IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Implementation Of Machine Learning Model For Classification Of Novel Amino Acids

Author:
Aastha Katiyar
Student of M.Tech., Department of Computer Science Engg
Institute of Engineering & Technology, Sage University, Indore

Guide:
Deepak K Yadav
Associate Professor, Department of Computer Science Engg
Institute of Engineering & Technology, Sage University, Indore

Abstract**—** Amino acids are traditionally categorized based on their biochemical attributes. In this study, the focus shifts to reorganizing amino acids solely according to structural statistics, thereby mitigating existing chemical biases. The proposed machine learning model is to propose classification of the novel amino acid based on their structural insights. Main aim is to study the protein sequences of the candidates, that have been crystallized and their X-ray structures are well known. These are stored in the Protein Data Bank (PDB). Expansion in available amino acids eventually slows down the speed of identifying the protein constituted by the number of amino acids. Hence, there is a requirement for an effective model to predict the secondary protein structure swiftly and accurately. This model has been designed to anticipate the protein's secondary structure through specific predefined tasks.

Our Aim is to study the protein sequences of the candidates who have been crystallized and their X-ray structures are known and deposited in Research Collaboratory for Structural Bioinformatics (RCSB) PDB. The sequences of all the known crystal structures are to be downloaded from PDB and then a table is to be prepared based on occurrence of each amino acid on independent positions in the range of secondary structure. Once the data is ready, the noise must be removed accordingly, and the inferences will be made by polishing the data above a reasonable threshold. This focuses on forecasting the required parameters by analyzing the arrangement of amino acids and their neighboring context.

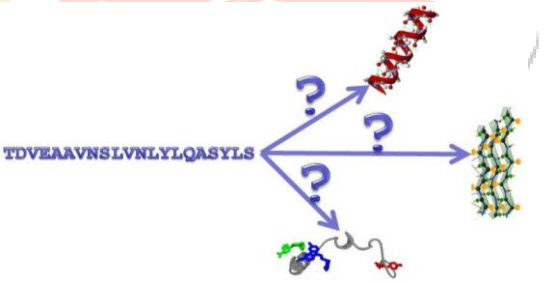

FIG. 1. PREDICTION BY PROTEIN FASTA SEQUENCE

Keywords**—** Bioinformatics, Protein structure, Amino acid classification, Random Forest Classifier, Decision Tree, Cross-validation, Feature extraction, Feature selection, Sequence analysis, Structural insights, FASTA Sequence, RCSB, Protein Data Bank, Protein sequence analysis, Predictive modeling, Data Science, Machine Learning Model.

I. INTRODUCTION

Amino acids are fundamental biomolecules, constituting organic compounds. Proteins are constructed from numerous amino acids linked through polypeptide bonds, creating a chain structure as indicated in Refer Table [1]. Currently, 20 amino acids are identified as the foundational elements of proteins. Out of these, nine are deemed essential, requiring dietary intake, while five are nonessential, as the human body can synthesize them. The

remaining six amino acids vital for protein synthesis are conditional, necessary only in specific life stages or health conditions.

A. *Molecular Composition of Amino Acid*

A typical amino acid consists of an amino group, a carboxyl group and a side chain known as the R group. specific to each amino acid. Helix-forming, sheet-forming, and coil-forming amino acids are basic Amino Acid conformations and secondary structure. Refer Fig [2].

Proteins are created through the sequential arrangement of acids, in a chain. An alpha helix is formed by twisting this chain of acids into the shape of a spiral. Beta Pleated sheets are the strands that consist of 3-10 Amino Acid residues. These are used in the representation of amino acids.

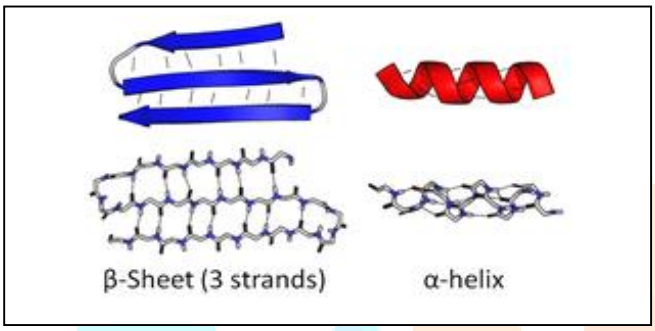

FIG. 2. BETA SHEETS AND ALFA HELIX STRUCTURE.

II. PROBLEM STATEMENT

In the existing system, there are a variety of machine learning models, deep learning algorithms and neural network models available for the secondary structure prediction of protein. Neural networks have proven to be the most effective computational approach for secondary structure prediction, surpassing other methods like statistical approaches, nearest neighbor methods, hidden Markov models (HMM), and support vector machines (SVM). This algorithm's efficiency, accuracy, and cost-effectiveness make it a superior choice when compared to other algorithms that employ more complex techniques, demanding larger computational resources for structure prediction.

On average about 70% accuracy was obtained. The focus of all these models is regrouping into acids. This is not much but is time consuming task and these algorithms and model need a high amount of computation power.

On the other hand, the initial step of the research is to State the problem and investigate the required data. Identification of data sources for data gathering and collection. The current study mainly focuses on regrouping into acids based on original structural statistics and thus removing the chemical bias. The Model aims to predict the type of protein for a particular sequence by analyzing the existing sequences. For all the analysis and prediction, it uses Random Forest Classifier (RFC) as it classifies the results appropriately. The model will provide a solution which is less time consuming and requires comparatively less computation power with equal and improved accuracy. Fig [3] shows how sequences can be transformed to Secondary structure of proteins. It will perform a set of predefined operations or tasks to determine the secondary structure of the protein:

i. Amino acid at frequency calculation along each position in the secondary structure. Refer Fig [8].

ii. Find the relative occurrence of each amino acid with respect to all individual amino acids Refer Fig [8].

iii. Pattern recognition of amino acids sets, and the effect of flanking in secondary structure.

iv. Regrouping of Amino acids free from chemical bias.

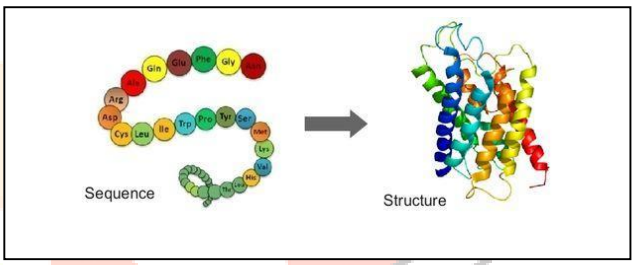

FIG. 3. SECONDARY STRUCTURE FROM FASTA SEQUENCE

III. METHODOLOGY

A. *Algorithm: Random Forest Classifier (RFC)*

Random Forest belongs to supervised learning kind of algorithm. It is widely used because of its simplicity and flexibility. Without parameter tuning Random Forest often produces excellent results. What makes it more appealing is that it can be applied to both classification and regression tasks.

B. *Ensemble learning and Decision Tree*

The forest created is a collection of Decision Trees, typically trained using the "bagging" method. This technique involves combining multiple learning models to enhance overall outcome. By merging decision trees predictions become more precise and reliable. Measuring the influence of each feature, on the prediction is a process. Additionally, it is straightforward to assess the neighbour and relative importance of each feature in the prediction process.

Of focusing on the most crucial feature, during the node splitting process the algorithm instead chooses the best feature, from a randomly selected subset. This method promotes diversity, which often results in the creation of an model.

C. *Advantages of RFC:*

1) One major advantage of using a forest is its versatility, in handling both classification and regression problems, which are commonly encountered in machine learning systems today.

2) Random Forest is also regarded as an user friendly algorithm as its default settings often yield accurate prediction results.

3) This algorithm is also an option when time constraints require us to develop a model.

4) Additionally it offers a measure of the significance it assigns to your features.

Protein is formed by the combination of multiple amino acids arranged in a chaining sequence. The aim of the model is to study the protein sequences of the candidates who have been crystallized and their X-ray structures are known and deposited in PDB.

The sequences of all the known crystal structures are to be downloaded from PDB. Downloaded FAST Alignment of all (FASTA) Sequence is blasted on Defined secondary structure of protein (DSSP) server to get sequence of structural arrangement of all the amino acid in a particular chain of secondary structure of protein. Fig [4] demonstrates the FASTA Sequence format. The DSSP method is commonly used to classify the structure of a protein. Afterward we need to create a table that shows the frequency of each amino acid at positions, within the secondary structure range. Once the data is ready, the noise must be removed accordingly, and the inferences will be made by polishing the data above a reasonable threshold.

```
>101M:A:sequence
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGA
ILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKEL
GYQG
>101M:A:secstr
  HHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHH GGGGGG TTTTT  SHHHHHH HHHHHHHHHHHHHHH
HHTTTT  HHHHHHHHHHHHTS  HHHHHHHHHHHHHHHHHH GGG SHHHHHHHHHHHHHHHHHHHHHHHHT
T
```

FIG. 4. FASTA SEQUENCE FORMAT

Methods for determining protein structure, such as X ray crystallography and nuclear magnetic resonance (NMR) are both expensive and time consuming, making them unsuitable for large scale applications on a level. However, there are cost methods, including machine learning that can be used to predict secondary structures. Predicting secondary structures is crucial for predicting 3 structures. Therefore, the development of more methods for predicting secondary structures remains a significant priority.

Furthermore, ensuring high prediction accuracy is equally essential to uncover the impact of the specific content and relevant positions of amino acids on the accuracy of predictions.

D. *Workflow of Machine Learning Model:*

The basic workflow model is demonstrated in the below diagram. That fairly describes a set of sequences of amino acids and respective proteins that are formed by the combination of them. This data will go through the classification task, then regrouping of amino acids will take place to form a new protein structure and this new structure is the final output.

Fig [5] demonstrates the machine learning model workflow used to achieve the data collection and model building.

1) Data collection and preparation: Collected a dataset containing amino acid sequences and corresponding labels. FASTA Sequences are extracted from PDB and are blasted on DSSP server to ingest into machine learning model Fig [4].

2) Data analysis: Data is organized and processed to handle the anamolies like missing data, bad data, currupt data and noise removal from dataset before fit into machine learning model.

3) Data cleansing: To get FASTA Sequence, structure details, amino acid, sequence and secondary structure.

4) Data preprocessing: Based on neighbour amino acid, their relative and total frequency is calculated from FASTA Sequence.

5) Training and Testing Data: Data is splitted to training and testing dataset. The training set teaches the Random Forest model, the validation set assists in adjusting settings for better performance and helps hyperparameter tuning, and the test set judges how well the final model works.

6) Model Generation and Selection: Dataset on training group is fit to different machine learning algorithms and model is generated to get the most accurate model.

7) Prediction: Predicted values are geneared againset real values from the testing dataset.

8) Calculate performance metrices: Accuracy estimation of model is tested based on confusion metrix and accuracy score.

9) Data Visuallisation: The generated result are visuallised with different bar chart and pie chart.

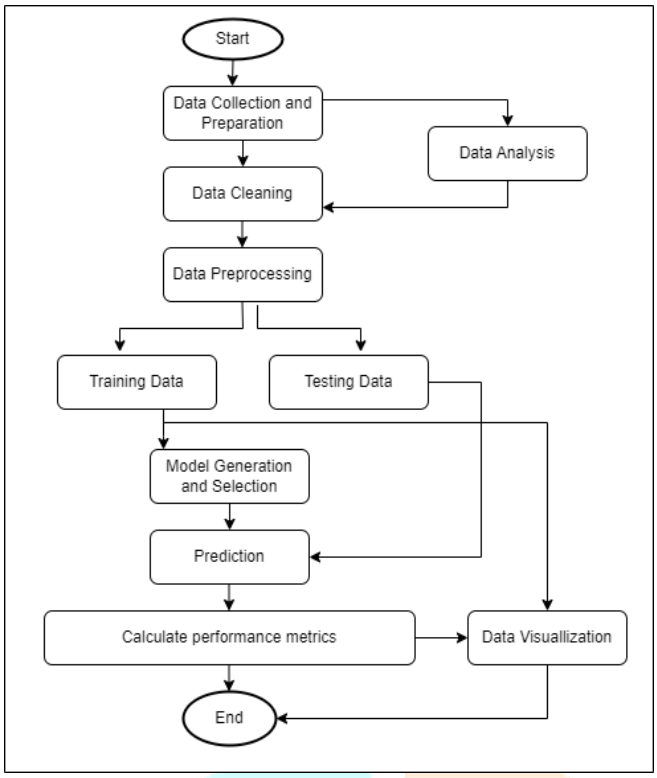

FIG. 5. MACHINE LEARNING MODEL WORKFLOW

IV. IMPLEMENTATION

This model is implemented on the concept of machine learning methodology. Models are trained in multiple machine learning methods to get better fit and accuracy. Random Forest machine learning model is the result of average accuracy of multiple decision trees on which the model is trained.

A. *Tools*

DSSP server and RCSB PDB is used to extract the amino acid and protein sequence in pearson format of Fasta sequence. The application code is build on Jupyter Notebook open source platform, web application for creating and sharing computational documents. Programming language is Python and its machine learning libraries, it is free and open source.

People, especially data miners and scientists, use Python programming language for tasks like analyzing data, performing statistical computations, and creating graphics. The machine learning model is built on Scipy, Skit learn, Matplotlib, Pandas and Numpy library.

1) Software Specifications:

Python Language Version	3.7.0
Jupyter Notebook Version	5.7.0
numpy Version	1.15.4
Pandas Version	0.23.4
Scikit Learn Version	0.20.0
matplotlib Version	1.2.0
Scipy version	1.1.0
DSSP	1.1.0
RCSB	1.1.0

2) Hardware Specifications:

System Type	Windows 64-bit operating system, x64-based processor
Processor	11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 1.38 GHz
Memory	16 GB
Graphics	Intel (R) Iris(R) Xe Graphics
Operating System	Windows 10 Pro

B. *Methodology*

Fundamental analysis aims to determine the existing chains of amino acids and the proteins formed by them. On the flip side, technical analysis involves examining the current chains and attempting to predict the kind of protein a new amino acid sequence will create. Refer Fig [5] for methodology and workflow of model building. The below description shows the experimental and implementation screenshots from Jupyter Notebook.

1) Import Dependencies and Dataset: The data used as a training dataset is a .txt file that contains the sequence of amino acids from which the protein is formed, structure id, sequence/secstr etc. To train the machine learning model, we often need dataset in CSV format so that it will be easy for the classification task.

	Id	Fasta_Sequence
0	>101M:A:sequence	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...
1	>101M:A:secstr	HHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHH GGGGGG TT...
2	>102L:A:sequence	MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSE...
3	>102L:A:secstr	HHHHHHHHH EEEEEE TTS EEEETTEEEESSS TTTHHHHH...

FIG. 6. INITIAL DATASET POST FASTA SEQUENCE SPLIT

2) Transformed Fast Sequence data to fit into machine learning model

Fasta_Sequence	Structure_Id	Amino_Acid	Sequence/Secstr
MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR...	101M	A	sequence
HHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHH GGGGGG TT...	101M	A	secstr
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAAKSE...	102L	A	sequence

Amino_Acid	Structure_Id	0	1	2	3	4	5	6	7	...	1221	1222	1223	1224	1225	1226	1227	1228	1229	Sequence/Secstr
7	1000	113	121	112	118	106	108	106	122	...	101	101	101	101	101	101	101	101		1
7	1000	100	100	100	100	109	109	109	109	...	101	101	101	101	101	101	101	101		0
7	1001	113	114	110	107	106	113	112	117	...	101	101	101	101	101	101	101	101		1
7	1001	100	100	109	109	109	109	109	109	...	101	101	101	101	101	101	101	101		0
7	1002	113	121	112	118	106	108	106	122	...	101	101	101	101	101	101	101	101		1

FIG. 7. DATA ENCODING AND NORMALIZATION

3) Relative Frequencey of Amino Acid:

	0	1	2	3	4	5	6	7	8	9	...	N	P	Q	R	S	T	V	W	X	Y
499			E	E	E	E	E	E	E	E	...	0.0	0.0	0.0	0.0	43.0	45.0	0.0	0.0	0.0	0.0
407				S	S					E	...	0.0	0.0	0.0	0.0	34.0	43.0	0.0	0.0	0.0	0.0
765											...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
405				S	S					E	...	0.0	0.0	0.0	0.0	34.0	43.0	0.0	0.0	0.0	0.0
403			S			T	T				...	0.0	0.0	0.0	0.0	37.0	45.0	0.0	0.0	0.0	0.0

	left	character	right
0	M	N	I
1	N	I	F
2	I	F	E
3	F	E	M

FIG. 8. DERIVING RELATIVE FREQUENCY

4) Fit the model to Random forest classifier

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini'
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=120, n_jobs=1,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

FIG. 9. FITTING RANDOM FIREST CLASSIFIER WITH PERFORMANCE PARAMETERS

V. PROCESS FLOW

Below are the steps to define process flow from Data collection to performance evaluation. Fig [10] demonstrates the working and steps involved to achieve the model.

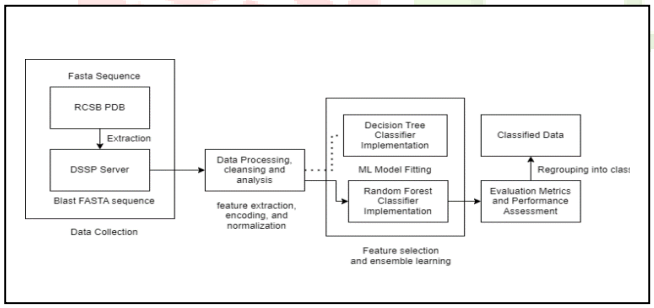

FIG. 10. MACHINE LEARNING MODEL PROCESS FLOW

A. *Data Collection:*

Dataset with classification categories or classes is require to analyse the sequences and amino acid arrangements. Fig [6] shows the image of raw dataset collected from PDB.

1) RCSB PDB: Protein Data Bank is used for getting Database of three dimensional structural data for large bimolecular (Protein). Scientist submits structural data on PDB. PDB is key in area of structural biology. Downloaded dataset of FASTA sequences.

2) DSSP Server: It is a web-based service that allows researchers to submit protein structure files in PDB format. It does not predict the secondary structure of Protein. DSSP stores chain ID of each residue under column 12 as a single character.

Data is collected from RCSB PBD official website. It is not necessary that all the strings of aminos will have the same length therefore in making the string lengths same, some NAN values had encounter which are bad for our model and must be removed. In this plot black rows shot the filled values, and the rest shows the NAN values which must be removed to apply Machine learning Model. Refer Fig [11] shows images from matplotlib for Blank data. Refer Fig [12] shows graph shows the correlation between all the caulis in dataset that is all the aminos.

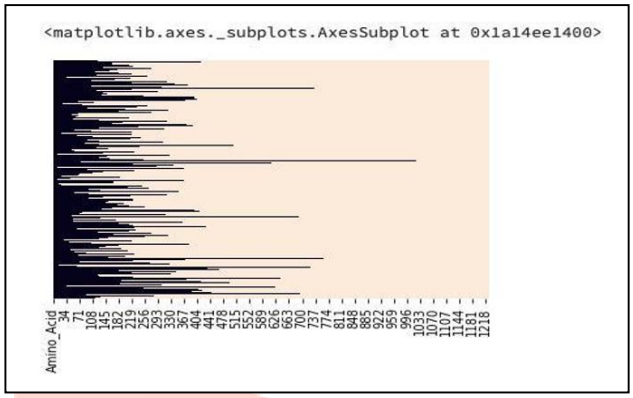

FIG. 11. PDB DATASET AMINO ACID PLOT

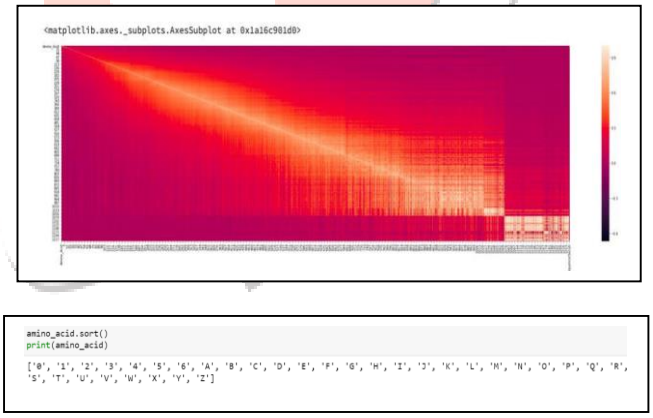

FIG. 12. CORRELATION GRAPGH AMONG AMINO ACIDS

B. *Data Preprocessing, Cleansing and Analysis:*

Data Processing steps feature extraction, encoding, and normalization Fig [7]. Preprocess the data by performing feature extraction to represent the amino acids in a suitable format for machine learning. It includes encoding amino acids. Fig [11] and Fig [12] shows analysis results.

C. *Feature Selection:*

Selected relevant features that capture the important characteristics of the amino acids for classification mainly their molecules. Refer Fig [8].

FIG. 13. UNIQUE AMINO ACID AS FEATURES

D. ML Model Fitting and Ensemble Learning:

Ensemble learning trains the model on multiple and various machine learning algorithms' achieve better accuracy, below two methodologies are implemented. Fig [9] demonstrates RFC Model.

1) Decision Tree Classifier: Compared the performance of RFC with one of the Classification method DTC for amino acid classification. Accuracy of DTC was quite low.

2) Random Forest Classifier: Initialized an RFC with appropriate hyperparameters (e.g., number of trees, depth of trees, etc.). Trained the Random Forest model on the training data using the selected features. Refer Fig [9].

E. Evaluation Metrics and Performance Assessment:

To assess the model's performance and refine its parameters, we evaluated its accuracy on the validation set utilizing the AccuracyScore metric and cross-validation technique..

F. Classified Data:

Tested and predicted values drawn as data points in Pie chart and Scatter plot. Refer Fig [16] and Fig [17].

VI. EXPERIMENTAL RESULTS

Random Forest was chosen for this model due to its significantly better outcomes compared to the Decision Tree classifier, as illustrated in the figure below. The study collected fundamental and technical data and conclusive reports from various online sources. We have gathered the data from Protein Data Bank (PDB), a repository that mainly deals with Biomolecules. The data which we have obtained is in the form of a txt file but for our Machine Learning model it must be in CSV file, so we have converted it to Comma Separated Value (CSV). Researchers have made many models for protein structure prediction like Swiss one of the best but not as good as they wanted it to be. These models do not have a good accuracy so bioinformatician cannot rely on them.

```
accuracy_score(y_test,predictions)
0.3837681612248086
```

FIG. 14. DECISION TREE PERFORMANCE ESTIMATION

```
accuracy_score(y_test,rfc.predict(X_test))
0.80776441181066
```

FIG. 15. RANDOM FOREST CLASSIFIER PERFORMANCE ESTIMATION

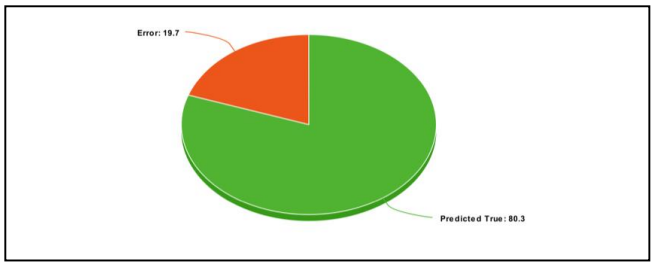

FIG. 16. PIE CHART VISUALIZATION

So, we tried to implement a model that can classify the proteins more accurately. The model uses RFC to predict the type of protein. In this LR we will first look at the data and its analysis then we will propose RFC ML Algorithm to perform prediction then finally predictive report can be generated.

Thus, Fig [14], Fig[15] clearly demonstrate the result of algorithms-Random Forest and Decision Tree, It concludes the impact of Random Forest (Accuracy = 0.80) Fig [12] is far superior to the Decision Tree (Accuracy = 0.38) algorithm Fig [11]. It indicates that the model developed using Random Forest has provided the reliable prediction.

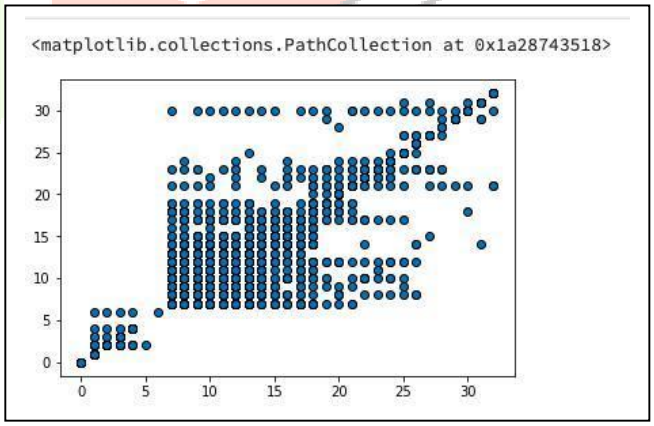

FIG. 17. SCATTER PLOT VISUALIZATION

VII. CONCLUSION

This will be an open-source program, anyone can install and pay for it. They just need to install dependencies with Python language and required scientific libraries to run this model. Anyone can install this in .exe format since it is not a GUI based application therefore user must store that file into some directory for its efficient use. To run the model, you must first open the console, get into the directory where locates the .exe file. Once you are in the directory you can just run the master file as a python program. We can get predicted results on the platform.

This allows us all the features with support of extension and well predicted model. This satisfies all the functional requirements. can work on large data. As of now this model mainly focuses on the secondary structure of the chain formed by the combination of different distinct or same kind of amino acids, but there are some more features that can be used as an Extension to this model.

A. *Achieving Accurate Classification:*

Due to RFC, this study successfully demonstrated the effectiveness of machine learning applications in classifying amino acids based on their structural properties.

B. *Insights into Amino Acid Properties:*

By examining the importance of features and their scores, we gained valuable insights into the importance of various amino acid properties in the classification process. This knowledge enhances our understanding of the connection between sequence structure and functional classification.

C. *Robustness and Generalization:*

The robust performance of the *RFC* across different datasets and cross-validation folds emphasizes its generalization capability. This suggests that the model is capable of accurate amino acid classification even when applied to novel, unseen sequences.

D. *Potential Applications in Biomedicine:*

The accurate classification of amino acids has promising implications in various biomedical domains, such as protein function prediction and disease classification. The model's ability to capture subtle sequence variations provides a foundation for further exploration and development of diagnostic and therapeutic tools. Along with secondary structure prediction, the model can find its huge involvement in bioinformatics fields to predict the ab initio with great accuracy. Its result serves as secondary structure function dictionary or researchers to have better prediction to choose mutation site. Collectively these help in identifying the medicines for the disease caused by different proteins that may help the doctors to treat their patient. We can also apply the trained model to classify new, unseen amino acid sequences.

TABLE I. LIST OF ABBREVIATIONS OF AMINO ACIDS

Amino Acid	"Three-Letter" Abbreviation	One-Letter Abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y

VIII. REFERENCES

[1] Tiwari, A., Kumar, A., & Dehzangi, A. (2020). Protein classification into fold classes using random forest and deep features. IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[2] Sajjan, S. S., & Kulkarni, B. D. (2020). A Review on Feature Extraction and Selection of Amino Acids for Protein Classification. Proceedings of the 2020 3rd International Conference on Data, Engineering and Applications (IDEA).

[3] Kar, P., & Scheinberg, K. (2020). Learning to Classify Proteins by Fold Class Using Multiple Alignments and Random Forests. Journal of Computational Biology.

[4] Sharma, A., & Gaur, V. (2020). Prediction of Subcellular Localization of Eukaryotic Proteins Using Chou's General Pseudo Amino Acid Composition and Random Forest Algorithm. Evolutionary Bioinformatics.

[5] Saravanan, R., & Anitha, R. (2020). Application of Random Forest Classifier for Classification of Amino Acids in Proteins. In Proceedings of the 2020 International Conference on Data Science and Applications (ICDSA).

[6] S. Saraswathi Battelle, J. L. Fernández-Martínez, A. Koliński, R. L. Jernigan, A. Kloczkowski Battelle, Distributions of amino acids suggest that certain residue types more effectively determine protein secondary structure, 2013 October

[7] Mingchuan Fu, Zhuoran Huang, Yuanhui Mao and Shiheng Tao, Neighbor Preferences of Amino

Acids and Context-Dependent Effects of Amino Acid Substitutions in Human, Mouse, and Dog, 27 August 2014 / Accepted: 2 September 2014

[8] Gianluca Pollastri and Aoife McLysaght, "Porter: a new, accurate server for protein secondary structure prediction"

[9] Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:211–215.

[10] Garnier J, Osguthorpe DJ, Robson B. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120

[11] Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584 –599. 220 –223