Protein Secondary Structure Prediction based on CNN and Machine Learning Algorithms

Romana Rahman Ema[1], Dr. Md. Nasim Adnan[3]
Assistant Professor, Dept. of Computer Science and Engineering
Jashore University of Science and Technology, Jashore, Bangladesh[1, 3]

Mt. Akhi Khatun[2]
(B.Sc Engg.), Dept. of Computer Science and Engineering
Jashore University of Science and Technology
Jashore, Bangladesh

Dr. Syed Md. Galib[5]
Professor, Dept. of Computer Science and Engineering
Jashore University of Science and Technology
Jashore, Bangladesh

Sk. Shalauddin Kabir[4]
Lecturer, Dept. of Computer Science and Engineering
Jashore University of Science and Technology
Jashore, Bangladesh

Dr. Md. Alam Hossain[6]*
Associate Professor, Dept. of Computer Science and Engineering, Jashore University of Science and Technology
Jashore, Bangladesh

Abstract—**One of the most important topics in computational biology is protein secondary structure prediction. Primary, secondary, tertiary, and quaternary structure are the four levels of complexity that can be used to characterize the entire structure of a protein that are totally ordered by the amino acid sequences. The polypeptide backbone of a protein's local configuration is referred to as a secondary structure. In this paper, three prediction algorithms have been proposed which will predict the protein secondary structure based on machine learning. These prediction methods have been improved by the model structure of convolutional neural networks (CNN). The Rectified Linear Units (ReLU) has been used as the activation function. The 2D CNN has been trained with machine learning algorithms, including Support Vector Machine, Naive Bays and Random Forest. The SVM is used to correctly classify the unseen data. Naïve Bays (NB) and Random Forest (RF) are also applied to solve the prediction problems for not only classification problems but also regression problems. The 2D CNN, hybrid of 2D CNN -SVM, CNN-RF and CNN-NB have been proposed in this experiment. These different methods are implemented with the RS126, 25PDB and CB513 dataset. Further, all prediction Q3 accuracy is compared and improved with their datasets.**

Keywords—Protein Secondary Structure Prediction (PSSP); Support Vector Machine (SVM); Naive Bays (NB); Random Forest (RF); Convolutional Neural network (CNN)

I. Introduction

Proteins are the building blocks of amino acid sequences [1] [2]. Generally, there are four types of protein structure's which are primary, secondary, tertiary, and quadratic structure. Secondary structure are the building blocks of the macromolecule structure [1]. Secondary structures can be divided into two categories- the regular and the irregular secondary structures. The regular secondary structure has two types, including α -helices (H) and β-sheet (E) and the irregular secondary structure has more types, including tight turns, Random coils, Bulges, etc. By using only their basic structure, the PSSP method is a series of bioinformatics technique aimed at predicting the secondary structure of protein sequences or residues[2] [3]. In molecular biology, the most essential and crucial problems are the prediction of the protein secondary structures using machine learning approaches [3]. This proposed work aims to predict the protein secondary structures and come up with a highly accurate solution that would be easily solved by computational biology. The purpose of PSSP is also to categorize the pattern of residues in amino acid sequences as a-helix, B-strand, or coil. To discover the secondary structure of proteins, the researchers must examine hydrogen bonding patterns and geometric limitations, as well as employ the DSSP tool [4]. The prediction of PSSP is done using a variety of machine learning and deep learning techniques[5] [6].

Researchers used a variety of strategies to predict the protein secondary structures in the early years [5][7][8]. Furthermore, compared to the prior years, the prediction accuracy has been improved in this paper. The understanding of protein folding mechanisms are frequently regarded as a crucial objective that will help structural biologists unravel the puzzling connection among the protein sequence, structure, and function. The scientific work will be aided by the ability to estimate protein folding speeds without the requirement for actual experimental study. In this study, the secondary structure prediction is merely enhanced, as it would be challenging to predict the tertiary structure of proteins in the absence of homology.

Convolutional neural network is a form of artificial neural network that is generally used in deep learning for image processing [1][2]. The hidden units of the CNN are frequently the same dimensions or size as the processed data [1]. The data is convolved by the hidden units, which then stored the information from the data. Each hidden unit's information will

*Corresponding Author.

be recorded as a feature map [1] [7]. The number of feature maps that produced equal to the number of hidden units are employed. After that, the pooling stage is performed on the existing feature maps, collecting dense information from them [1][9].

Support vector machine is a machine learning algorithm based on a supervised technique that has been used for different types of classification problems [1]. If a model of SVM is given that the sets of labeled schooling data for each section, it will be enabled to classify the new data. It has been used to analyze the data for not only classification problems but also regression problems. It has also been used to categorize the unlabeled data. The main aim of this is to search hyperplane that can be utilized as a decision surface to close the gap between two classes [1] [10].

Naïve bays is also a supervised machine learning model based on Bays Theorem [3]. Typically, this is a classification technique that predicted with an estimation of independence. Naïve bays model is easy to create and especially applicable to solve the classification problems and huge complex data sets. For that, it can make as a fast prediction [11].

Popular machine learning algorithm, random forest belongs to the supervised approach. This type of algorithm is also applied for both classification as well as regression problems. In RF, each individual tree has been spread as a class prediction and the most voted class can be estimated for prediction and it predicts the final output. It improves the prediction accuracy [12].

In this paper, a prediction method has been represented for the secondary structure of proteins based on CNN and machine learning techniques. In addition, this paper is exploded into six sections. The Section II discusses about the related works. The Section III represents the preliminaries for the proposed models. The Section IV introduces and describes the proposed techniques smoothly. The Section V shows the simulation results. Finally, conclusion is presented in Section VI.

II. Related Work

Vincent Michael Sutanto et al. [1] introduced a hybrid 1-D CNN and SVM for the prediction of protein secondary structures. They fine-tuned CNN and then changed it, therefore, instead of giving an orthogonal label as output, model feature map production. By doing this step, they projected the data into higher levels. They aimed to absorb high dimensional space of SVM as classifier. The modified CNN model created the 3-D array properties as an output and therefore, a conversion must be applied. The modified CNN models used feature maps for the training of SVM models.

Shangxin Xie et al. [10] introduced a new algorithm formed on the increased fuzzy support vector machine (FSVM) for the prediction of secondary structure of proteins. They applied the different classification rules for the prediction. They used this model to improve the fuzzy membership value and prediction accuracy. They implemented this model to derive the approximate optimal division hyperplane in the feature space.

Masood Zamani et al. [13] proposed the evolutionary-based computation method of protein secondary structure classifiers. They evaluated the performance prediction by using the amino acid sequence with the help of the clustering technique. K-means clustering technique is applied to reduce the dimension of the classifier's inputs on sequence component. They also presented PSS classifier stand on genetic programming technique. They evaluated this approach to improve the performance prediction.

Pooja Jain et al. [14] explored the structural classification of protein with the help of supervised learning technique. Firstly, they chose domains for learning the structure classification. This technique considered two types of domains, known and unknown structural classification. They assigned known domains into unknown domains for classification of protein.

Ying Xu et al. [15] proposed a prediction method of secondary structure based on convolutional neural network (CNN) and random forest. After each convolutional layer, Rectified Linear Units (ReLU) activation layer was used to solve the gradient disappearance problems. They used deep CNN to extract the protein features from amino acid sequence. Here, the fully connected layer and SoftMax layer were used to predict the three types of secondary structure (C, H, E).

III. Preliminaries

A. Protein

The three-dimensional arrangement of atoms in an amino acid-chain molecule is known as protein structure [16] [17]. Proteins are polymers – especially polypeptides – made up of amino acid sequences, which are the polymer of monomers [16]. Proteins are involved in a wide range of biological processes from accelerating chemical reactions to construct the architecture of all living things [18]. Despite their diverse activities, all proteins are built up of the same twenty amino acid building blocks. The folding of the protein into its unique final form and function is determined by how these twenty amino acids are organized [8][19][20].

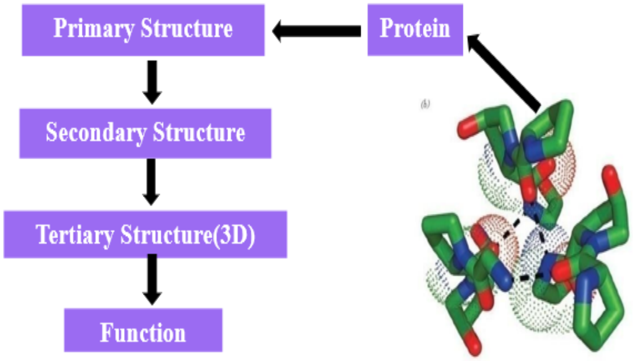


Fig. 1. Protein Structure.

Fig. 1 shows the structure of protein and its function.

B. Primary Structure

The sequence of amino acids in a polypeptide chain is known as primary structure [18]. The primary structure of a protein is described the starting from the amino-terminal (N) end to the carboxyl-terminal (C) end. Ribosomes are the most frequent organelles in cells that perform protein production [21]. Peptides can also be manufactured in the lab. Primary structures of proteins can be sequenced directly or deduced from DNA sequences [18] [20].

C. Secondary Structure

The pattern of hydrogen bonds between the amino hydrogen and carboxyl oxygen atoms in the peptide backbone is technically characterized as a secondary structure [17] [18]. The hydrophobic side chains are those facing inward. A third of every amino acid has hydrophobic properties [17]. There are hydrogen bonds between the two distinct zones (carbonyl & amino groups). Beta sheets are made up of 5–10 amino acids in each region.

Secondary structure, the next level of protein structure, refers to the local folded structures formed inside a polypeptide as a result of interactions between backbone atoms [22]. (It does not involve R group atoms; the backbone only refers to the polypeptide chain away from the R groups.) Helixes and pleated sheets are the two most common types of secondary structures [18][21][23].

D. Tertiary Structure

The term "tertiary structure" refers to the overall three-dimensional structure of a polypeptide. There are various types of polypeptide chains in it. It interacts with R-groups. Polar R-groups are capable of forming hydrogen bonds. Dipole-dipole interactions and hydrophobic interactions play key roles in the three-dimensional structure [8].

E. Convolutional Neural Network (CNN)

The most popular kind of neural network is the convolutional neural network which is used to solve the image processing problems [1]. CNN is divided into two sections: the hidden layers, also known as the feature extraction section, and the classification section. A series of convolutions and pooling operations are carried out by the hidden layers [7]. To create a map, convolution is applied to the input data using a filter or kernel [7][21].The classification parts assign a probability for the object on the image being what the algorithm is predicted. Multilayer perceptron is regularized variants of CNNs. Normally, these networks are fully connected, which means that every neuron in one layer is connected to every neuron in the layer below. It is susceptible to data overfitting because of the network's "full connectivity." By utilizing the hierarchical structure in the data and assembling patterns of increasing complexity using smaller and simpler patterns imprinted in their filters, CNNs develop a new method of regularization [20]. The bottom end of the connectivity and complexity spectrum is where CNNs fall due to this.

Fig. 2 shows the architecture of CNN, including with three input layers, hidden layer section 1 and section 2 (both containing 4 layers) and finally output layer.

F. Convolutional Layer

It is the principal components of CNN. It has a collection of filters or kernels, which are parameters that must be learned during training. The convolution 2D layer was employed in this investigation. The size of the filters is typically smaller than the size of the image. An activation map is created when each filter is applied to an image. This has the advantage of reducing parameter usage and allowing the convolution kernel to extract features more effectively. The preprocessed protein data with a size of (13, 20) is used as the input data. To obtain the output, the convolution kernel of 3× 3 is twisted in steps of 1 from left to right, from top to bottom, from the upper left corner of the data with the input size of (13, 20). If padding is not used, the output size is (13 – 3 + 1)/ (20 – 3+ 1) [7]. The number of parameters is minimized by the CNN feature and considerably the training speed has been enhanced through it. The data is convolved in the same way by 128 convolution kernels, and each convolution kernel extracts features automatically. Different convolution kernels extract picture edge information, shading information, contours, and other image features automatically in the image domain. The model can extract 128 features automatically in theory. Supposing that the convolution kernel’s width is ck and the height is dk. The 2D convolutional equation is:

$$m_{i,j}=\begin{bmatrix} n_{i,j} & n_{i+1,j} \cdots\cdot & n_{i+ck,j} \\ n_{i,j+1} & n_{i+1,j+1} \cdots & n_{i+ck,j+1} \\ n_{i,j+ck} & n_{i+1,j+dk} \cdots & n_{i+ck,j+dk} \end{bmatrix} [7][15] \quad (1)$$

The ReLU is used as the activation function within the model. The ReLU function has the following expression:

$$F(y)=\max(0, y) [7] \quad (2)$$

G. SoftMax Layer

It is mostly used in artificial neural network as the activation function in the output layer. This is used as classification problems where more than two class labels require class membership [9]. The SoftMax layer has been also used in proposed model as the activation function [7].

H. MaxPooling Layer

It is a procedure for retrieving features from the convolutional layer that reduces their dimensionality [18]. It offers the advantage of reducing the size of the featured image while maintaining the number of feature maps, allowing for data processing after the convolutional layer is output. Simultaneously, significant feature information is stored, reducing the model's computation complexity and increasing calculation speed. Median and average pooling are two other pooling algorithms. The kernel size of MaxPooling layer is 2× 2 [9].

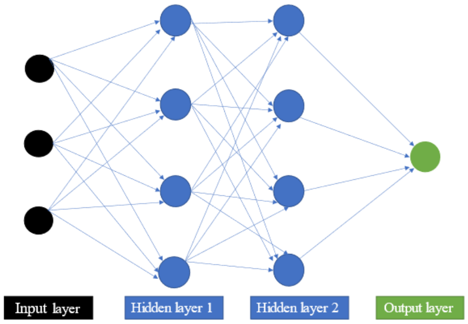


Fig. 2. Convolutional Neural Network.

I. Flatten Layer

It is frequently used to convert multidimensional information into one-dimensional input during the transition from the convolution layer to the fully linked layer. Flattening is the process of converting data into a one-dimensional array for use in the layer below. We flatten the convolutional layer output to create a single, extensive feature vector. The final classification model, sometimes referred to as a fully-connected layer, is connected to it [3].

TABLE I. POOLED FEATURE MAP AND FLATTENING STEPS

1	1	0
4	2	1
0	2	1

Pooled
Feature Map

Flattening →

1
1
0
4
2
1
0
2
1

Table I shows that the pooled feature map and flattening steps. The flattening process is used to make a single feature vector and get the next layer.

J. Support Vector Machine (SVM)

It is a type of supervised learning model that analyzes and linearizes data for classification as well as regression in machine learning [18]. SVM algorithm classifies data that is linearly divisible. If it is not linearly divisible, we have to follow the kernel strategy to build decision [24]. This decision program is called support vector. Support Vector uses a subset of training points that makes it efficient in memory [25]. Support vector machine works in following steps:

- Firstly, we have to import the dataset.
- Need to analyze the data.
- Then we have to preprocess the data.
- Split up the dataset.
- Sort out the dataset into training and testing sections.
- Train and test the support vector machine algorithm.
- Build some predictions.
- Finally, compute the accuracy.

Fig. 3 shows the architecture of SVM. In this figure, a hyperplane with maximized margin is created which refers to the distance between the data points.

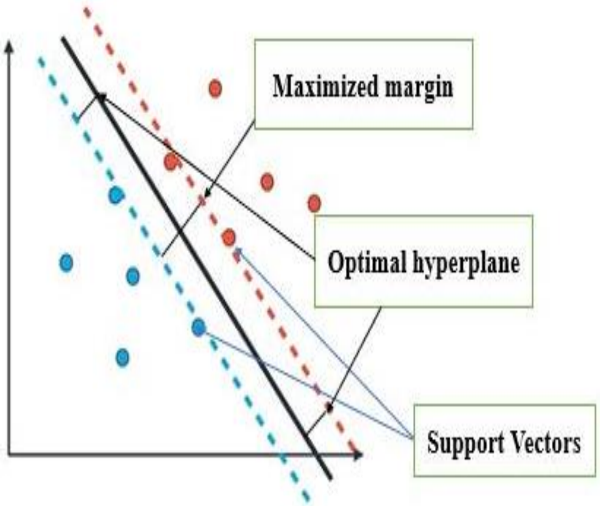


Fig. 3. The Architecture of Support Vector Machine.

K. Naïve Bays (NB)

Naïve bays is a general machine learning strategy that is supervised using Bayes Theorem [3]. It is a statistical classification technique. The basic innocent way is to assume that each feature makes a distinct and equal one for which it is called a “naïve”. This is a strong idea for genuine and unrealistic data. This type of conjecture is indicated as class conditional independence. Naïve Bays algorithm works in following steps such as:

Fig. 4 shows the naïve bays classifier. It is one kind of linear classifier. It’s named “Naïve” because it assumes that the feature of datasets is mutually independent.

Naïve Bays algorithm works in following steps such as:

- Divide the class.
- Sum up the dataset.
- Again, sum up data according to class.
- Then calculate density function according to Gaussian Probability.
- Finally compute class probabilities.

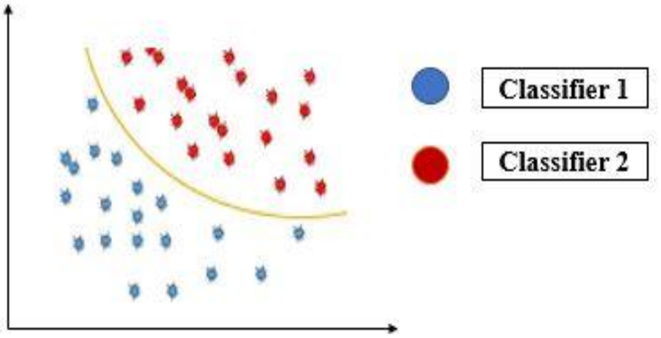


Fig. 4. The Architecture of Naïve Bays Classifier.

L. Random Forest (RF)

Generally, random forest is a simple, supervised learning technique. It is used for not only classification troubles but also regression troubles like support vector algorithm. It can be learned how to classify data randomly. It extracts several samples from the original sample using the bootstrap resampling approach that works by training a large number of decision trees [23]. The RF output is the class chosen by the majority of trees for the classification task [18].

The random forest algorithm can be explained in following steps:

- Import the dataset.
- Choose the random samples.
- Compute the vote for the predictive result. It can be used "Mode" for "Classification" troubles, and also can be used "Mean" for "Regression" troubles.
- Ultimately, choose the peak voted value for predictive result. And it is the most precious final prediction.

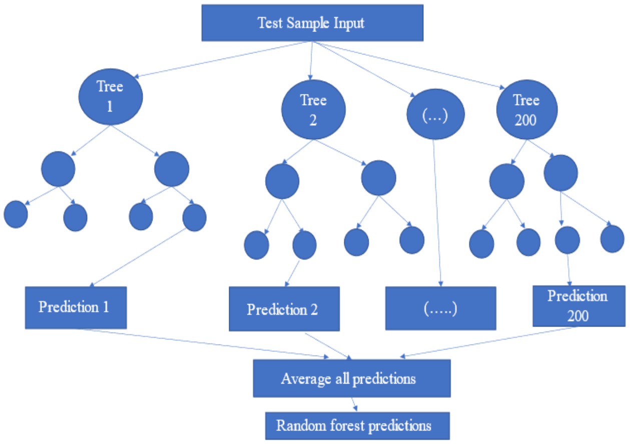


Fig. 5. The Architecture of Random Forest Classifiers.

Fig. 5 shows the random forest classifier. In this figure, the test sample input is tree root. Random forests generate decision trees from randomly chosen data samples, obtain predictions from each tree, then vote on the best option.

IV. Proposed Method

In this research, the PSSP method has been proposed based on CNN and machine learning algorithms. 2-D CNN and machine learning algorithms i.e., RF, SVM and NB have been used in this paper. Here, the Rectified Linear Unit (ReLU) has been used as an activation function in 2D-CNN. It has been used after each convolutional layer to solve the gradient disappearance problems. This paper used deep CNN to extract the protein features from the amino acid sequence. The fully connected layer, MaxPooling, flatten and SoftMax layer have been used to predict the three types of secondary structure (C, H, E). Further, the pooling function have been used in these layers. The 2D CNN has been combined and trained with SVM, RF, NB. Secondly, the datasets have been trained and tested with SVM, RF and NB. The 2D CNN, the hybrid of 2D CNN-SVM, CNN-RF and CNN-NB have been implemented. Finally, the prediction accuracy has been calculated.

Fig. 6 shows the flow diagram of protein secondary structure prediction. Firstly, the input (amino acid sequence) has been processed. Secondly, the Con2D layer has been used. Here, the filter has been changed with the process. The softmax layer has been used as an activation function. Also, the maxpooling and other layers have been used in this process. Further, the 2D-CNN has been combined and trained algorithm such as SVM, RF and NB algorithm. Finally, the secondary structure sequence such as Helix (H), Strand (E) and Coil (C) have been predicted from the amino acid sequence.

Secondary structure prediction datasets.

1) RS126 dataset.
2) CB513 dataset.
3) 25PDB dataset.

The first and best-known dataset for the PSSP is the RS126 dataset. Electric Sander as well as Rost came up with the concept [26]. It is one of the best non-homologous datasets for predicting the structure of protein [26]. This dataset is applied for the prediction of protein secondary structure. Its maximum carrying capacity is 23,347 residues, and its typical protein sequence length is 185. RS126 is composed of 47% random coil, 21% -sheet, and 32% -helices (H).[25] Furthermore, the CB513 and 25PDB datasets have also been used. The 2-D structure is predicted using the CB513 dataset. This is a crucial dataset that is excellent for enhancing algorithms and forecasting secondary structures. The CB513 dataset was created by Cuff and Barton and contained 513 sequences and 84,107 residues [24]. This kind of dataset is used in this study to enhance the performance of protein secondary structure prediction. One of the biggest datasets is this. Furthermore, it is utilized to classify data and fold the protein structure. Additionally, the 25PDB dataset was included. The PDB dataset's accuracy is better than some other types of datasets in each of the datasets. All protein datasets have been collected from Research Collaboratory for Structural Bioinformation (RCSB).

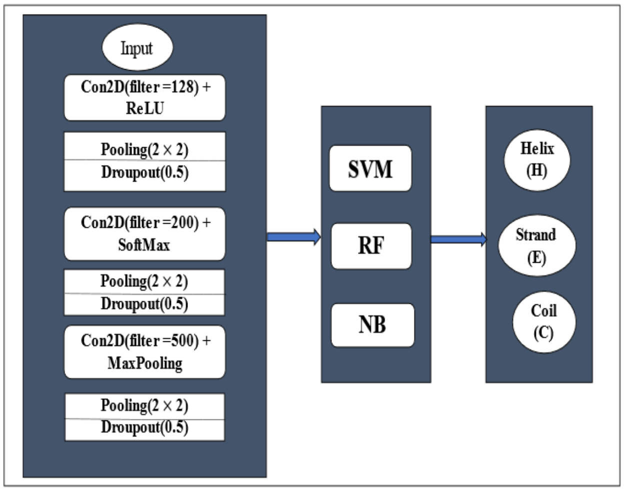


Fig. 6. The Flowchart of Proposed Model of Protein Secondary Structure Prediction.

TABLE II. PROTEIN SECONDARY STRUCTURES OF 8 AND 3 CLASSES.[22]

8class symbols	8 class names	3class symbols	3class names
H	α-helix	H	Helix
L	Loop/Irregular	C	Coil/Loop
T	B-Turn	C	Coil/Loop
S	Bend	C	Coil/Loop
G	3_{10}-helix	H	Helix
B	B-bridge	E	Sheet
I	π-helix	H	Helix

Table II shows the protein secondary structure of eight and three classes. According to DSSP (Define Secondary Structure of Proteins), eight class secondary structures are converted to three class of secondary structures of proteins.[19].

Pseudocode 1: CNN [1]

```
Input: Amino acid sequence
Step 1: Sliding window process
Step 2: ESF → Extract Shadow Feature
Step 3: Normalize by the ESF
Step 4: Feature, data size 13*20
Step 5: Repeat
Step 6: Forward Propagation
CD → Convolution2D (ESF)
MP → MaxPooling 2D (CD)
FC → Fully Connected (MP)
Class label → SoftMax (FC)
Class label → Dense(D)
Class label → Flatten (F)
Step 7: Backward Propagation
Conduct backward propagation with Adam;
Step 8: Use the trained algorithm with SVM, RF and NB.
```

Pseudocode 2: SVM

```
Input: M and n filled with schooling labeled data, $\beta \Leftarrow 0$ or $\beta \Leftarrow$ partially trained support vector machine.
Step 1: P ⇐some value (20 for example)
Step 2: repeat
Step 3: for all $\{m_i,n_i\}$, $\{m_j,n_j\}$ do
Step 4: optimize $\beta_i$ and $\beta_{i.}$
Step 5: End for
Step 6: until no changes in $\beta$ or other element constraint criteria met.
Ensure: Just contain the support vectors ($\beta_i > 0$).
Output: Accuracy prediction.
```

Pseudocode 3: RF [12] [15]

```
Step 1: Randomly select the "K" attribute from the total "P" attributes. K < P
Step 2: Calculate the node "r" in the "K" properties by applying the largest partition points.
Step 3: Need to use the greatest split node for splitting the node into daughter mode.
Step 4: Steps are repeated 1 through 3 until the "F" number of node is reached.
Step 5: The forest is created by repeating steps 1 to 4 for "N" number times for to build "N" number of trees.
```

Pseudocode 4: NB [3]

```
Input: Training dataset D,
P= ($p_{1,}$ $p_2$, $p_{3...,}p_n$) // value of the predictor variable in testing dataset.
Output: A category of testing dataset.
Step 1: Firstly, read the training dataset D.
Step 2: Secondly, compute the mean deviation and standard deviation of the predictor variables in every group;
Step 3: Repeat. Compute the probability of $p_i$ applying the gauss density equation in every group;
Until the probability of predictor variables ($p_1$, $p_2$, $p_3$………,$p_n$) has been computed.
Step 4: Compute the likelihood for every class or group;
Step 6: Finally, get the best likelihood.
```

V. RESULT ANALYSIS AND DISCUSSION

The Q3 approach was employed in this paper to assess the algorithm's effectiveness. Q3 is used to represent the number of residues and is computed by dividing the number of accurately predicted protein residues by the total number of residues in a known protein's secondary structure sequence.

$$Q_3 = \frac{Q_C+Q_H+Q_E}{Q} \tag{3}$$

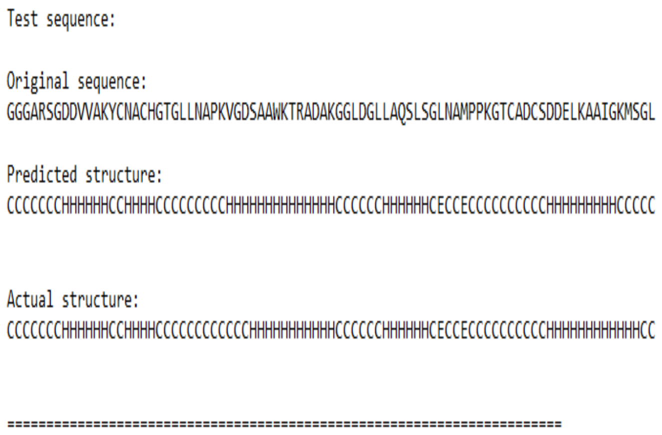
Where Q_C, Q_H and Q_E are the number of accurately predicted protein structural class of C, H and E. The total number of amino acids are denoted by Q [7].

$$Q_j = \frac{Q_k}{Q} \tag{4}$$

Where,

Q_k represents the total number of amino acid residues. $K \in \{C, H, E\}$. [7].

Q_C, Q_H and Q_E are used to evaluate the experimental result in this research. The Eq. 4 can be used to calculate the values of Q_C, Q_H and Q_E.



```
Test sequence:

Original sequence:
GGGARSGDDVVAKYCNACHGTGLLNAPKVGDSAAWKTRADAKGGLDGLLAQSLSGLNAMPPKGTCADCSDDELKAAIGKMSGL

Predicted structure:
CCCCCCCHHHHHHCCHHHHCCCCCCCCHHHHHHHHHHHHHHCCCCCCHHHHHHCECCECCCCCCCCCCHHHHHHHHHCCCCC


Actual structure:
CCCCCCCHHHHHHCCHHHHCCCCCCCCCCCHHHHHHHHHHHCCCCCCHHHHHHCECCECCCCCCCCCCHHHHHHHHHHHHCC


=====================================================================
```

Fig. 7. Prediction from Original Sequence.

Fig. 7 shows the actual structure predicted from the predicted structure and predicted structure predicted from original primary sequence. [Test sequence collected from RS126 dataset].

Table III, IV, V shows the Q_C , Q_H , Q_E and Q_3 accuracy based on RS126, CB513 and PDB25 datasets.

Where Q_C , Q_H and Q_E are used to accurately predict the protein structural class of C, H and E. In these tables, CNN, CNN-SVM, CNN-RF and CNN-Bays methods have been used for the prediction of protein secondary structure. In Q_3 all prediction accuracy, the CNN-SVM method has been achieved the highest prediction accuracy. These achievements are 82.34%, 84.32%, 83.76% based on RS126, CB513 and 25PDB datasets.

Table VI and VII shows the comparison between the previous model and the proposed model. Here, 25PDB and CB513 datasets have been used for the protein secondary structure prediction. It is seen that the previous work of datasets based on the model of CNN-SVM, CNN-RF and CNN-Bays are not satisfactory. The proposed model has improved the performance of protein secondary structure prediction in Q3 accuracy. The CNN-RF, CNN-Bays, CNN and CNN-SVM have been achieved the highest Q3 accuracy of 81.73%, 79.35%, 80.57% and 84.32 based on 25PDB and CB513 datasets which are 2.34%, 2.45%, 1.48%, 2.83% higher than those datasets. It is seen that the proposed model has been achieved the highest prediction Q_3 accuracy than the previous model.

TABLE III. THE ACCURACY OF RS126 DATASET

Method	Accuracy (%)			
	Q_C	Q_H	Q_E	Q_3
CNN	79.24	76.35	77.43	80.09
CNN-SVM	81.27	78.32	80.35	82.34
CNN-RF	79.76	75.03	78.45	80.24
CNN-Bays	78.20	73.23	75.76	79.27

TABLE IV. THE ACCURACY OF CB513 DATASET

Method	Accuracy (%)			
	Q_C	Q_H	Q_E	Q_3
CNN	79.54	73.65	77.83	80.35
CNN-SVM	83.08	78.67	81.07	84.32
CNN-RF	78.80	74.09	76.53	79.24
CNN-NB	78.36	74.26	77.34	79.80

TABLE V. TABLE V. THE ACCURACY OF 25PDB DATASET

Method	Accuracy (%)			
	Q_C	Q_H	Q_E	Q_3
CNN	80.32	74.30	76.04	80.57
CNN-SVM	82.37	75.31	81.09	83.76
CNN-RF	79.05	72.53	74.23	81.73
CNN-NB	77.45	70.63	73.83	79.35

TABLE VI. THE PERFORMANCE COMPARISON RESULTS BETWEEN THE PREVIOUS AND PROPOSED MODEL

		Q_3 Accuracy (%)	
Dataset	Method	Previous prediction	Proposed prediction
CB513	CNN-SVM	81.49 [1]	84.32
	CNN-RF	-	79.24
	CNN-NB	-	79.80
	CNN	-	80.35

TABLE VII. THE PERFORMANCE COMPARISON RESULTS BETWEEN THE PREVIOUS AND PROPOSED MODEL

Dataset	Method	Q_3 Accuracy (%)	
		Previous prediction	Proposed prediction
25PDB	CNN	79.09 [15]	80.57
	CNN-RF	79.39 [15]	81.73
	CNN-Bays	76.90 [3]	79.35
	CNN-SVM	-	83.76

VI. CONCLUSION AND FUTURE WORK

The issue of protein structure prediction must be resolved in the realm of bioinformatics. The protein secondary structure has been utilized in this paper. To understand the protein sequence using CNN and machine learning algorithms was the initial step for solving the PSSP problem. In this work, three identical and separate datasets—25PDB, RS126, and CB513 are used in this study. These datasets are sufficient for solving the problems of prediction. The ReLU layer is used as the activation function (CNN). Further, MaxPooling, SoftMax, dense and flatten layer have been used. The CNN has been integrated with RF, SVM, and NB in this study. The proteins have been categorized using these techniques. These hybrid techniques have successfully overcome the gradient disappearance problems by retaining the significance of original features data to the maximum extent possible. It is demonstrated that the proposed model has been enabled to capture the long-range interdependencies between the sequence residues. The proposed models reached in high accuracy. The CNN-SVM model has achieved the highest Q3 accuracy of 82.35%, 84.32% and 83.76% on the RS126, CB513 and 25PDB datasets than the previous work. Besides, CNN, CNN-RF and CNN-NB has achieved the highest prediction accuracy. The secondary structure of protein takes part in an important role in protein function and folding. It identifies the similar function where protein sequence varies (only ~50% remote homologies may be identified based on sequence). The proposed method can experimentally perform better than other previous work and this work could be easily understandable by researchers for solving the problem of protein secondary structure prediction. Further, it can help to explain the disease (the effect of mutations and design drugs). As a future scope of research, it would be possible to propose the three-dimensional protein structure prediction using deep learning algorithms.

Acknowledgment

This research did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] V. M. Sutanto, Z. I. Sukma, and A. Afiahayati, “Predicting Secondary Structure of Protein Using Hybrid of Convolutional Neural Network and Support Vector Machine,” Int. J. Intell. Eng. Syst., vol. 14, no. 1,pp.232–243,2020, doi: 10.22266/IJIES2021.0228.23.

[2] R. R. Ema, A. Khatun, M. A. Hossain, M. R. Akhond, N. Hossain, and M. Y. Arafat, “Protein Secondary Structure Prediction using Hybrid Recurrent Neural Networks,” J. Comput. Sci., vol. 18, no. 7, pp. 599–611, 2022, doi: 10.3844/jcssp.2022.599.611.

[3] Y. Liu, Y. Chen, and J. Cheng, “Feature extraction of protein secondary structure using 2D convolutional neural network,” Proc. - 2016 9th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2016, pp. 1771–1775, 2017, doi: 10.1109/CISP-BMEI.2016.7853004.

[4] G. B. De Oliveira and H. Pedrini, “Ensemble of Template-Free and Template-Based Classifiers for Protein Secondary Structure Prediction,” 2021.

[5] W. Wardah, M. G. M. Khan, A. Sharma, and M. A. Rashid, “Protein secondary structure prediction using neural networks and deep learning: A review,” Comput. Biol. Chem., vol. 81, no. December 2018, pp. 1–8, 2019, doi: 10.1016/j.compbiolchem.2019.107093.

[6] W. Yang, K. Wang, and W. Zuo, “A fast and efficient nearest neighbor method for protein secondary structure prediction,” 2011 3rd Int. Conf. Adv. Comput. Control. ICACC 2011, no. Icacc, pp. 224–227, 2011, doi: 10.1109/ICACC.2011.6016402.

[7] Y. Zhao, H. Zhang, and Y. Liu, “Protein secondary structure prediction based on generative confrontation and convolutional neural network,” IEEE Access, vol. 8,pp.199171–199178,2020,doi: 10.1109/ACCESS.2020.3035208.

[8] P. Kumar, S. Bankapur, and N. Patil, “An enhanced protein secondary structure prediction using deep learning framework on hybrid profile based features,” Appl. Soft Comput. J., vol. 86, p. 105926, 2020, doi: 10.1016/j.asoc.2019.105926.

[9] J. Cheng, Y. Liu, and Y. Ma, “Protein secondary structure prediction based on integration of CNN and LSTM model,” J. Vis. Commun. Image Represent., vol. 71, p. 102844, 2020, doi: 10.1016/j.jvcir.2020.102844.

[10] S. Xie, Z. Li, and H. Hu, “Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization,” Gene, vol. 642, no. September 2017, pp. 74–83, 2018, doi: 10.1016/j.gene.2017.11.005.

[11] M. O. F. Science, “Prediction of Protein Secondary Structure using Binary Classification Trees , Naive Bayes Classifiers and the Logistic Regression Classifier,” no. January, 2015.

[12] C. Kathuria, D. Mehrotra, and N. K. Misra, “Predicting the protein structure using random forest approach,” Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1654–1662, 2018, doi: 10.1016/j.procs.2018.05.134.

[13] M. Zamani and S. C. Kremer, “Protein secondary structure prediction using an evolutionary computation method and clustering,” 2015 IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2015, 2015, doi: 10.1109/CIBCB.2015.7300327.

[14] P. Jain, J. M. Garibaldi, and J. D. Hirst, “Supervised machine learning algorithms for protein structure classification,” Comput. Biol. Chem., vol. 33, no. 3, pp. 216–223,2009,doi: 10.1016/j.compbiolchem.2009.04.004.

[15] Y. Xu and J. Cheng, Protein secondary structure prediction using cnn and random forest, vol. 1254 CCIS. 2020. doi: 10.1007/978-981-15-8101-4_25.

[16] M. H. Zangooei and S. Jalili, “Protein secondary structure prediction using DWKF based on SVR-NSGAII,” Neurocomputing, vol. 94, pp. 87–101, 2012, doi: 10.1016/j.neucom.2012.04.015.

[17] W. Pirovano and J. Heringa, “Protein secondary structure prediction.,” Methods Mol. Biol., vol. 609, pp. 327–348, 2010, doi: 10.1007/978-1-60327-241-4_19.

[18] M. Patel and H. Shah, “Protein secondary structure prediction using support vector machines (SVMs),” Proc. - 2013 Int. Conf. Mach. Intell. Res. Adv. ICMIRA 2013,pp.594–598,2014,doi: 10.1109/ICMIRA.2013.124.

[19] P. Kountouris et al., “A comparative study on filtering protein secondary structure prediction,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. 9, no. 3, pp. 731–739, 2012, doi: 10.1109/TCBB.2012.22.

[20] S. Long and P. Tian, “Protein secondary structure prediction with context convolutional neural network,” RSC Adv., vol. 9, no. 66, pp. 38391–38396, 2019, doi: 10.1039/c9ra05218f.

[21] A. K. Mandle, P. Jain, and S. K. Shrivastava, “P Rotein S Tructure P Rediction U Sing,” vol. 3, no. 1, pp. 67–78, 2012.

[22] Y. Guo, B. Wang, W. Li, and B. Yang, “Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks,” J. Bioinform. Comput. Biol.,vol.16,no.5,2018,doi:10.1142/S021972001850021X.

[23] Z.Aydin,“Combining Classifiers for Protein Secondary Structure Prediction,” pp. 6–10, 2017, doi: 10.1109/CICN.2017.9.

[24] Q. Jiang, X. Jin, S. J. Lee, and S. Yao, “Protein secondary structure prediction: A survey of the state of the art,” J. Mol. Graph. Model., vol. 76, pp. 379–402, 2017, doi: 10.1016/j.jmgm.2017.07.015.

[25] Y. F. Chin, R. Hassan, and M. S. Mohamad, “Optimized local protein structure with support vector machine to predict protein secondary structure,” Commun. Comput. Inf. Sci., vol. 295 CCIS, no. January 2012, pp. 333–342, 2012, doi: 10.1007/978-3-642-32826-8_34.

[26] B. Rost and C. Sander, “Prediction of protein secondary structure at better than 70% accuracy,” Journal of Molecular Biology, vol. 232, no. 2. pp. 584–599, 1993. doi: 10.1006/jmbi.1993.1413.