IJBPAS, December, Special Issue, 2023, 12(12): 598-612
ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)
'A Bridge Between Laboratory and Reader'

www.ijbpas.com

EMPIRICAL ANALYSIS FOR CLASSIFICATION AND PREDICTION OF PROTEIN FAMILY USING MACHINE LEARNING

RASHMI TS[1], VEENA M R[2], KAMARAJ R[3] AND JYOTHI NM[4*]

1: Department of Biotechnology, Government Science College, Chitradurga, Karnataka, India

2: Department of Biotechnology, Government Arts and Science College, Karwar, Karnataka, India

3: Department of Biotechnology, School of Arts and Science, PRIST Deemed to be University, Thanjavur, Tamil Nadu, India

4: Dept. of Computer Science Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswarm, AP, India

[*]Corresponding Author: Dr. Jyothi N M

Received 15th July 2023; Revised 19th Aug. 2023; Accepted 1st Dec. 2023; Available online 15th Dec. 2023

https://doi.org/10.31032/IJBPAS/2023/12.12.1073

ABSTRACT

Proteins are fundamental to life, and understanding their structures is crucial for deciphering their functions. Despite the efforts that have unveiled around 100,000 unique protein structures, this represents a small fraction of the vast protein sequence space. The laborious and time-consuming process of determining a protein's structure has been a bottleneck. To bridge this gap and enable large-scale structural bioinformatics, computational methods are essential. The challenge of predicting a protein's three-dimensional structure from its amino acid sequence, known as the 'protein folding problem,' has persisted for over five decades. Existing methods have limitations, especially when there are no structurally similar proteins as references. Recently, a groundbreaking machine learning approach was introduced, capable of consistently predicting protein structures with atomic accuracy, even in cases with no structural homologs. This approach leverages both physical and biological knowledge about protein structure and incorporates multiple sequence alignments into the machine learning algorithm's design. This research focuses on empirical analysis of protein structure and classification and prediction of protein family using machine learning algorithm and attained high accuracy of 95%.

Keyword: Protein family, DNA, protein sequence, K-Nearest Neighbors

I. INTRODUCTION

Proteins are indispensable for the sustenance of life as they fulfill a wide array of vital functions within organisms. They play pivotal roles in processes like DNA replication, catalyzing essential metabolic reactions, facilitating molecule transportation between cells, and contributing to reproduction, among others. A protein's sequence is constructed from twenty distinct amino acids arranged in a specific order, a critical feature that holds significant importance in gene encoding and the efficient functioning of proteins. Even a single gene mutation can lead to the incorporation of an incorrect amino acid into the sequence.

What sets proteins apart is their remarkable versatility, all based on a common foundation – the use of twenty amino acids to construct diverse proteins. This inherent diversity has spurred extensive research into understanding proteins, encompassing aspects like their structure, function, and composition. This focused exploration of proteins has been driven by the quest to answer fundamental questions, such as the origins of diseases like cancer, the mechanisms behind aging, the development of pharmaceutical interventions for various ailments, and the evolution of life on Earth. These inquiries find their answers through investigations into protein folding, functionality, complex formation, and related aspects of these remarkable biomolecules.

The composition of a protein sequence relies on a set of twenty different amino acids, and the arrangement of these amino acids, as well as the specific types utilized, fundamentally determine the structure and function of each protein. The connection between a protein's sequence and its functional attributes is a longstanding focal point in the field of molecular biology, given its profound implications. Traditionally, the determination of a protein's function has involved time-consuming techniques like crystallography for structural studies or biochemical studies. In recent years, computational methods have emerged as a valuable tool for predicting protein function. A more general and accessible approach to predicting protein function is based on the concept of inheritance. This concept implies that protein sharing sequence similarity are likely to share similar functions, leading to the formation of protein families that group together proteins with common functionalities.

Classifying a protein sequence into its respective family can provide a deeper understanding of the protein's structure, function, and metabolic activities. This approach proves particularly useful for identifying and characterizing unknown or challenging-to-study proteins that may not be easily characterized using pairwise alignments. Typically, proteins are categorized based on their structural or

sequence similarities, often relying on well-characterized proteins with known functions. When a novel protein is encountered, its functional properties can be inferred by considering the group to which it belongs. Furthermore, it's observed that proteins tend to alter their structure or sequence while still maintaining their functions. Protein sequence classification offers an effective means to extract valuable biological insights from vast datasets, which has become increasingly relevant with the substantial growth in biological data generation, especially related to proteins. Therefore, there is a pressing need for the development of new methods employing advanced tools and techniques for the classification of proteins.

The objectives of the research are preprocessing of the protein data, feature visualization, empirical analysis for deep understanding of the features, classification and prediction of protein family using Machine Learning (ML) models and comparison of the results. The rest of the paper is organized into data preprocessing, literature survey, experimentation, results, and discussion finally conclusion.

II. LITERATURE SURVEY

Recently introduced deep learning models like DeepFam and ProtCNN have been designed for the classification of proteins into their respective families. However, these models predominantly focus on non-hierarchical classification. In this study, we introduce a novel deep learning neural network named DeepHiFam, specifically devised for hierarchical protein classification, achieving high accuracy across different levels simultaneously **[1]**. Utilizing neural network-based amino acid representation learning, we investigate the performance of annotated protein families, considering limited characterized proteins and incorporating amino acid location information. Furthermore, we extend the application of this method to assess our approach on all reviewed protein sequences and their respective families obtained from the UniProt database **[2]**. The PCNM model employs a combination of a Convolutional Neural Network (CNN) module and a Global and Local feature (GL) module for feature extraction at three distinct levels: conserved domains, amino acids, and protein sequences. Demonstrating robust performance, PCNM attains an accuracy of 85.12% on the GPCR sub-subfamily dataset and 81.42% on the POG dataset **[3]**. Conventional methods, which involve comparing an unseen sequence with all identified protein sequences and determining the category index with the highest similarity, are often time-consuming. Consequently, there is an urgent need to develop an efficient protein sequence classification system. This study investigates the effectiveness of using Single Hidden Layer Feedforward Networks (SLFNs) for protein sequence classification **[4]**. The

DeepPPF framework demonstrates a notable capability in uncovering intricate motifs for functional classes using minimal sets of protein sequences. Experimental findings underscore the significance of rich motif discovery as a crucial factor in enhancing the modeling performance of protein families through the application of deep learning techniques **[5]**. Identifying membership in a known family is a fundamental process in various bioinformatics analyses, encompassing tasks such as protein structure and function prediction, as well as metagenomic taxon identification and abundance profiling. This step is crucial when dealing with new biological sequences and has widespread applications in the field **[6]**. In the present era, there is a continuous growth in the quantity of protein sequences archived in global protein databases, contributed by laboratories worldwide. However, only a subset of these proteins undergoes experimental analysis for the identification of their structure and, consequently, their functional roles within the respective organisms **[7]**.

Currently, the main techniques used to determine protein 3D structure are X-ray crystallography and nuclear magnetic resonance (NMR). In X-ray crystallography the protein is crystallized and then using X-ray diffraction the structure of protein is determined.

These techniques have major drawbacks which are listed below.

1.Crystallization Challenges: X-ray crystallography relies on the ability to crystallize the protein of interest. However, not all proteins can be easily crystallized, which limits the applicability of this method.

2.Sample Size: NMR requires a relatively large amount of purified protein, which can be a limitation when working with proteins that are difficult to produce in large quantities.

3.Protein Dynamics: Both methods provide static snapshots of protein structures. They may not capture dynamic changes or conformational flexibility, which can be essential for understanding protein function.

4.Size Limitations: NMR is typically limited to proteins with smaller molecular weights due to the technical challenges associated with larger molecules. This restricts its applicability for studying larger protein complexes.

5.Resolution: The resolution of X-ray crystallography and NMR structures may not always be sufficient to resolve fine details, especially for large, complex proteins or those with multiple domains.

6.Time and Cost: Both methods can be time-consuming and costly. Protein crystallization and data collection in X-ray crystallography, for example, can be labor-intensive and may take months or even years.

7.Destructive Nature: In X-ray crystallography, the protein crystals are often subjected to intense X-ray radiation, which can damage the crystals and potentially alter the protein's structure.

8.Limited to Static Structures: Both methods provide a snapshot of a protein's structure under specific conditions. They do not capture the full range of structural variations that a protein may exhibit in different cellular environments.

III. EXPERIMENTATION

The Protein Sequence Classification methodology is a multi-step process that involves the conversion of raw data into high-level information. The following steps are executed within this methodology:

(a) Data Selection: The initial step involves the collection of the dataset from Kaggle, which consisted of a substantial 467,305 amino acid sequences. This dataset's large size made it well-suited for the project. The dataset was originally available in two separate CSV files.

(b) Data Cleaning and Pre-processing: Subsequently, the collected data underwent cleaning and pre-processing. This is a crucial step because raw data is often not immediately usable. It included tasks such as merging the two CSV files, eliminating unnecessary rows and columns, and handling missing values.

(c) Data Transformation: The data was transformed using various techniques, including TF-IDF, Keras Word Embedding, and chi-square as a feature selection method to generate an optimal feature representation. To address the challenge of imbalanced data, random under-sampling was employed.

(d) Data Mining: In this phase, different machine learning models, specifically Random Forest K- nearest neighbor (KNN), Support Vector Machine (SVM), Convolution Neural Network (CNN) were implemented on the transformed data.

(e) Evaluation: Finally, the performance of all the models was evaluated using standard metrics such as Accuracy, Precision, Recall, and F1-score. The models were compared based on their accuracy scores to assess their effectiveness.

Figure 1 illustrates the step-by-step Protein Sequence Classification methodology, which encompasses these stages for efficient data analysis and model development. **Figure 2** shows the overall architecture of the proposed research.

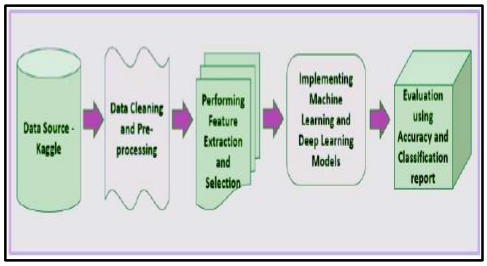

Figure 1: Classification steps of protein structure

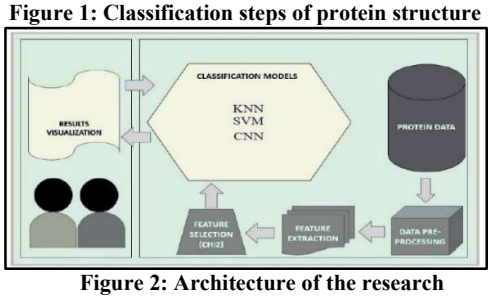

Figure 2: Architecture of the research

The dataset is obtained from Kaggle which consists of 14,0911 unique values. The original source of the dataset is http://www.rcsb.org/pdb/. It consists of 15 features. Preparing and refining data is crucial when delving into the intricacies of protein structures. The possession of a protein's structural blueprint offers profound insights into its functionalities, paving the way for the formulation of hypotheses on manipulation, control, or modification strategies. An illustrative instance lies in the ability to engineer site-specific mutations, strategically altering protein functions. Presently, the primary methodologies employed to unveil the 3D configuration of proteins are X-ray crystallography and nuclear magnetic resonance (NMR). In X-ray crystallography, the protein undergoes crystallization, followed by X-ray diffraction analysis to elucidate its structural arrangement.

Dataset is available in a well-structured format and thus, requires only data processing. The following preprocessing steps are applied.

i. Eliminating duplicate rows- The dataset is cleaned to obtain unique rows. **Figure 3** shows the heatmap of 14-feature in the dataset. This gives the maximum and minimum number of variations in each column of the dataset and null values are eliminated.

ii. Read and Analyze the data – The **Figure 4 and 5** show the analysis of the protein data. **Figure 5** shows the shape of the of the protein data structure. The analysis shows the parameter values of the protein data which are residue Count x, resolution, structure Molecular Weight, crystallization Temp K, density Matthews, density Percent Sol, pH Value and residue Count_y

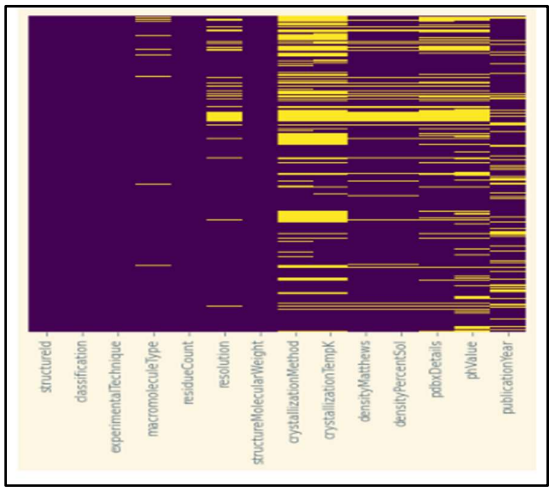

Figure 3: Heatmap of the dataset after eliminating null values and duplicate rows

	count	mean	std	min	25%	50%
residueCount_x	471811.000000	6249.411993	23602.912835	0.000000	456.000000	1140.00000
resolution	449845.000000	3.020053	3.090108	0.480000	2.000000	2.500000
structureMolecularWeight	471811.000000	924930.327041	3016951.476731	314.380000	52614.735000	130834.430
crystallizationTempK	317806.000000	290.882456	8.903673	4.000000	291.000000	293.000000
densityMatthews	390156.000000	2.850614	0.824283	0.000000	2.320000	2.670000
densityPercentSol	390278.000000	54.196381	10.269266	0.000000	46.890000	53.950000
phValue	340901.000000	6.830511	2.461170	0.000000	6.100000	7.000000
publicationYear	414031.000000	2010.458932	7.035084	201.000000	2007.000000	2012.00000
residueCount_y	471149.000000	6257.931820	23618.383810	0.000000	456.000000	1140.00000

Figure 4: Visualization of Protein data

iii. Protein structure visualization- The complete protein structure visualization with respect to residue Count_x, resolution, structure Molecular Weight, crystallization Temp K, density Matthews, density Percent Sol, pH Value and residue Count_y values are shown in **Figure 6**.

```
The shape of the protein dataset is: (471811, 17)
structureId                                    9XIM
experimentalTechnique       X-RAY DIFFRACTION, SOLUTION NMR
residueCount                                313236
resolution                                    70.0
structureMolecularWeight                97730536.0
crystallizationTempK                          398.0
densityMatthews                                99.0
densityPercentSol                              92.0
phValue                                       724.0
publicationYear                              2018.0
dtype: object
```

Figure 5: Shape of the protein data structure

iv. Protein crystallization Temp- contributes significantly to the classification of protein families by providing detailed structural information. This information enables to explore relationships between proteins, infer functional characteristics, Crystallization plays a crucial role in protein family classification. Protein crystallization is the process of forming a crystal lattice of protein molecules, allowing scientists to obtain a highly ordered three-dimensional structure of the protein. This process is essential for determining the atomic-level details of a protein's structure through techniques such as X-ray crystallography. Once the crystal structure of a protein is determined, it can be compared with structures of other proteins. This comparative analysis helps identify similarities and differences, aiding in the classification of proteins into families based on structural homology.

v. The Matthews coefficient- This itself is not a direct classifier of protein families, it provides valuable information about the packing of proteins in a crystal lattice. Differences in the Matthews coefficient among crystals of proteins from different families may reflect variations in their crystallization behavior, aiding in the overall understanding of structural biology and potentially influencing decisions related to crystallography experiments. It is even referred to as Matthews correlation coefficient or densityMatthews, is a parameter used in the analysis of X-ray crystallography data, specifically in the context of protein crystallography. It plays a role in estimating the solvent content of a crystal and, indirectly, can be relevant to protein family classification. In protein crystallography, the Matthews coefficient is also linked to considerations of crystal symmetry and pseudo-symmetry. Understanding these aspects is crucial for accurate structure determination. Different protein families may show different propensities for specific crystal symmetries, and the Matthews coefficient aids in assessing the validity of these symmetries. Comparison of Matthews coefficients across different protein crystals can give insights into the packing efficiency and overall characteristics of crystals formed by proteins from various families. This information can be valuable for assessing the general behavior of protein families in the context of crystallization.

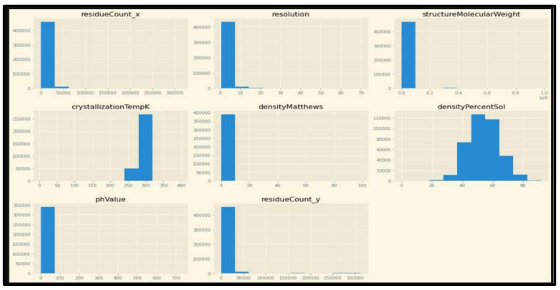

Figure 6: Visualization of protein structure

vi. Residue Count- The count of residues in a protein structure signifies the total number of amino acid residues comprising the protein chain. Each amino acid residue contributes significantly to the overall structure and functionality of the protein. This count directly impacts the protein's size and length, and larger proteins, characterized by a higher residue count, tend to possess intricate structures, potentially fulfilling diverse functions within a biological context. The number of residues plays a pivotal role in influencing the stability and folding of proteins. As proteins undergo folding to attain specific three-dimensional structures crucial for their functional states, the interactions among amino acid residues become paramount in determining the stability of these structures. Additionally, understanding the residue count aids in the identification of structural motifs and domains inherent in the protein. Moreover, the residue count holds significance in the realm of drug discovery and design. It is an essential parameter to develop targeted therapies, as the structural insights gained from analyzing residue count contribute to understanding the intricate relationship between a protein's structure and function. In essence, the analysis of residue count offers valuable perspectives on the intricate interplay between the structure and function of proteins.

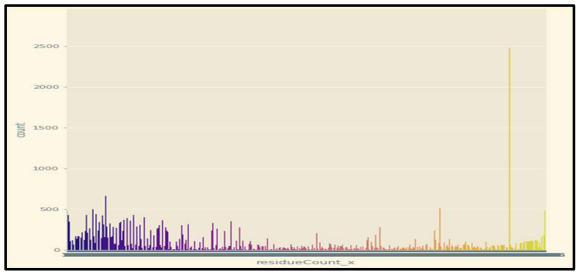

Figure 7: Residue count protein structure

vii. pH value visualization -The stability of a protein's structure is intricately tied to the surrounding pH conditions. Understanding the response of proteins within a family to variations in pH yields valuable insights into their diverse functionalities, unique structural traits, and potential roles in specific biological processes. Variations in pH can induce conformational changes in proteins, impacting their overall structure and function. Such pH-dependent structural features often serve as distinguishing factors for specific protein classes. For instance, proteins designed to function in acidic cellular compartments may showcase structural adaptations tailored to lower pH conditions. The impact of pH on protein classification is complex, given that the environmental pH plays a pivotal role in shaping the intricate interplay of structural and functional characteristics inherent to proteins.

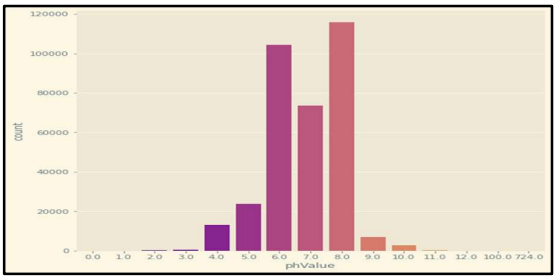

Figure 8: pH value of the protein structure

The correlation matrix of the selected protein structure features is shown in **Figure 9**. Th features residue Count_ x, residue Count_y and density Matthews how high correlation. Correlated features may provide redundant information, leading to overfitting and reduced generalization performance. By eliminating correlated features, the classification model is less likely to be influenced by noise and can better focus on the unique information relevant to distinguishing between protein families. This results in improved accuracy and robustness of the classification model. They contribute to increased dimensionality in the dataset. Removing these features reduces the number of input variables, making the classification task computationally more efficient. This is especially important when working with large datasets, as it can speed up training and testing processes. Removing correlated features can simplify the interpretation of the underlying biological characteristics that differentiate protein families. It allows researchers to identify and focus on independent features that play distinct roles in classifying proteins into different families. This elimination of correlated features can reduce the computational complexity associated with training and deploying classification models. It addresses issues related to redundancy, dimensionality, and multicollinearity, leading to more accurate and meaningful classification results.

Hence, residue Count_x, residue Count_y and density Matthews features are dropped and the resultant correlation value are shown in **Figure 10**.

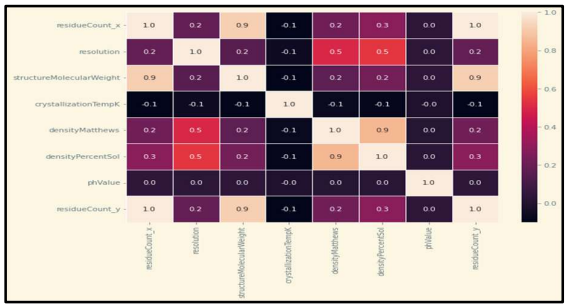

Figure 9: Correlation matrix

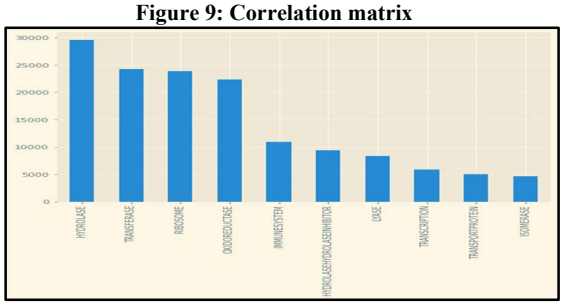

Figure 10: After removing correlated data

Protein families with greater than 15000 correlations are selected. **Figure 11** shows classification count of protein families greater than 15000. They are hydrolase, transferase, ribosome, oxidoreductase

All categorical features are converted to numerical measured in Armstrong units. **Figure 12** shows the boxplot of the resolution features of hydrolase, transferase, ribosome, oxidoreductase protein families.

viii. Protein crystallization - Top 4 most common protein crystallization methods are Microbatch, Vapor Diffusion, Vapor Diffusion Hanging Drop, Vapor Diffusion Sitting Drop. Figure 13 shows the Pie Chart for visualizing these protein crystallization methods. Protein crystallization methods contribute significantly to protein classification by providing high-resolution structural information. These methods are primarily associated with structural biology and X-ray crystallography. The obtained crystal structures offer insights into the similarities and differences within and between protein families, facilitating a more comprehensive understanding of their biological roles and relationships. Microbatch crystallization provides small-scale conditions for protein crystallization. The resulting crystals offer detailed structural information about the protein. This information is valuable for classification purposes as it helps to identify unique structural features that distinguish protein families. Whether using the hanging drop or sitting drop configuration, vapor diffusion methods are fundamental in producing high-quality protein crystals. The obtained crystal structures contribute directly to the understanding of protein architecture, aiding in the classification of proteins based on their structural similarities and differences.

```
HYDROLASE                          29559
TRANSFERASE                        24236
RIBOSOME                           23858
OXIDOREDUCTASE                     22287
IMMUNESYSTEM                       10876
HYDROLASEHYDROLASEINHIBITOR         9396
LYASE                               8354
TRANSCRIPTION                       5890
TRANSPORTPROTEIN                    5010
ISOMERASE                           4624
Name: classification, dtype: int64

HYDROLASE          29559
TRANSFERASE        24236
RIBOSOME           23858
OXIDOREDUCTASE     22287
Name: classification, dtype: int64
```

Figure 11: Classification count

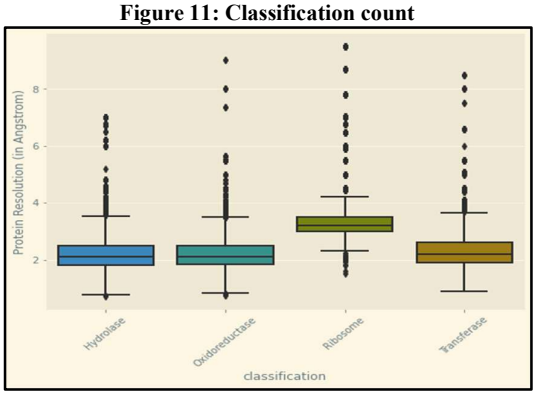

Figure 12: Boxplot of protein resolution feature

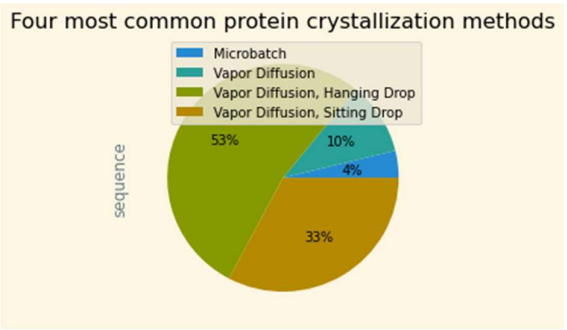

Figure 13: Four most common protein crystallization methods

ix. Molecular weight- Molecular weight plays a crucial role in the classification of protein families, offering valuable insights into a protein's size, structure, and potential functional roles. It serves as a fundamental characteristic for categorizing and organizing proteins into distinct families. Proteins belonging to the same family often exhibit

similar molecular weight ranges, forming the basis for size-based classification. An additional facet of molecular weight is its ability to shed light on a protein's oligomeric state—whether it exists as a monomer, dimer, trimer, or forms a larger complex. This information is pivotal in comprehending the structural and functional organization of proteins within a given family. Moreover, molecular weight acts as a parameter for the comparative analysis of proteins, both within a family and across different families. Variations in molecular weight can signify evolutionary divergence or convergence within related protein groups, providing valuable evolutionary insights. The identification of clusters of molecular weights within datasets of related proteins holds significance in the development of classification algorithms. These clusters contribute to the nuanced understanding of protein relationships and aid in the establishment of robust classification frameworks. The mean molecular weight of Ribosome proteins is approximately 20 times higher than that of other protein families. Notably, the Ribosome family is unique in that it lacks any outliers in the molecular weight column. Consequently, this distinct characteristic makes it more straightforward for the model to accurately classify proteins belonging to the Ribosome family.

Comparing the molecular weights of the top 4 protein families using a boxplot is shown in **Figure 14**.

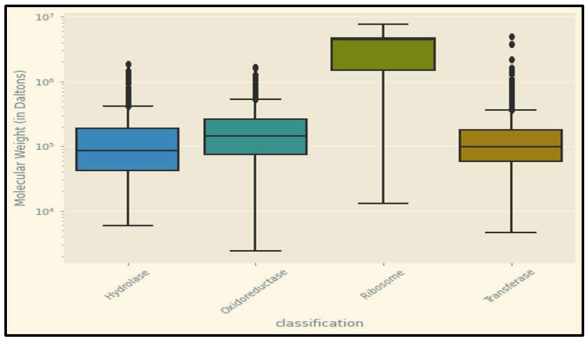

Figure 14: Boxplot of protein molecular weight

A. Protein categorization using KNN model

K-Nearest Neighbors (KNN) is a simple and intuitive machine learning algorithm used for both classification and regression tasks. In KNN, the prediction for a data point is based on the majority class or average value of its k nearest neighbors in the feature space. The term "k" refers to the number of neighbors considered. The algorithm operates on the principle that similar data points should have similar outcomes. To make a prediction for a new data point, KNN identifies the k training instances with the closest feature values and assigns the new point the most common class label (for classification) or the average value (for regression) among its neighbors. One notable characteristic of KNN is its flexibility, as it doesn't assume any underlying structure in the data. However, the choice of parameter k is crucial, as it influences the algorithm's sensitivity to local variations and noise. KNN is widely used, especially in scenarios where

the decision boundaries are complex and not easily represented by parametric models.

There are advantages of using KNN in protein family classification. KNN is a straightforward and intuitive algorithm. It doesn't assume a specific structure in the data, making it particularly appealing when the relationships between proteins in different families are complex and not easily captured by parametric models. It is non-parametric, meaning it doesn't make assumptions about the underlying distribution of the data. This flexibility is advantageous in protein family classification, where the characteristics of different families may vary widely. It does not involve a training phase, which can be beneficial when dealing with protein datasets that may be dynamic or subject to frequent updates. It allows for real-time adaptation to changes in the dataset.

Proteins within the same family may exhibit local patterns or clusters in feature space. KNN, by considering the nearest neighbors, can capture these local patterns effectively, making it suitable for scenarios where proteins within a family share similar characteristics. It does not assume linearity in the relationships between features, making it well-suited for cases where the boundaries between protein families are nonlinear or difficult to define using linear models. The decision-making process of KNN is transparent. The classification of a protein is based on the majority class among its neighbors. This transparency can be valuable for interpreting and validating classification results in the context of protein families. KNN does not assume independence between features, allowing it to handle complex relationships and dependencies among various protein features, which is crucial for accurate classification in protein family analysis.

Manual feature selection was conducted on the 'clean protein dataset' through a process involving exploratory data analysis (EDA), intuition, and domain knowledge. The objective was to reduce the number of feature inputs to five key variables: structure molecular weight, pH, crystallization method, crystallization temperature, and sequence. This selection aimed to streamline the dataset by focusing on the most relevant features based on a thorough understanding of the domain and insights gained from exploratory analysis.

IV. RESULTS AND DISCUSSION

Figure 14 shows the confusion matrix of protein family classification. The KNN model performed with high rate of True Positive. The classification report is shown in **Figure 15**. The evaluation metrics precision, recall, f1-score for each protein family is shown. Overall, 85 % accuracy is obtained.

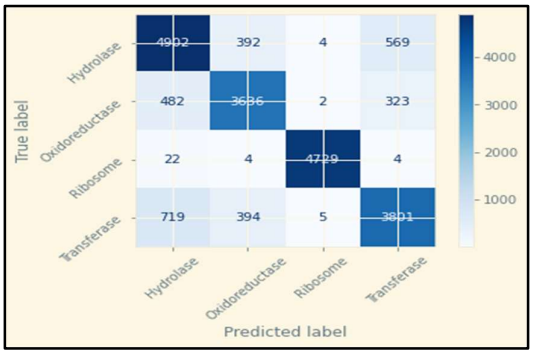

Figure 14: Confusion matrix

A. *Optimization of KNN using Grid Search CV*

Leveraging Grid Search CV, which stands for Grid Search Cross-Validation, to identify the optimal k value in K-Nearest Neighbors (KNN) entails a systematic assessment of various k values within a predefined range. The objective is to pinpoint the k value that results in the most favorable performance for a specific dataset. The optimized accuracy of classification obtained after applying GridsearchCV is shown in **Figure 16**. The accuracy of the classification reached 94.75%. as shown in **Figure 16**.

	precision	recall	f1-score	support
Hydrolase	0.80	0.84	0.82	5867
Oxidoreductase	0.82	0.82	0.82	4443
Ribosome	1.00	0.99	1.00	4759
Transferase	0.81	0.77	0.79	4919
accuracy			0.85	19988
macro avg	0.86	0.86	0.86	19988
weighted avg	0.85	0.85	0.85	19988

Figure 15: Classification Report

```
The best k parameter is : {'n_neighbors': 1}
Accuracy: 0.947518511106664


0.93741244293191
Total runtime of the knn program is 140.17503380775452 seconds
```

Figure 16: Optimized Accuracy after applying Grid search CV

In the K-Nearest Neighbors (KNN) algorithm, the selection of the k value, which represents the number of neighbors to consider, plays a crucial role in determining the model's performance. The evaluation of the model's accuracy often relies on the error rate. A prevalent strategy involves assessing the error rate across different k values and selecting the one that minimizes errors, thus identifying an optimal configuration for the model on a specific dataset. In this experiment, mean k-error rate of 0.25 is incurred and the graph is shown in the **Figure 17**.

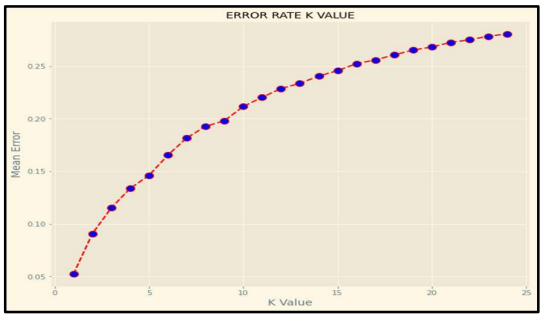

Figure 17: Mean Error rate in k value

As per the experiment, K=1 gives us the lowest mean error with 5 features. Figure 18 shows the confusion matrix after applying using GridSearchCV. The Tru Positive rate of the model prediction improved in comparison with the earlier prediction using only KNN.

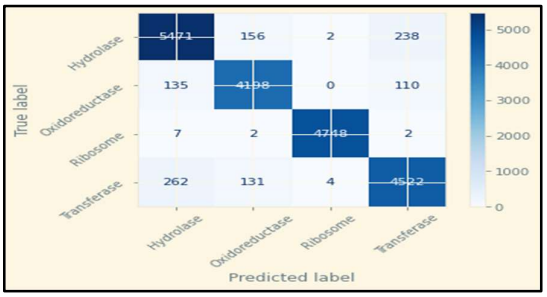

Figure 18: Confusion matrix after optimization

Figure 19 shows the classification report of KNN after applying GridsearchCV with an overall accuracy of 95% in the classification.

	precision	recall	f1-score	support
Hydrolase	0.93	0.93	0.93	5867
Oxidoreductase	0.94	0.94	0.94	4443
Ribosome	1.00	1.00	1.00	4759
Transferase	0.93	0.92	0.92	4919
accuracy			0.95	19988
macro avg	0.95	0.95	0.95	19988
weighted avg	0.95	0.95	0.95	19988

Figure 19: Classification report after optimization

V. CONCLUSION

The empirical data analysis of the protein structure helps in understanding the protein structure, analysis of correlated features. An accurate protein classification prediction system is introduced, leveraging the physiochemical attributes of proteins. In this study, a focused and practical feature subset was identified for predicting the families and functions of four distinct protein types. Through fine-tuning the hyperparameter K in the K-Nearest Neighbors (KNN) model, notable improvements were observed. The classification accuracy surged from 85% to 95% when K was set to 1, and the number of features considered was reduced to 5. The working of the KNN is optimized using GridSearchCV to obtain 95 % accuracy in classification.

Limitations

K-Nearest Neighbors (KNN) exhibits sensitivity to outliers or noise within the data, a characteristic particularly pertinent in protein family classification where datasets often possess diverse characteristics. KNN is vulnerable to irrelevant features, which is particularly problematic in protein family classification. The selection of the k-value, representing the number of neighbors, emerges as a critical consideration in optimizing KNN's performance. The quest for the optimal k-value demands meticulous evaluation and consideration of the dataset's characteristics.

Future Enhancements

Future enhancements for protein family classification could involve integrating advanced feature engineering techniques to capture more nuanced characteristics. Additionally, exploring ensemble methods or hybrid models that combine the strengths of multiple algorithms may further boost classification accuracy. Finally, incorporating deep learning approaches, such as neural networks, could unveil intricate patterns within protein datasets, potentially leading to more sophisticated and accurate classification models.

VI. REFERENCES

[1] Sandaruwan PD, Wannige CT. An improved deep learning model for hierarchical classification of protein families. PLoS One. 2021 Oct 20;16(10):e0258625. doi: 10.1371/journal.pone.0258625. PMID: 34669708; PMCID: PMC8528337

[2] Zhang D, Kabuka MR. Protein Family Classification from Scratch: A CNN Based Deep Learning Approach.

IEEE/ACM Trans Comput Biol Bioinform. 2021 Sep-Oct;18(5):1996-2007. doi: 10.1109/TCBB.2020.2966633. Epub 2021 Oct 8. PMID: 31944984.

[3] G. Zhou and W. Chen, "Protein Functional Family Classification Based on Multilevel Feature Information," 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 2022, pp. 1836-1839, doi: 10.1109/BIBM55620.2022.9994943.

[4] Jour, Huang, Tao,Cao, Jiuwen, Xiong, Lianglin , 2014 Protein Sequence Classification with Improved Extreme Learning Machine Algorithms, BioMed Research International, SP 103054, VL - 2014, 2314-6133, 10.1155/2014/103054

[5] Shehu Mohammed Yusuf, Fuhao Zhang, Min Zeng, Min Li, DeepPPF: A deep learning framework for predicting protein family, Neurocomputing, Volume 428, 2021, Pages 19-29, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2020.11.062.

[6] Nguyen, Np., Nute, M., Mirarab, S. et al. HIPPI: highly accurate protein family classification with ensembles of HMMs. BMC Genomics 17 (Suppl 10), 765 (2016). https://doi.org/10.1186/s12864-016-3097-0

[7] Diplaris, S., Tsoumakas, G., Mitkas, P.A., Vlahavas, I. (2005). Protein Classification with Multiple Algorithms. In: Bozanis, P., Houstis, E.N. (eds) Advances in Informatics. PCI 2005. Lecture Notes in Computer Science, vol 3746. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11573036_42