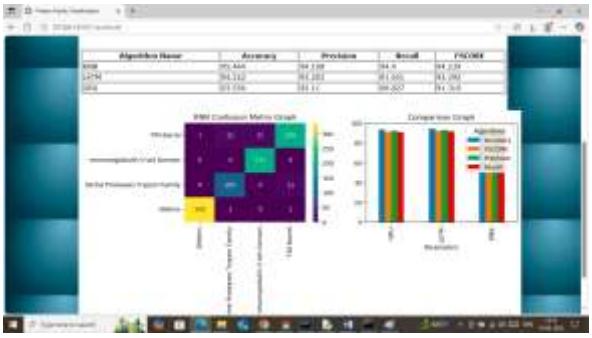
To implement this project we have designed following modules

1) Registration Here: user can sign up with the application
2) User Login: user can login to system
3) Load & Process Protein Sequences Dataset: using this model will load and normalize all dataset sequences and then split into train and test where

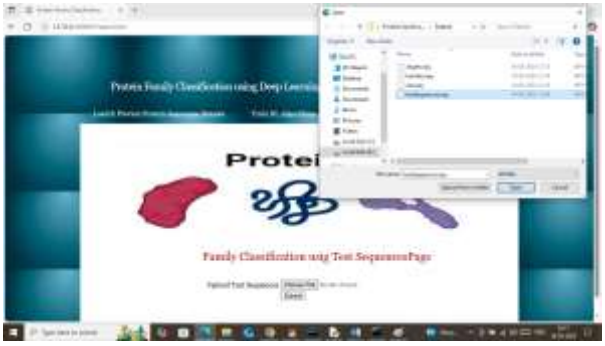
application using 80% sequences for training and 20% for testing

4) Train DL Algorithms: 80% training sequences will be input to DL algorithms to trained a model and this model will be applied on 20% test sequences to calculate prediction accuracy
5) Family Classification using Test Sequences: using this module user can upload test sequences and then best performing DL model will classify type of family.

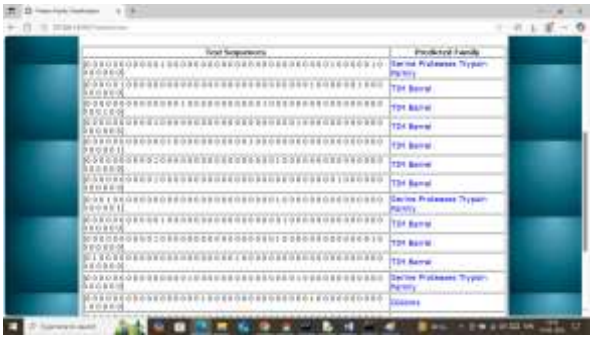
In above screen can see number of sequences loaded and processed from dataset and can see number of features available in each sequences and then can different families class labels found in dataset and then can see train and test size. Now click on 'Train DL Algorithms' link to train DL algorithms and then will get below page.



In above screen in table format can see LSTM, RNN and GRU algorithms performance where RNN got 95% accuracy and LSTM got 94% and GRU got 93% accuracy and can see other metrics like precision, recall and FSCORE. In confusion matrix graph x-axis represents 'Predicted Family Labels' and y-axis represents True labels and then different colour boxes in diagonal represents correct prediction count and remaining blue boxes represents incorrect prediction count which are very few. In second graph showing comparison graph between all algorithms where x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. Now click on 'Family Classification using Test Sequences' link to get below page



In above screen selecting and uploading test sequences file and then click on buttons to get below page



In above screen in first column showing few sequences from test data and in second column displaying classified family name

V. CONCLUSION

The proposed system for protein family classification using deep learning techniques provides an efficient and accurate solution for analyzing biological sequences. By leveraging advanced models such as RNN, LSTM, and GRU, the system effectively captures sequential patterns within protein data, leading to high classification accuracy. Among the implemented models, RNN achieved the highest accuracy, followed by LSTM and GRU, demonstrating the effectiveness of deep learning in bioinformatics applications. The use of GPU-based preprocessing and feature extraction significantly improved computational efficiency and enabled handling of large-scale datasets like SideChainNet. Additionally, the system's modular design and user-friendly interface allow easy dataset processing, model training, and classification of new sequences. Performance evaluation using metrics such as accuracy, precision, recall, and F1-score confirms the reliability of the approach. Overall, this project highlights the potential of deep learning in protein classification, contributing to advancements in computational biology, drug discovery, and understanding of protein functions.

REFERENCES

[1] J. Altschul *et al.*, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[2] S. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder," in *Proc. EMNLP*, 2014.

[6] A. Graves, "Supervised sequence labelling with recurrent neural networks," Springer, 2012.

[7] D. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *Journal of Molecular Biology*, 1999.

[8] M. Remmert *et al.*, "HHblits: Lightning-fast iterative protein sequence searching," *Nature Methods*, 2012.

[9] R. AlQuraishi, "End-to-end differentiable learning of protein structure," *Cell Systems*, 2019.