Protein Enzyme Sequence Class Prediction using Computational Model

Satyendra Kumar Bhardwaj[1], Shivam Vishwakarma[1], Anand Bihari[2], Sudhakar Tripathi[1], Sanjay Agrawal[3], and Puneet Joshi[3]

ARTICLE INFO

Article history:
Received: 21 June 2022
Revised: 17 October 2022
Accepted: 11 November 2022

Keywords:
Protein Sequence
Enzyme class
Machine Learning
Computational Model
me Profile

ABSTRACT

Proteins are the most necessary and multipurpose macromolecules of human life and the functions of proteins having a major impact in the development of new drugs and helps in understanding the new disease. So, Experimental approaches for protein function prediction are integrally low output and too much time consuming and costly too. One of the difficult issues in bioinformatics nowadays is predicting the protein function of an unknown protein. As the number of proteins grows, the prediction of Protein class opens up new opportunities for bioinformatics researchers. This research implements the machine learning techniques to predict the appropriate class of the proteins. In this research investigation, the SVM, Logistic Regression, Decision Tree, Random Forest, Adaboost, Naive Bayes, and KNN has been employed on 12285 protein data that are taken from the Kaggle Data Repository and classified into 27 classes. Protein data is high-order sequence data containing up to various features but in this article protein Sequence feature is used. The proposed method highlights that the Random Forests outperforms when we compare the outcome of the other machine learning techniques.

1. INTRODUCTION

Protein function prediction is a technique used by bioinformatics engineers to determine the biological or biochemical role of proteins. Protein is a macromolecule that functions as a component and functional component of cells, and accounts for the second largest proportion of cell weight after water. Currently, there are over 550 fully sequenced cell biology genomes, contributing to over 5 million unique protein sequences in publicly available databases [43].

To understand the biological role of these protein sequences, we need to understand their function. This is a goal that leads to many experimental and computational methods for assessing protein function. Protein Function Prediction is moving from using individual data sources for prediction to integrating different data sources and methods to achieve this goal. In general, computational protein function prediction relies on two foundations: data sources and predictive models / methods. Computational Protein Function Prediction Standards are not widely accepted, and each existing method has its own limitations, so the trend in this area of study is to combine available data sources and methods to enhance function. It is to predict effectively. This combination refers to data source integration, model / method integration, or both [44].

In this article protein function prediction is done using the protein sequence feature. The sequence of a protein is usually written as a sequence of letters corresponding to the order of amino acids from the amino terminus to the carboxy terminus of the protein. We can use a one-letter or three-letter code to represent each amino acid in the sequence [45-46, 48-52].

In this article features are extracted from protein sequence and then applied various machine learning algorithms then after all result analysis and comparison are done.

2. DATASETS

Dataset is the primary thing that is more necessary for building a model. The dataset must have the large and various numbers of proteins and their family type as well as sequence of protein. In this research two datasets are used which is downloaded from Kaggle [47]. The dataset consists of many different types of macromolecules of biological significance. The majority of the data records are of proteins. Both datasets have different number of features and data. The first dataset have 141401 rows × 14 columns and second dataset have 467304 rows × 5 columns. The raw dataset required data preprocessing to implement the model. Since first dataset has classification

[1]*Department of Information Technology, Rajkiya Engineering College Ambedkarnagar, U. P., India.*
[2]*Department of Computational Intelligence, School of Computer Sciene and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.*
[3]*Department of Electrical Engineering, Rajkiya Engineering College Ambedkarnagar, U. P., India.*
**Corresponding author:* Anand Bihari, Email: anand.bihari@vit.ac.in, csanandk@gmail.com.

and second dataset has protein sequence so the both datasets are joined together. Dataset has various raw attributes but we required two attributes first one is sequence of protein and second is classification of protein. So we filter all the protein and drop the other attributes from the dataset. Then null values from the dataset are removed and we have found value counts of each protein and keep those proteins only which value count is greater than 100. So the final filtered dataset Description is given following in Table 1.

Table 1: Dataset descriptions

Attributes' Name	Data count
Classification	12285
Sequence	12285

Table 2: Class wise total sequence data

Class ID	Classification/Class	Total Count
C1	HYDROLASE	2214
C2	OXIDOREDUCTASE	1956
C3	TRANSFERASE	1497
C4	LYASE	848
C5	HYDROLASE INHIBITOR/HYDROLASE	691
C6	IMMUNE SYSTEM	583
C7	ELECTRON TRANSPORT	466
C8	ISOMERASE	329
C9	TOXIN	326
C10	LIGASE	295
C11	SIGNALING PROTEIN	294
C12	VIRUS	258
C13	OXYGEN TRANSPORT	247
C14	VIRAL PROTEIN	235
C15	TRANSCRIPTION	223
C16	IMMUNOGLOBULIN	202
C17	DNA BINDING PROTEIN	193
C18	CHAPERONE	183
C19	HORMONE/GROWTH FACTOR	172
C20	OXYGEN STORAGE/TRANSPORT	165
C21	SUGAR BINDING PROTEIN	146
C22	PHOTOSYNTHESIS	144
C23	TRANSPORT PROTEIN	132
C24	MEMBRANE PROTEIN	130
C25	APOPTOSIS	127
C26	METAL BINDING PROTEIN	117
C27	CYTOKINE	112

The Table 1 depicts that the dataset has total 12564 sequence data also the same number of data with the classification of the data with 27 unique classes. That means total sequence data are categories into 27 distinct classes. Table 2 describes the class and their corresponding count of the proteins.

Next, we have extracted all the features using CountVectorizer approach. In this approach, data were split into several small chunk based on given delimiter and use the total count the each chuck for further processing.

3. METHODOLOGY

After extracting the features now, we have to implement the machine learning algorithm to find the optimal learning algorithm to classify and predict the protein sequence. In this research, we have implemented SVM, Random forest, Logistic Regression, Naïve Bayes, KNN, Decision tree and Adaboost.

3.1 Support Vector Machine (SVM)

The SVM is a popular supervised learning technique for classification and regression problems. [1-6].

The goal of the SVM method is to offer the optimal lines or decision boundaries for categorizing n-dimensional space, making it easy to assign further data points to the correct category in the future.

SVM selects an extreme/vector that benefits in the construction of a hyperplane. The method is called a support vector machine because these extreme cases are known as support vectors.

3.2 Random Forest (RF)

Random forest is a well-known machine learning technique that employs the supervised learning approach. [16–19]. It can be used to solve classification and regression problems in machine learning. It is based on the concept of ensemble learning, which combines different classifiers to address complex difficulties and improve model functioning.

"Random forest is a classifier that takes a set of decision trees for various subsets of a specific dataset and then takes the average to enhance the predicted accuracy of that dataset." In this model, multiple trees are generated based on the data and getting prediction from each tree. Then final result has been predicted based on the vote of the prediction of the initial trees. But one of the major issues with the random forests is the higher precision and overfitting, if the number of tree is more in the forests.

3.3 Logistic Regression

The logistic regression [7–10] forecasts the likelihood of a target variable. One of the few ML algorithms, it is utilized

for a variety of classification issues, including spam identification, diabetes prediction, and cancer diagnosis.

Calculations like logistic regression are used to foretell binary outcomes. Either something happens, or nothing happens. These options include Yes/No, Pass/Fail, Alive/Dead, and so forth. The binary outcomes, which can be classified into one of two groups, are determined by analyzing the independent variables.

Dependent variables are usually categorical variables, but independent variables might be either categorical or quantitative.

3.4 Naive Bayes

The Nave Bayes algorithm, which is based on the Bayes theorem, is a supervised learning technique for classification problems. [25-29].

One of the simplest and most effective classification algorithms known at the moment is the Naive Bayes Classifier. It helps with the rapid creation of machine learning models that are capable of producing reliable predictions. Because it uses a probabilistic classifier, it bases its predictions on the likelihood that a certain event will take place.

3.5 K-Nearest Neighbours Algorithm:

K-Nearest Neighbor is one of the most basic supervised learning-based machine learning algorithms.

The K-NN method assumes that the new case and the previous cases are comparable, and it places the new instance in the category that is most similar to the existing categories. [30-35].

After saving all prior data, a new data point is classified using the K-NN algorithm based on similarity. This suggests that utilizing the K-NN technique, fresh data may be consistently and quickly classified. [42].

3.6 Decision Tree:

It is called a supervised machine learning method and used to solve the classification and regression related problem. Basically, it is a tree based classification technique. In this method, the nodes represent the data and an edge represents the decision parameters and results are stored in the leaf nodes [11–15].

In this model, training has been done by the splitting of data into several groups based on the similarity of the attribute value. This model required to train the data in several steps that s called recursive partitioning. This process will end when there is no further division is possible or have the same value for the targeted output. It does not need parameter setting or domain expertise, decision tree classifier construction is appropriate for exploratory knowledge discovery. Decision tree is capable to handle high-dimensional data.

3.7 AdaBoost:

AdaBoost [20–24], commonly referred to as adaptive boosting, is an ensemble machine learning approach. A single-level decision tree, also known as a one-part decision tree, is the most frequent algorithm employed in AdaBoost.

Decision stumps are another name for these trees. This algorithm creates a model that equally weights each piece of information. Then give the incorrectly categorized points more weight. All points with high weights will be heavier in this model.

3.8 Performance Evaluation

To gauge the effectiveness of our machine learning model we have used the confusion matrix. It offers a very straightforward and effective way to measure the model's performance [36].

The confusion matrix allows us to employ a variety of performance criteria, some of which are listed below:

3.8.1 Accuracy

Based on the percentage of all samples that the classifier correctly identified, we may get the model's overall accuracy from all of the data [37]. The following equation can be used to determine Accuracy :

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

3.8.2 Precision (PRE)

The proportion of correct positives to all positives is known as PRE [38]. The following equation can be used to determine Precision:

$$\textbf{\textit{Precision}}\ (\text{PRE}) = TP / (TP+FP) \quad (2)$$

3.8.3 Recall (Recl)

This demonstrates [39] What percentage of all positive samples was correctly recognized as positive by the classifier. The following equation can be used to determine recall:

$$\text{Recall (RC)} = TP / (TP + FN) \quad (3)$$

3.8.2 F1

It is the combination of recall and PRE. Mathematically, it is a harmonic mean of accuracy and recall [40]. The following equation can be used to determine F1:

$$\text{F1 Score} = 2*(\text{PRE} * \text{Recl}) / (\text{PRE} + \text{Recl}) \quad (4)$$

3.8.2 AUROC

AUROC [41] is a performance metrics that shows the discrimination ability of the in case of positive or negative samples. The Model AUROC is high that means models are correctly discriminate the sample data. The following equation can be used to determine recall:

$$\text{AUROC}=((TP/(TP+FN))+(TN/(TN+FP)))/2 \quad (5)$$

4. RESULTS AND DISCUSSION

There are various machine learning algorithms implemented and these algorithms are SVM, Logistic Regression, Decision Tree, Random Forest, Adaboost, Naïve Bayes, and KNN. The result of these implemented algorithms can be measured by the value of accuracy, PRE, recall, and F1 score. The result of the SVM model is shown in table 3. The SVM overall accuracy is 0.90, and the PRE, recall and f1 score is 0.94, 0.86 and 0.89 respectively.

Table 3: Result of the SVM model.

	PRE	recall	F1
HYDROLASE	0.79	0.9	0.84
OXIDOREDUCTASE	1	0.97	0.99
TRANSFERASE	1	0.62	0.76
LYASE	1	0.73	0.85
HYDROLASE INHIBITOR/HYDROLASE	0.96	0.88	0.92
IMMUNE SYSTEM	1	0.89	0.94
ELECTRON TRANSPORT	0.86	0.92	0.89
ISOMERASE	0.55	0.89	0.68
TOXIN	0.89	0.92	0.9
LIGASE	0.95	0.8	0.87
SIGNALING PROTEIN	1	0.93	0.97
VIRUS	1	0.9	0.95
OXYGEN TRANSPORT	0.99	0.97	0.98
VIRAL PROTEIN	0.9	0.69	0.78
TRANSCRIPTION	1	0.65	0.79
IMMUNOGLOBULIN	0.98	0.95	0.96
DNA BINDING PROTEIN	0.91	0.7	0.79
CHAPERONE	0.89	0.92	0.9
HORMONE/ GROWTH FACTOR	0.96	0.9	0.93
OXYGEN STORAGE/ TRANSPORT	0.92	0.77	0.84
SUGAR BINDING PROTEIN	1	0.9	0.95
PHOTOSYNTHESIS	1	0.87	0.93
TRANSPORT PROTEIN	0.95	0.86	0.9
MEMBRANE PROTEIN	0.95	0.93	0.94
APOPTOSIS	0.9	0.78	0.84
METAL BINDING PROTEIN	1	0.88	0.94
CYTOKINE	0.96	0.96	0.96
Overall accuracy			**0.9**

Table 3 shows that the SVM model gives high PRE; Recl and f1 score for most of the classes. Also giving an average acceptable accuracy. Next we have model the system using Decision tree. The decision tree model result is shown in table 4. The Decision Tree overall accuracy is 0.22, the average PRE, recall and f1 score is 0.11, 0.08 and 0.7 respectively.

Table 4: Result of the Decision Tree model

	PRE	recall	F1
HYDROLASE	0	0	0
OXIDOREDUCTASE	0	0	0
TRANSFERASE	0	0	0
LYASE	0	0	0
HYDROLASE INHIBITOR/HYDROLASE	0	0	0
IMMUNE SYSTEM	0	0	0
ELECTRON TRANSPORT	0.19	1	0.32
ISOMERASE	0.94	0.1	0.19
TOXIN	0.44	0.19	0.26
LIGASE	0	0	0
SIGNALING PROTEIN	0	0	0
VIRUS	0	0	0
OXYGEN TRANSPORT	0	0	0
VIRAL PROTEIN	0	0	0
TRANSCRIPTION	0	0	0
IMMUNOGLOBULIN	0	0	0
DNA BINDING PROTEIN	0.71	0.17	0.27
CHAPERONE	0.78	0.7	0.74
HORMONE/GROWTH FACTOR	0	0	0
OXYGEN STORAGE/TRANSPORT	0	0	0
SUGAR BINDING PROTEIN	0	0	0
PHOTOSYNTHESIS	0	0	0
TRANSPORT PROTEIN	0	0	0
MEMBRANE PROTEIN	0	0	0
APOPTOSIS	0	0	0
METAL BINDING PROTEIN	0	0	0
CYTOKINE	0	0	0
Overall accuracy			**0.22**

Table 4 shows that he decision tree based model is not giving good result. The accuracy and other measures value are very less. So, we cannot recommend this model for classification and prediction. Next we have model the system using Logistic Regression. The Logistic Regression model result is shown in table 5. The Logistic Regression

overall accuracy is 0.90, the average PRE, recall and f1 score is 0.94, 0.89 and 0.86 respectively.

Table 5: Result of the Logistic Regression model

	PRE	recall	F1
HYDROLASE	0.8	0.95	0.87
OXIDOREDUCTASE	1	0.97	0.99
TRANSFERASE	1	0.62	0.76
LYASE	1	0.71	0.83
HYDROLASE INHIBITOR/HYDROLASE	0.96	0.89	0.92
IMMUNE SYSTEM	1	0.89	0.94
ELECTRON TRANSPORT	0.87	0.94	0.9
ISOMERASE	0.65	0.84	0.74
TOXIN	0.86	0.94	0.9
LIGASE	0.97	0.82	0.89
SIGNALING PROTEIN	1	0.93	0.97
VIRUS	1	0.93	0.96
OXYGEN TRANSPORT	0.99	0.98	0.98
VIRAL PROTEIN	0.9	0.69	0.78
TRANSCRIPTION	1	0.7	0.82
IMMUNOGLOBULIN	0.97	0.96	0.97
DNA BINDING PROTEIN	0.91	0.67	0.77
CHAPERONE	0.87	0.92	0.89
HORMONE/GROWTH FACTOR	0.93	0.9	0.92
OXYGEN STORAGE/TRANSPORT	0.94	0.77	0.85
SUGAR BINDING PROTEIN	1	0.9	0.95
PHOTOSYNTHESIS	0.98	0.87	0.92
TRANSPORT PROTEIN	0.95	0.86	0.9
MEMBRANE PROTEIN	0.92	0.96	0.94
APOPTOSIS	0.9	0.78	0.84
METAL BINDING PROTEIN	1	0.88	0.94
CYTOKINE	0.96	0.98	0.97
Overall Accuracy			**0.90**

Table 5 shows that the Logistic Regression model gives high PRE; recall and f1 score for most of the classes. Also giving an average acceptable accuracy. Next we have model the system using Naive Bayes. The Naive Bayes model result is shown in table 6. The Naive Bayes overall accuracy is 0.87, the average PRE, recall and f1 score is 0.79, 0.85 and 0.81 respectively.

Table 6: Result of the Naive Bayes model

	PRE	recall	F1
HYDROLASE	0.77	0.95	0.85
OXIDOREDUCTASE	0.9	0.97	0.93
TRANSFERASE	0.39	0.71	0.51
LYASE	0.83	0.71	0.76
HYDROLASE INHIBITOR/HYDROLASE	0.88	0.9	0.89
IMMUNE SYSTEM	0.66	0.82	0.73
ELECTRON TRANSPORT	0.96	0.82	0.89
ISOMERASE	0.77	0.84	0.81
TOXIN	0.89	0.74	0.81
LIGASE	0.64	0.87	0.74
SIGNALING PROTEIN	0.96	0.93	0.95
VIRUS	0.98	0.93	0.96
OXYGEN TRANSPORT	0.96	0.97	0.97
VIRAL PROTEIN	0.45	0.81	0.58
TRANSCRIPTION	0.45	0.75	0.57
IMMUNOGLOBULIN	0.98	0.93	0.95
DNA BINDING PROTEIN	0.6	0.5	0.55
CHAPERONE	0.79	0.92	0.85
HORMONE/GROWTH FACTOR	0.9	0.93	0.92
OXYGEN STORAGE/TRANSPORT	0.68	0.77	0.72
SUGAR BINDING PROTEIN	0.9	0.9	0.9
PHOTOSYNTHESIS	0.89	0.91	0.9
TRANSPORT PROTEIN	0.64	0.92	0.76
MEMBRANE PROTEIN	0.97	0.89	0.93
APOPTOSIS	0.68	0.83	0.75
METAL BINDING PROTEIN	0.95	0.85	0.9
CYTOKINE	0.93	1	0.96
Overall Accuracy			**0.87**

Table 6 shows that for most classes, the Naive Bayes model provides average PRE, recall, and f1 score. Providing a satisfactory average accuracy as well.

Table 7 displays the results of the AdaBoost model. The average PRE, recall, and f1 score for AdaBoost are 0.20, 0.15, and 0.14, respectively, with 0.24 being the total accuracy. This model and the decision tree-based model are identical.

Table 7: Result of the AdaBoost model

	PRE	recall	F1
HYDROLASE	0	0	0
OXIDOREDUCTASE	0	0	0
TRANSFERASE	0	0	0
LYASE	0	0	0
HYDROLASE INHIBITOR/HYDROLASE	0	0	0
IMMUNE SYSTEM	0	0	0
ELECTRON TRANSPORT	0.2	0.97	0.33
ISOMERASE	0	0	0
TOXIN	0.42	0.23	0.29
LIGASE	0.45	0.22	0.3
SIGNALING PROTEIN	0	0	0
VIRUS	0	0	0
OXYGEN TRANSPORT	0	0	0
VIRAL PROTEIN	0	0	0
TRANSCRIPTION	0	0	0
IMMUNOGLOBULIN	0.57	0.09	0.15
DNA BINDING PROTEIN	0.7	0.23	0.35
CHAPERONE	0.79	0.73	0.76
HORMONE/GROWTH FACTOR	0.65	0.73	0.69
OXYGEN STORAGE/TRANSPORT	0	0	0
SUGAR BINDING PROTEIN	1	0.35	0.52
PHOTOSYNTHESIS	0	0	0
TRANSPORT PROTEIN	0	0	0
MEMBRANE PROTEIN	0.1	0	0.01
APOPTOSIS	0.5	0.39	0.44
METAL BINDING PROTEIN	0	0	0
CYTOKINE	0	0	0
Overall Accuracy			**0.24**

Table7 shows that the AdaBoost based model is not giving good result. The accuracy and other measures value are very less. So, we cannot recommend this model for classification and prediction. This is behaving similar to the Decision tree. Next we have model the system using Random Forest. The Random Forest model result is shown in table 8. The Random Forest overall accuracy is 0.91, the average PRE, recall and f1 score is 0.94, 0.86 and 0.89 respectively.

Table 8: Result of the Random Forests model

	PRE	recall	F1
HYDROLASE	0.79	0.9	0.84
OXIDOREDUCTASE	1	0.97	0.99
TRANSFERASE	1	0.62	0.76
LYASE	0.97	0.73	0.83
HYDROLASE INHIBITOR/HYDROLASE	0.96	0.89	0.92
IMMUNE SYSTEM	1	0.89	0.94
ELECTRON TRANSPORT	0.87	0.92	0.89
ISOMERASE	0.59	0.89	0.71
TOXIN	0.88	0.92	0.9
LIGASE	0.95	0.78	0.85
SIGNALING PROTEIN	1	0.93	0.97
VIRUS	1	0.9	0.95
OXYGEN TRANSPORT	0.99	0.96	0.98
VIRAL PROTEIN	0.9	0.69	0.78
TRANSCRIPTION	1	0.65	0.79
IMMUNOGLOBULIN	0.97	0.96	0.96
DNA BINDING PROTEIN	0.92	0.77	0.84
CHAPERONE	0.89	0.92	0.9
HORMONE/GROWTH FACTOR	0.96	0.9	0.93
OXYGEN STORAGE/TRANSPORT	0.96	0.77	0.86
SUGAR BINDING PROTEIN	1	0.9	0.95
PHOTOSYNTHESIS	1	0.89	0.94
TRANSPORT PROTEIN	0.96	0.88	0.91
MEMBRANE PROTEIN	0.94	0.94	0.94
APOPTOSIS	0.9	0.83	0.86
METAL BINDING PROTEIN	1	0.88	0.94
CYTOKINE	0.96	0.96	0.96
Overall Accuracy			**0.91**

Table 8 shows that the Random Forests model gives average PRE; recall and f1 score for most of the classes. Also giving an average acceptable accuracy. Next we have model the system using KNN. The KNN model result is shown in table 9. The overall accuracy of the KNN is 0.79, while the average PRE, recall, and f1-score are 0.91, 0.71, and 0.79, respectively. The KNN based model not giving the acceptable accuracy, recall and f1-score. So, we cannot recommend this model for classification and prediction.

Table 9: The KNN model's output

	PRE	recall	F1
HYDROLASE	0.71	0.71	0.71
OXIDOREDUCTASE	1	0.89	0.94
TRANSFERASE	0.8	0.38	0.52
LYASE	1	0.61	0.76
HYDROLASE INHIBITOR/HYDROLASE	0.95	0.76	0.84
IMMUNE SYSTEM	1	0.82	0.9
ELECTRON TRANSPORT	0.5	0.96	0.66
ISOMERASE	0.86	0.66	0.74
TOXIN	0.88	0.74	0.8
LIGASE	0.84	0.69	0.76
SIGNALING PROTEIN	1	0.77	0.87
VIRUS	1	0.74	0.85
OXYGEN TRANSPORT	0.99	0.87	0.93
VIRAL PROTEIN	0.94	0.58	0.71
TRANSCRIPTION	1	0.4	0.57
IMMUNOGLOBULIN	0.99	0.82	0.9
DNA BINDING PROTEIN	0.54	0.73	0.62
CHAPERONE	0.88	0.63	0.74
HORMONE/GROWTH FACTOR	0.91	0.67	0.77
OXYGEN STORAGE/TRANSPORT	0.95	0.56	0.71
SUGAR BINDING PROTEIN	1	0.81	0.89
PHOTOSYNTHESIS	1	0.72	0.84
TRANSPORT PROTEIN	0.97	0.65	0.78
MEMBRANE PROTEIN	1	0.8	0.88
APOPTOSIS	0.93	0.61	0.74
METAL BINDING PROTEIN	1	0.76	0.86
CYTOKINE	0.96	0.84	0.9
Overall Accuracy			**0.79**

Based on given result in table 3 to table 9, we observed that the Decision tree and Adaboost based models are not suitable for the protein classification and prediction. The KNN based model also giving not acceptable accuracy, recall and f1-score. So, this model also be not suited in this environment. Rest of the models are giving an acceptable accuracy that is 90% or above. To validate the above statements, we have computed the average PRE, recall, f1-score and AUROC for each model. The proposed comparative result is given in table 10 and same has been given in Figure 1.

Table 10: The comparative result of the all proposed model.

Model	PRE	Recall	F1-score	Accuracy	AUROC
SVM	0.94	0.86	0.88	0.90	**0.93**
Naïve Bayes	0.79	0.85	0.81	0.87	0.92
Adaboost	0.20	0.15	0.14	0.24	0.56
KNN	0.91	0.71	0.79	0.79	0.85
Random forest	**0.94**	**0.86**	**0.89**	**0.91**	**0.93**
Decision tree	0.11	0.08	0.07	0.22	0.52
Logistic Regression	0.94	0.86	0.88	0.90	**0.93**

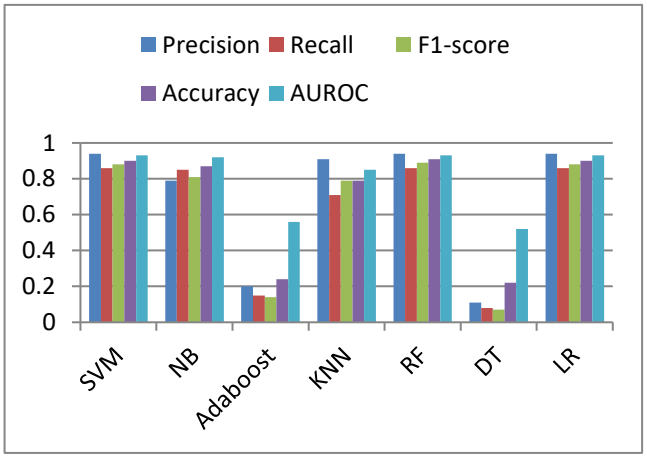

Figure 1: Comparison of the effectiveness of each model.

From the table 10 and figure 1, we can see that the AUROC value of SVM, Random Forests and Logistic regression is equal. So, based on only AUROC, we cannot come to the final outcome. If we consider, accuracy and AUROC together, then the Random forests based models is giving better result than the other models. Based on the above discussion, we can conclude that the Random Forest based models can be used for protein sequence classification and prediction.

5. CONCLUSIONS

Much research has been done in the field of computational biology to determine the most meaningful and accurate functions for predicting protein function. This section describes different types of classification techniques such as SVM, Logistic Regression, Decision Trees, Random Forests, Adaboost, Naive Bayes, and KNN. Experimental analysis of large numbers of sample data from proteins in the human category was performed for classification and prediction and state that the Random forests based model

can be used. The random forests based model is giving highest accuracy i.e. 91% and the AUROC score is 0.93.

REFERENCES

[1] Asaly S, Gottlieb L-A, Inbar N, Reuveni Y. Using Support Vector Machine (SVM) with GPS Ionospheric TEC Estimations to Potentially Predict Earthquake Events. *Remote Sensing*. 2022; 14(12):2822.

[2] Xu G, Zhang M, Zhu H and Xu J: A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene 604: 33-40, 2017.

[3] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, U.K., Cambridge:Cambridge Univ. Press, 2000.

[4] S. Afifi, H. GholamHosseini, R. Sinha, A system on chip for melanoma detection using FPGA-based SVM classifier, Microprocess. Microsyst. 65 (Mar2019) 57–68.

[5] Samb et al., 2012. Samb, M. L., Camara, F., Ndiaye, S., Slimani, Y., and Esseghir, M. A. (2012). A novel rfe-svm-based feature selection approach for classification. International Journal of Advanced Science and Technology, 43(1):27–36.

[6] Mohammadreza Sheykhmousa, Masoud Mahdianpari, Hamid Ghanbari, Fariba Mohammadimanesh, Pedram Ghamisi, Saeid Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol.13, pp.6308-6325, 2020.

[7] Mahadevan A, Mathioudakis M. Certifiable Unlearning Pipelines for Logistic Regression: An Experimental Study. *Machine Learning and Knowledge Extraction*. 2022; 4(3):591-620.

[8] Wright, R. E. (1995). Logistic regression. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 217–244). American Psychological Association.

[9] hristian Genest, Aristidis K. Nikoloulopoulos, Louis-Paul Rivest, Mathieu Fortin, Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas, Brazilian Journal of Probability and Statistics, 10.1214/11-BJPS165, **27**, 3, (2013).

[10] C. E. Davis, J. E. Hyde, S. I. Bangdiwala, and J. J. Nelson. An example of dependencies among variables in a conditional logistic regression. In S. H. Moolgavkar and R. L. Prentice, editors, *Modern Statistical Methods in Chronic Disease Epi*, pages 140–147. Wiley, New York, 1986.

[11] Mohebbanaaz, Kumari, L.V.R. & Sai, Y.P. Classification of ECG beats using optimized decision tree and adaptive boosted optimized decision tree. *SIViP* **16**, 695–703 (2022).

[12] Emmanuel Aldovino, Yi Wang, Applying Decision Tree in Fast Fashion Process, Advanced Manufacturing and Automation XI, 10.1007/978-981-19-0572-8_83, (653-660), (2022).

[13] Tran Quang-Huy, Phuc Thinh Doan, Nguyen Thi Hoang Yen, Duc-Tan Tran, Shear wave imaging and classification using extended Kalman filter and decision tree algorithm, Mathematical Biosciences and Engineering, 10.3934/mbe.2021378, 18, 6, (7631-7647), (2021).

[14] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., AND STONE, C. J. 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

[15] P. Argentiero, R. Chin and P. Beaudet, "An automated approach to the design of decision tree classifiers", *IEEE Trans. Patt. Analy. Mach. Intell.*, vol. PAMI-4, pp. 51-57, 1982.

[16] Palimkar, P., Shaw, R.N., Ghosh, A. (2022). Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach. In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol 218. Springer, Singapore.

[17] Breiman L Random forests—random features. 1999 Technical Report 567, Statistics Department, University of California, Berkeley, ftp://ftp.stat.berkeley.edu/pub/users/breiman.

[18] C. Luo, Z. Wang, S. Wang, J. Zhang and J. Yu, "Locating facial landmarks using probabilistic random forest", *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2324-2328, Dec. 2015.

[19]. M. Oshiro, P. S. Perez and J. A. Baranauskas, "How many trees in a random forest?", *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.*, pp. 154-168, 2012.

[20] E. Sevin¸c, An empowered AdaBoost algorithm implementation: A COVID-19 dataset study, Computers & Industrial Engineering, vol.165, DOI: 10.1016/j.cie.2021.107912, 2022.

[21] R. E. Schapire and Y. Singer, "Improved Boosting Algorithms Using Confidence-ratedPredictions", *Machine Learning*, vol. 37, no. 3, pp. 297-336, December 1999.

[22] Sanjay Saxena, Suraj Shama, "Brain Tumor Segmentation by Texture Feature Extraction with the Parallel Implementation of Fuzzy C-Means using CUDA on GPU", *2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pp.580-585, 2018.

[23] T. G. Dietterich, "An Experimental Comparison of Three Methods for ConstructingEnsembles of Decision Trees: Bagging Boosting and Randomization", *MachineLearning*, vol. 40, no. 2, pp. 139-157, August 2000.

[24] V. Anitha and S. Murugavalli, "Brain tumor classification using two-tier classifier with adaptive segmentation techniques", *IET computer vision*, no. 10, pp. 9-17, 2015.

[25] Sethi, J.K., Mittal, M. Efficient weighted naive bayes classifiers to predict air quality index. *Earth Sci Inform* **15**, 541–552 (2022).

[26] Aggarwal, CC, Zhai, C. A survey of text classification algorithms. In: Aggarwal, CC, Zhai, C (ed.) Mining text data. Berlin: Springer, 2012, pp. 163–222.

[27] John, GH, Langley, P. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the 11th international conference on uncertainty in artificial intelligence, San Francisco, CA, 18–22 August 1995, pp. 338–345. San Mateo, CA: Morgan Kaufmann.

[28] McCallum, A, Nigam, K. A comparison of event models for Naïve Bayes text classification. In: ICML/AAAI-98 workshop on learning for text categorization, Madison, WI, 26–27 July 1998, pp. 41–48. Palo Alto, CA: AAAI.

[29] Rennie, JDM, Shih, L, Teevan, J. Tackling the poor assumptions of naive Bayes text classifiers. In: Proceedings of the 20th international conference on machine learning, Washington, DC, 21–24 August 2003.

[30] Yunhua Zhou, Peiju Liu, and Xipeng Qiu. 2022. KNN-Contrastive Learning for Out-of-Domain Intent Classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5129–5141, Dublin, Ireland. Association for Computational Linguistics.

[31] Wang, H.: Nearest Neighbours without k: A Classification Formalism based on Probability, technical report, Faculty of Informatics, University of Ulster, N.Ireland, UK (2002).

[32] Kubat, M., Jr., M.: Voting Nearest-Neighbour Subclassifiers. In: Proceedings of the 17th International Conference on Machine Learning, ICML 2000, pp. 503–510, Stanford, CA, June 29-July 2 (2000).

[33] Enrico Blanzieri and Farid Melgani. 2008. Nearest neighbor classification of remote sensing images with the maximal margin principle. IEEE Trans. Geosci. Remote Sens. 46, 6 (2008), 1804--1811.

[34] S. Zhang, M. Zong, K. Sun, Y. Liu and D. Cheng, "Efficient kNN algorithm based on graph sparse reconstruction", *Proc. ADMA*, pp. 356-369, 2014.

[35] E. Blanzieri and F. Melgani, "Nearest neighbor classification of remote sensing images with the maximal margin principle", *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1804-1811, Jun. 2008.

[36] Gupta, Chhote Lal Prasad, Anand Bihari, and Sudhakar Tripathi. "Protein Classification using Machine Learning and Statistical Techniques: A Comparative Analysis." arXiv preprint arXiv:1901.06152 (2019).

[37] Poux et al., 2017. Poux, S., Arighi, C. N., Magrane, M., Bateman, A., Wei, C.-H., Lu, Z., Boutet, E., Bye-A-Jee, H., Famiglietti, M. L., Roechert, B., et al. (2017). On expert curation and scalability: Uniprotkb/swiss-prot as a case study. Bioinformatics, 33(21):3454–3460.

[38]Rentzsch and Orengo, 2013. Rentzsch, R. and Orengo, C. A. (2013). Protein function prediction using domain families. In BMC bioinformatics, volume 14, page S5. BioMed Central.

[39]J.Ren,G.Yang. Facial Expression Recognition based on Gabor Transform and AdaBoost Algorithm. Chinese Journal of MicroComputer Information. 23(3-1):290-292,2007.

[40]Schwartz et al., 2011. Schwartz, C. E., Sprangers, M. A., Oort, F. J., Ahmed, S., Bode, R., Li, Y., and Vollmer, T. (2011). Response shift in patients with multiple sclerosis: an application of three statistical techniques. Quality of Life Research, 20(10):1561–1572.

[41]Narkhede, Sarang. "Understanding auc-roc curve." *Towards Data Science* 26.1 (2018): 220-227.

[42]Rohmat Indra Borman, Riduwan Napianto, Nurhasan Nugroho, Donaya Pasha, Yuri Rahmanto, Yohanes Egi Pratama Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants", 2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), pp.46-50, 2021.

[43]Tiwari and Srivastava, 2014. Tiwari, A. K. and Srivastava, R. (2014). A survey of computational intelligence techniques in protein function prediction. International journal of proteomics, 2014.

[44] E. Karunapala, Protein function prediction using machine learning Ph.D. dissertation.

[45]S. Das, I. Sillitoe, D. Lee, J.G. Lees, N.L. Dawson, J. Ward, and C.A. Orengo, "CATH FunFHMMer web server: protein functional annotations using functional family assignments", Nucleic Acids Res., vol. 43, no. W1, 2015.W148-53.

[46]A. Garg, and G.P. Raghava, "A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search", In Silico Biol. (Gedrukt), vol. 8, no. 2, pp. 129-140, 2008.

[47]https://www.kaggle.com/code/abharg16/predicting-protein-classification/data.

[48]Gupta, Chhote LP, Anand Bihari, and Sudhakar Tripathi. "Protein Classification Using Machine Learning and Statistical Techniques." Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science) 14.5 (2021): 1616-1632.

[49]Gupta, C. P., Bihari, A., & Tripathi, S. (2019). Human Protein Sequence Classification using Machine Learning and Statistical Classification Techniques. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 2, pp. 3591–3599). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijrte.b3224.078219

[50]Gupta, C. L. P., Bihari, A., & Tripathi, S. (2019). Rat Protein's Enzyme Class Classification Using Machine Learning. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 6, pp. 655–663). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijeat.f8098.088619

[51]Uoc, N. Q., Duong, N. T., Son, L. A., & Thanh, B. D. A (2022) Novel Automatic Detecting System for Cucumber Disease Based on the Convolution Neural Network Algorithm. GMSARN International Journal 16 (2022) 295-301

[52]KS, A. K., Sarita, K., Kumar, S., Saket, R. K., & Swami, A.(2022) Machine Learning-based Approach for Prevention of COVID-19 using Steam Vaporizer GMSARN International Journal 16 (2022) 399-404.