PROTEIN FOLD PREDICTION

A PROJECT REPORT II (IP1302)

Submitted By

Jerome B (CS22B1019)

N Kathiravan (CS22B1036)

Guided By

Dr. Sanjay S. Bankapur

Assistant Professor
Department of Computer Science and Engineering
National Institute of Technology Puducherry
Karaikal – 609609

DEPARTMENT OF
COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY PUDUCHERRY
KARAIKAL – 609 609
NOVEMBER 2024

BONAFIDE CERTIFICATE

Certified that this Project II Report (IP1302) “PROTEIN FOLD PREDICTION” is the Bonafide work of “JEROME B (CS22B1019), N KATHIRAVAN (CS22B1036)” who carried out the project work under my supervision at National Institute of Technology Puducherry during the period from July -2024 to November-2024.

Dr. Sanjay S. Bankapur	**Dr.Vani V**	**Dr. Ansuman Mahapatra**
Supervisor	Project Coordinator	Head of the Department
Assistant Professor	Assistant Professor	Assistant Professor
Department of CSE	Department of CSE	Department of CSE
NITPY, Karaikal	NITPY, Karaikal	NITPY, Karaikal

Abstract

Protein fold prediction plays a crucial role in understanding protein structure and functions, with significant implications for biological research and drug discovery. This study explores the use of advanced machine learning models for classifying protein structures based on sequence data. Protein sequences are first encoded into embeddings using the pretrained Prot-BERT model, capturing the biochemical characteristics of each protein. These embeddings fed into various supervised models, few are XGBoost, SVM, MLP and selected model was hybridized with Quantum, which is designed to enhance classification accuracy by incorporating quantum computing techniques. The performance of each model is evaluated across multiple datasets using metrics such as accuracy, precision, recall, and F1-score. Experimental results show that XGBoost and Quantum Hybrid MLP outperform simpler models like Decision Trees, with the Quantum Hybrid MLP achieving notable improvements in accuracy on challenging datasets. Future work aims to further improve prediction performance by exploring deep learning architectures like RNNs, CNNs and Transformers, as well as expanding the use of quantum-enhanced models for protein fold prediction. These advancements could significantly impact the understanding of protein structures, aiding in areas such as drug discovery.

TABLE OF CONTENTS

List of Tables

List of Figures

Abbreviations

1. **SVM** – Support Vector Machine
2. **XG-Boost** – Extreme Gradient Boosting
3. **KNN** – K-Nearest Neighbors
4. **RF** – Random Forest
5. **LR** – Logistic Regression
6. **MLP** – Multi Layer Perceptron
7. **CASP -** Critical Assessment of Protein Structure Prediction
8. **BERT -** Bidirectional Encoder Representations from Transformers
9. **PSSP -** Protein Secondary Structure Prediction
10. **PDB -** Protein Data Bank

1.0 INTRODUCTION

Proteins are essential molecules in biology, made up of chains of amino acids that fold into specific shapes, allowing them to perform their functions. This folding process is crucial because a protein’s structure determines its role, whether it’s building cellular structures, carrying out chemical reactions, or signaling between cells.

Protein folding occurs in stages, starting with the primary structure, which is simply the linear sequence of amino acids. This sequence naturally begins to form local shapes, known as the secondary structure, like coils (alpha helices) and flat sheets (beta sheets). As folding continues, these secondary elements interact to form a more complex three-dimensional tertiary structure, which is the functional shape for many proteins. In some cases, multiple protein chains join together, creating a quaternary structure.

In this project, we aim to predict two key aspects of protein structure: the overall fold (PFP) and the local shapes (PSSP) within the protein. To achieve this, we used supervised machine learning (ML) models trained on known protein structures to learn the patterns between sequences and their resulting shapes. Additionally, we explored quantum computing by comparing a classical multi-layer perceptron (MLP) model with a hybrid quantum MLP. This comparison helps us assess whether quantum approaches might offer benefits over traditional ML in complex tasks like protein structure prediction.

Protein Secondary Structure Prediction (PSSP) and protein fold prediction are crucial for understanding protein 3D structure, which determines function. PSSP identifies structural motifs like alpha-helices, while fold prediction classifies proteins into structural categories. Together, they support drug discovery, disease research, and functional annotation.

1.1 MOTIVATION

Understanding and predicting protein structures is a key challenge in biology, with significant implications in drug design, disease research, and biotechnology. Proteins play a critical role in nearly all cellular processes, and their specific shapes determine their function. Misfolding of proteins is linked to several diseases, such as Alzheimer's, Parkinson's, and certain cancers, making accurate prediction of protein structures essential for developing new therapies and understanding disease mechanisms.

Despite advances, predicting protein structures based solely on amino acid sequences remains a complex problem, especially when tackling large proteins with intricate folding patterns. Traditional methods like X-ray crystallography and cryo-electron microscopy, while accurate, are time-consuming and costly. This motivates the development of computational approaches like protein fold prediction (PFP) and protein secondary structure prediction (PSSP), which can provide faster insights into protein structure using machine learning.

In this project, we explore supervised ML techniques to predict these structures, leveraging large datasets of known protein shapes. Additionally, with the recent growth of quantum computing, we aim to assess whether quantum-enhanced models could bring new capabilities to protein structure prediction. By comparing classical and hybrid quantum approaches, we hope to contribute to a new generation of computational tools that can more effectively address the challenges of protein structure prediction, paving the way for advances in biomedical research and therapeutic development.

2.0 LITERATURE SURVEY

Table 2.1. A Literature review on Protein fold prediction

S.No	Title	Author	Methodology	Outcome
1	Self-attention and asymmetric multi-layer perceptron-gated recurrent unit blocks for protein secondary structure prediction	Dewi Pramudi Ismi, Reza Pulungan, Afiahayati	The paper introduces the SADGRU-SS model, which combines self-attention mechanisms with asymmetric multi-layer perceptron (MLP) and gated recurrent unit (GRU) blocks to improve prediction accuracy	This model gives 82.78 percent accuracy in predicting Q3 and gives 70.74 percent accuracy in predicting Q8 states

2	Accurate structure prediction of biomolecular interactions with AlphaFold 3	Josh Abramson, Jonas Adler, Jack Dunger, Richard Evansnger, 43 more	AlphaFold 3 uses a diffusion-based architecture capable of predicting the joint structure of complexes, including proteins, nucleic acids, small molecules, ions, and modified residues	Achieved higher precision in atomic-level accuracy across various protein families
3	Highly accurate protein structure prediction with AlphaFold	John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov,25 more	Deep Learning concepts like CNN and RNN are used. Stages like input sequence, databases searches, MSA generation, Representation generation, Evoformer, Structure Module, Recycling and Output are involved	Improved prediction accuracy to near-experimental levels
4	ColabFold: making protein folding accessible to all	Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, Martin Steinegger	MMseqs2 for Homology Search, AlphaFold2 and RoseTTAFold for Structure Prediction	Enabled broad access to high-accuracy protein structure prediction for users without advanced hardware, democratizing access to structural biology tools

5	Protein folds vs. protein folding Differing questions, different challenges	Shi-Jie Chen, Mubashir Hassan, Robert L. Jernigan,21 more	Differentiates the words Protein fold and folding	Folding is about the process and dynamics, Fold refers to the specific, stable three dimensional structure
6	Accurate prediction of protein structures and interactions using a three-track neural network	Minkyung Baek, Frank DiMaio, Ivan Anishchenko and many more	RoseTTAFold employs a three-track network that simultaneously processes and integrates information from protein sequences, distance matrices, and coordinate frames. This multi-track approach allows the model to capture complex relationships and dependencies within protein structures	The model achieves high accuracy in predicting protein structures, rivaling experimental methods such as X-ray crystallography and cryo-electron microscopy
7	An enhanced protein secondary structure prediction using deep learning framework on hybrid profile based features	Prince Kumar, Sanjay Bankapur, Nagamma Patil	The paper proposes using a deep learning model to enhance the prediction accuracy of protein secondary structures, which are essential for understanding protein functions and interactions	This outperforms traditional prediction methods by leveraging the strengths of deep learning in processing complex data and making predictions with higher precision

8	An Enhanced Protein Fold Recognition for Low Similarity Datasets Using Convolutional and Skip-Gram Features With Deep Neural Network [3]	Sanjay Bankapur, Nagamma Patil	The paper employs a hybrid model of Convolutional Neural Network and skip gram model	The model is evaluated on three popular and publicly available benchmark datasets such as DD, EDD, and TG and obtained 85.9%, 95.8%, and 88.8% fold accuracies, respectively

3.0 METHODOLGY

The Protein Fold Prediction model uses protein sequences as input to predict the specific structural folds of proteins. The process begins by encoding these sequences into meaningful embeddings using the Prot_BERT model, which captures the biochemical characteristics of each protein. These embeddings are then used as inputs for various supervised machine learning models, which aim to classify proteins into distinct fold categories. Additionally, a Quantum Hybrid Model is introduced to further enhance classification accuracy by incorporating quantum computing techniques. This hybrid approach seeks to leverage quantum capabilities to handle complex features in the protein data, potentially yielding better predictive power. The performance of each model is evaluated based on metrics such as accuracy, precision, recall, and F1-score, providing insights into their effectiveness for protein fold prediction.

For Protein Secondary Structure Prediction (PSSP), the model focuses on predicting secondary structure elements such as alpha-helices and beta-sheets using data from the CASP dataset. This dataset undergoes a series of preprocessing steps, including cleaning and formatting, to ensure the data is suitable for machine learning. Key features are then selected from the preprocessed data to improve model efficiency and relevance, reducing dimensionality while retaining important structural information. These selected features are input into supervised machine learning models designed to predict secondary structure classes. Finally, the performance of each model is analyzed using a similar set of metrics, allowing for a comparative assessment of their predictive capabilities.

3.1 PRE-PROCESSING

In the data preprocessing for protein fold prediction, Prot_BERT plays a crucial role in transforming protein sequences into high-dimensional embeddings that encode meaningful biochemical and structural information. Prot_BERT is a transformer-based language model specifically trained on large amounts of protein sequence data, inspired by the BERT (Bidirectional Encoder Representations from Transformers) model used in natural language processing. Just as BERT learns contextual relationships between words in a sentence, Prot_BERT learns relationships between amino acids within a protein sequence, capturing the complex dependencies and properties inherent in protein structures.

When protein sequences are passed through Prot_BERT, it generates a numerical representation, or embedding, for each sequence. These embeddings are essentially feature-rich vectors where each element reflects specific biochemical and structural characteristics of the protein. For example, the embeddings may capture information about amino acid composition, secondary structural motifs, and evolutionary conservation, which are key for accurately predicting the 3D fold of the protein.

This embedding process replaces manual feature engineering, which can be challenging and time-consuming for protein data. Prot_BERT's embeddings thus provide a standardized, comprehensive feature set that allows machine learning models to learn from detailed protein characteristics. Once generated, these embeddings are often normalized to a consistent scale, which enhances the training stability and performance of downstream models, allowing them to more effectively classify proteins into fold categories based on the encoded structural and functional patterns learned by Prot_BERT.

Overall, Prot_BERT streamlines data preprocessing by automating feature extraction and delivering powerful, high-dimensional representations tailored for protein analysis.

3.2 PROPOSED ARCHITECTURE

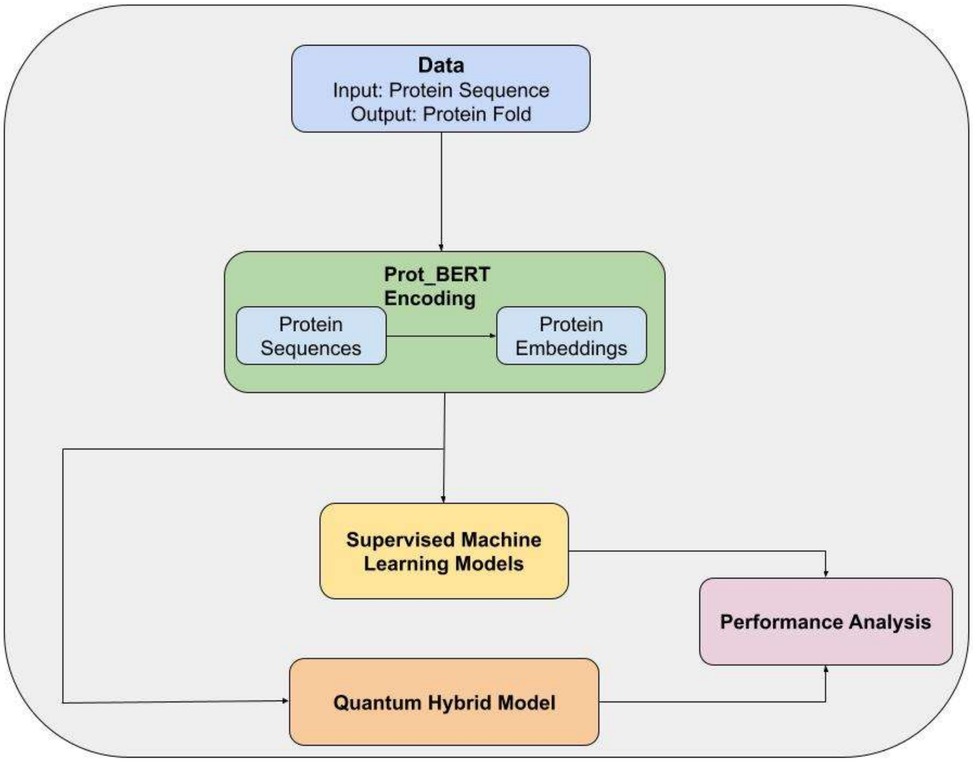

Fig. 3.2.1 Architecture Diagram of Protein fold prediction model

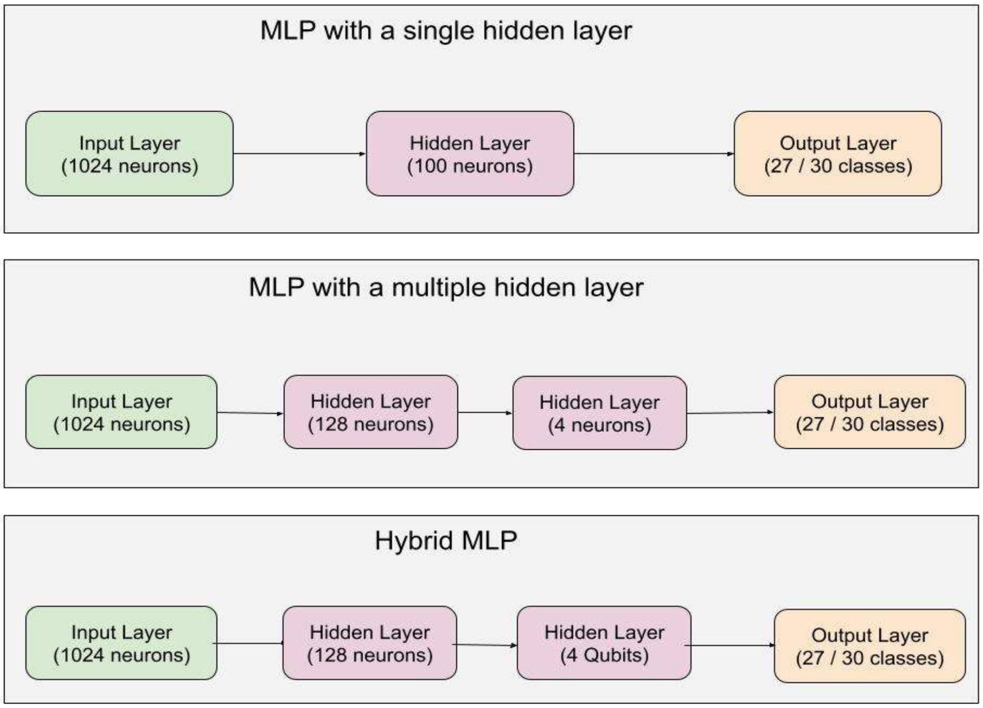
MLP with a single hidden layer
Input Layer (1024 neurons)
Hidden Layer (100 neurons)
Output Layer (27 / 30 classes)
MLP with a multiple hidden layer
Input Layer (1024 neurons)
Hidden Layer (128 neurons)
Hidden Layer (4 neurons)
Output Layer (27 / 30 classes)
Hybrid MLP
Input Layer (1024 neurons)
Hidden Layer (128 neurons)
Hidden Layer (4 Qubits)
Output Layer (27 / 30 classes)

Fig. 3.2.2 Architecture of MLP and Hybrid MLP

Single-Layer MLP: This model consists of an input layer with 1024 neurons, followed by a single hidden layer with 100 neurons, and an output layer that performs the final classification. This simple structure is intended to capture basic patterns in the data.

Multiple-Layer MLP: This architecture includes two hidden layers, with the first layer containing 128 neurons and the second layer containing 4 neurons. By adding an additional layer, the model gains the capacity to learn more complex features, which can potentially enhance classification accuracy.

Hybrid MLP: In this configuration, the second hidden layer comprises 4 quantum qubits, creating a hybrid model that integrates quantum computing with classical neural networks. The hybrid approach leverages quantum properties to capture intricate patterns, with the aim of improving model performance for complex classification tasks.

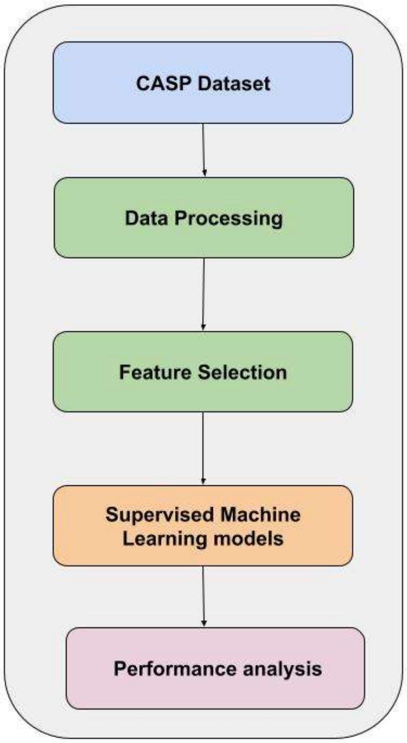

Fig. 3.2.3 Architecture of Protein Secondary Structure Prediction model

4.0 PERFORMANCE ANALYSIS

4.1 DATASET

4.1.1 Protein Fold Prediction

The DD, EDD, and TG datasets are commonly used for protein fold prediction. The DD dataset, constructed from SCOP 1.63, consists of 311 protein sequences in the training set and 383 in the test set, representing 27 different folds with less than 40% similarity between sequences. The EDD dataset, derived from SCOP 1.75, is an extended version of DD and includes 3,418 protein sequences across the same 27 folds, exhibiting less than 40% similarity. The TG dataset, sourced from SCOP 1.73, contains 1,612 protein sequences across 30 different folds with less than 25% similarity. These datasets are essential for training and evaluating models focused on protein fold recognition and structure prediction.

4.1.2 Protein Secondary Structure Prediction

The CASP (Critical Assessment of Structure Prediction) datasets are widely used in the field of protein structure prediction. They provide experimentally determined structures of proteins, which serve as the ground truth for evaluating prediction models. These datasets include the features like PDB(Protein Data Bank) id , amino acid, chain, Q3 label, Q8 label, Accessible Surface Area, Relative Solvent Accessibility, Phi (φ) and Psi (ψ) Angles.

CASP10: Focuses on proteins from the 10th round of the CASP competition. It includes a variety of protein folds and structures, with a mixture of soluble and membrane proteins. This dataset has 9525 records.

CASP11: Includes data from the 11th CASP competition, featuring more complex folds and longer sequences. It also provides a good mix of both well-characterized and less understood protein structures. This dataset has 16232 records.

CASP12: A more recent dataset from the 12th CASP competition, showcasing advancements in structure prediction, including some proteins with challenging and novel folds. The dataset is considered a key benchmark in evaluating prediction models. This dataset has 19778 records.

4.2 EXPERIMENTAL SETUP

- **Environment:** NVidia A100 AI Server
- **Programming Language**: Python.
- **Libraries:** Pytorch**,** Pandas, Sklearn, Numpy, Pennylane-Qiskit.
- **Models:** Random forest, Decision tree, K- Nearest Neighbour, Multi-Layer Perceptron, Naive Bayes, Support Vector Machine, Logistic Regression, Extreme Gradient Boosting.

4.3 EVALUATION METRICS

In this study, we employed several classifiers to evaluate the performance on three datasets: DD, EDD, and TG. The models were assessed using the following key metrics:

1. **Accuracy**: This metric measures the overall correctness of the model, representing the proportion of true results (both true positives and true negatives) among the total number of cases evaluated.

   Accuracy = (True Positives + True Negatives) / Total Instances

2. **Precision**: Precision quantifies the number of correct positive predictions made out of all positive predictions. A higher precision indicates fewer false positives.

   Precision = True Positives / (True Positives + False Positives)

3. **Recall (Sensitivity)**: This metric evaluates the model's ability to correctly identify positive instances. It is the ratio of true positives to the sum of true positives and false negatives.

   Recall = True Positives / (True Positives + False Negatives)

4. **F1-Score**: The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both concerns, particularly useful when dealing with imbalanced datasets.

   F1-Score = 2*Precision*Recall / (Precision + Recall)

4.4 RESULTS & ANALYSIS

Table 4.4.1: Protein Fold Prediction on DD Dataset

Classifiers	Accuracy	Precision	Recall	F1-score
Decision Tree	35%	40%	35%	35%
Naïve Bayes	54%	61%	54%	54%
KNN	50%	56%	50%	47%
SVM	70%	74%	70%	69%
Logistic Regression	61%	68%	62%	61%
Random Forest	63%	71%	63%	61%
MLP	77%	78%	77%	76%
XG-Boost	**90%**	**95%**	**85%**	**90%**

Table 4.4.2: Protein Fold Prediction on EDD Dataset

Classifiers	Accuracy	Precision	Recall	F1-score
Decision Tree	40%	32%	35%	33%
Naïve Bayes	52%	53%	58%	51%
KNN	61%	60%	57%	56%
SVM	86%	89%	84%	85%
Logistic Regression	79%	81%	71%	74%
Random Forest	66%	73%	54%	57%
MLP	**89%**	**90%**	**89%**	**89%**
XG-Boost	76%	77%	72%	72%

Table 4.4.3: Protein Fold Prediction on TG Dataset

Classifiers	Accuracy	Precision	Recall	F1-score
Decision Tree	66%	54%	70%	56%
Naïve Bayes	66%	54%	70%	56%
KNN	79%	72%	68%	70%
SVM	86%	76%	59%	63%
Logistic Regression	86%	81%	60%	66%
Random Forest	84%	76%	53%	58%
MLP	**87%**	**74%**	**71%**	**73%**
XG-Boost	86%	81%	61%	67%

Table 4.4.4 Comparison Between Classical MLP and Quantum MLP

Models	**Classical MLP**				**Quantum Hybrid MLP**			
Datasets	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
DD	56.12%	59.85%	56.12%	53.64%	70.50%	73.80%	70.50%	70.21%
EDD	82.89%	85.34%	82.89%	83.51%	87.13%	88.86%	87.13%	87.33%
TG	58.82%	60.46%	58.82%	58.40%	78.02%	79.34%	78.02%	77.26%

The evaluation across the DD, EDD, and TG datasets reveals that models like XG-Boost, SVM, and MLP achieved superior performance, with XG-Boost standing out, especially on the DD dataset with 90% accuracy and F1-score. In contrast, simpler models such as Decision Tree showed weaker performance. Additionally, a comparison between Classical MLP and Quantum Hybrid MLP highlights that the Quantum Hybrid model consistently outperformed the Classical approach across all datasets, achieving a significant accuracy boost in the DD dataset (70.5% vs. 56.12%). Notably, the Quantum Hybrid MLP also demonstrated higher precision and recall on the EDD dataset, underscoring the potential of quantum techniques to enhance classification performance in complex data scenarios

Table 4.4.5 Protein Secondary Structure Prediction on CASP 12

	Q3 (3 classes)				**Q8 (8 classes)**			
Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Decision Tree	84%	83%	84%	84%	77%	60%	60%	60%
Naïve Bayes	76%	74%	76%	73%	64%	34%	33%	31%
KNN	88%	88%	88%	88%	81%	63%	61%	62%
SVM	74%	72%	71%	71%	64%	28%	30%	28%
Logistic Regression	76%	74%	73%	73%	65%	35%	33%	32%
Random Forest	**90%**	**90%**	**90%**	**90%**	**84%**	**82%**	**66%**	**71%**
MLP	82%	81%	80%	80%	71%	45%	41%	41%
XG-Boost	89%	88%	88%	88%	81%	79%	61%	66%

Table 4.4.6 Protein Secondsary Structure Prediction on CASP 11

	Q3 (3 classes)				Q8 (8 classes)			
Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Decision Tree	83%	83%	83%	83%	72%	58%	58%	58%
Naïve Bayes	72%	73%	75%	71%	62%	34%	34%	32%
KNN	85%	85%	86%	86%	75%	70%	58%	60%
SVM	72%	72%	74%	72%	59%	28%	3%	29%
Logistic Regression	74%	74%	75%	74%	65%	35%	33%	32%
Random Forest	**89%**	**89%**	**89%**	**89%**	**81%**	**79%**	**65%**	**69%**
MLP	83%	83%	83%	83%	71%	58%	46%	47%
XG-Boost	86%	86%	87%	87%	77%	71%	60%	63%

Table 4.4.7 Protein Secondary Structure Prediction on CASP 10

	Q3 (3 classes)				Q8 (8 classes)			
Classifiers	**Accuracy**	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Precision**	**Recall**	**F1-score**
Decision Tree	76%	76%	77%	77%	62%	45%	45%	45%
Naïve Bayes	66%	68%	72%	67%	54%	29%	30%	28%
KNN	82%	82%	82%	82%	70%	59%	57%	58%
SVM	67%	67%	70%	68%	55%	29%	30%	27%
Logistic Regression	70%	70%	72%	70%	56%	29%	31%	29%
Random Forest	**83%**	**83%**	**83%**	**83%**	**74%**	**62%**	**53%**	**56%**
MLP	78%	78%	79%	78%	63%	50%	39%	38%
XG-Boost	82%	82%	82%	82%	70%	68%	52%	55%

The results compare the performance of different classifiers on the CASP 11 and CASP 12 datasets for both Q3 (3-class) and Q8 (8-class) tasks. Across both datasets and tasks, Random Forest and XGBoost show consistently high accuracy, precision, recall, and F1-scores, particularly in the Q3 classification. MLP also performs well, but slightly lower than Random Forest and XGBoost, especially in Q8. Decision Tree, Naïve Bayes, and Logistic Regression generally perform the worst, with lower scores in both Q3 and Q8. The classifiers generally achieve higher scores in the Q3 task compared to Q8 due to the reduced complexity of fewer classes.

5.0 CONCLUSION

The evaluation across various datasets highlights the strong performance of advanced models like XGBoost, SVM, and MLP in protein structure classification tasks, with XGBoost achieving particularly high accuracy on the DD dataset (90%). Simpler models, such as Decision Trees, underperformed, emphasizing the need for more sophisticated approaches to tackle the complexity of these datasets. Notably, the Quantum Hybrid MLP outperformed the Classical MLP across all datasets, achieving a significant accuracy improvement on the DD dataset (70.5% vs. 56.12%), suggesting that quantum computing techniques can enhance classification performance in high-dimensional data scenarios. Additionally, Random Forest and XGBoost were top performers in the CASP 11 and CASP 12 datasets, excelling in simpler Q3 tasks, while classifiers like Naïve Bayes and Logistic Regression showed relatively weaker performance.

Looking ahead, future work should explore more advanced deep learning models, such as Convolutional Neural Networks (CNNs) and Transformer-based architectures, to further improve classification accuracy. The continued integration of quantum-enhanced models, such as Quantum Hybrid architectures, could unlock greater potential in handling complex protein data. Furthermore, expanding research into protein folding prediction models, inspired by advancements like AlphaFold, could provide deeper insights into protein structures and offer valuable contributions to biological research and practical applications in areas such as drug discovery.

REFERENCES

1. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 1-3.

2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... & Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557), 871-876.

3. Bankapur, S., & Patil, N. (2020). An enhanced protein fold recognition for low similarity datasets using convolutional and skip-gram features with deep neural network. IEEE Transactions on NanoBioscience, 20(1), 42-49.

4. Chen, S. J., Hassan, M., Jernigan, R. L., Jia, K., Kihara, D., Kloczkowski, A., ... & Rose, G. D. (2023). Protein folds vs. protein folding: Differing questions, different challenges. Proceedings of the National Academy of Sciences, 120(1), e2214423119.

5. Ismi, D. P., & Pulungan, R. (2024). Self-attention and asymmetric multi-layer perceptron-gated recurrent unit blocks for protein secondary structure prediction. Applied Soft Computing, 159, 111604.

6. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. nature, 596(7873), 583-589.

7. Kumar, P., Bankapur, S., & Patil, N. (2020). An enhanced protein secondary structure prediction using deep learning framework on hybrid profile based features. Applied Soft Computing, 86, 105926.

8. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nature methods, 19(6), 679-682.