IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2023, accepted 25 May 2023, date of publication 5 June 2023, date of current version 13 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3282702

RESEARCH ARTICLE

Multi-S3P: Protein Secondary Structure Prediction With Specialized Multi-Network and Self-Attention-Based Deep Learning Model

M. M. MOHAMED MUFASSIRIN[1,2], M. A. HAKIM NEWTON[3,4], JULIA RAHMAN[1], AND ABDUL SATTAR[1,3]

[1]School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia
[2]Department of Computer Science, South Eastern University of Sri Lanka, Oluvil 32360, Sri Lanka
[3]Institute for Integrated and Intelligent Systems, Griffith University, Nathan, QLD 4111, Australia
[4]School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia

Corresponding author: M. M. Mohamed Mufassirin (m.mufassirin@griffithuni.edu.au)

This work was supported in part by the Australian Research Council under Grant DP180102727, and in part by the Accelerating Higher Education Expansion and Development (AHEAD) Operations Project of Sri Lanka.

ABSTRACT Protein structure prediction (PSP) is a vital challenge in bioinformatics, structural biology and drug discovery. Protein secondary structure (SS) prediction is critical since three-dimensional (3D) structures are primarily made up of secondary structures. With the advancement of deep learning approaches, SS classification accuracy has been significantly improved. Many existing methods use an ensemble of complex neural networks to improve SS prediction. Because of the high dimensionality of the hyperparameter space, deep neural networks with complex architectures are typically challenging to train effectively. Also, predicting secondary structures in the boundary regions between different types of SS is challenging. This study presents Multi-S3P, which employs bidirectional Long-Short-Term-Memory (BILSTM) and Convolutional Neural Networks (CNN) with a self-attention mechanism to improve the secondary structure prediction using an effective training strategy to capture the unique characteristics of each type of secondary structure and combine them more effectively. The ensemble of CNN and BILSTM can learn both contextual information and long-range interactions between the residues. In addition, using a self-attention mechanism allows the model to focus on the most important features for improving performance. We used the SPOT-1D dataset for the training and validation of our model using a set of four input features derived from amino acid sequences. Further, the model was tested on four popular independent test datasets and compared with various state-of-the-art predictors. The presented results show that Multi-S3P outperformed the other methods in terms of Q3, Q8 accuracy and other performance metrics, achieving the highest Q3 accuracy of 87.57% and a Q8 accuracy of 77.56% on the TEST2016 test set. More importantly, Multi-S3P demonstrates high performance in SS boundary regions. Our experiment also demonstrates that the combination of different input features and a multi-network-based training strategy significantly improved the performance.

INDEX TERMS Deep learning, convolutional neural network, protein structure prediction, protein secondary structure, recurrent neural network.

I. INTRODUCTION

Proteins are crucial for living organisms because of their diverse functions, such as functioning as a catalyst in cell metabolism, constructing antibodies and generating cell architecture and live tissues. The functions of a protein are determined by its native 3D structure, which has the lowest free energy. Misfolded proteins have the potential to cause acute illnesses in living organisms. If a drug molecule could dock on a disease protein, its function may be inhibited.

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

As a result, understanding protein structures is critical in medicine, structural bioinformatics, drug design, and other related domains [1], [2]. It typically takes a long time and is expensive to determine protein structure experimentally using techniques like X-ray crystallography, Cryo-EM, and NMR. Numerous computational prediction techniques have been developed to address these issues [3], [4]. By gaining outstanding results during the 14^{th} Critical Assessment of Structure Prediction (CASP14) competition in 2020, AlphaFold2 [5] showed the full potential of deep-learning approaches for the prediction of protein structures from amino acid sequences [6]. The pre-calculated AlphaFold database [7] provides 3D models for over a million distinct proteins as of March 2022, significantly enhancing our knowledge of protein structure and function.

Despite the fact that the novel deep learning techniques have substantially increased the accuracy of protein structure prediction [8], many difficulties still exist because of the complexity of protein 3D structures. The massive computational demands of running the AlphaFold model, both in terms of processing power and runtime, are indeed challenging requirements. As a result, it is still necessary for prediction algorithms that can predict protein structure quickly and accurately. Researchers generally divide this problem into a number of manageable sub-problems, such as the secondary structure prediction [9], [10], backbone angle prediction [11], [12], [13], distance map prediction [14], [15], [16] and contact map prediction [17], in order to make it easier to understand.

When Pauling and Corey suggested sheet and helical conformations (structures) of protein backbones based on hydrogen bonding patterns in 1951 [18], it marked the beginning of the secondary structure (SS) prediction. The secondary structures of a protein refer to the local, repetitive arrangements of its amino acid residues into specific conformations, such as alpha-helices or beta-sheets. These confirmations, in turn, determine the overall folding and stability of the protein, as well as its interactions with other molecules. The local structure of the polypeptide backbone is described at a coarser scale by the secondary structures.

Secondary structure prediction, despite its long history, remains an active area of research because 3D structures are primarily made up of secondary structures. When proteins fold, secondary structures are generated initially, followed by the formation of 3D native conformations. The task of predicting a protein's 3D structure using only its primary amino acid sequence is a challenging one. However, simplifying the prediction process by utilizing the basic definitions of secondary structures can significantly aid in achieving this goal [19]. If the secondary structure of a protein can be built correctly, it can be used to predict numerous structural features necessary for 3D structure prediction. Secondary structures can provide valuable insights into a protein's activity, functions, and relationships [20].

Protein secondary structures can be classified into either Q3 (3-state) or Q8 (8-state) [21]. The 3-state main types of secondary structures are helix (H), strand (E), and coil (C), with helix and strand structures being the most prevalent in nature, according to the various hydrogen bonding modes [18]. A more detailed description of secondary structures was put forth later in 1983. The previous 3-states (Q3) are expanded to 8-states (Q8) in the new classification determined by the DSSP algorithm [21].

The advancement of practical machine learning algorithms and the discovery of novel features have played a significant role in protein SS prediction over the last few decades [22], [23], [24]. To identify SS in proteins, early efforts used statistical propensities of particular amino acids derived from known structures [25]. The integration of sequence evolutionary profile features derived from multiple sequence alignment (MSA), such as position-specific scoring matrices (PSSM) [26], [27], resulted in subsequent improvements.

In addition to the PSSM profile, the Hidden Markov model (HMM) feature obtained from HHblits [28] was employed for predicting protein structure and properties [24]. Some studies used Atchley's factors to determine the similarity of amino acid types [24]. Also, seven physicochemical properties (7PCP) [29] have been used in many secondary structure prediction methods [9], [10], [30], [31], [32].

This paper presents Multi-S3P, which utilizes three different deep-learning techniques and four input features (PSSM, HMM, 7PCP and PSP19) in order to achieve a better classification accuracy of protein SS. Initially, our approach involved constructing an ensemble network consisting of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BILSTM) networks. The utilization of this hybrid architecture allowed for the learning of short-range contextual information as well as long-range interactions. To effectively train our models to classify different types of protein secondary structures, we employed three specialized distinct combinations of CNN and BILSTM networks with a self-attention mechanism to focus on the most informative parts of the feature vectors and make a final prediction for both Q3 and Q8 classification. Furthermore, the proposed method used the same training and validation datasets as the state-of-the-art method SPOT-1D [32]. We conducted our experiments on four independent test sets (TEST2016, TEST2018, CASP12 and CASP13) to evaluate and facilitate a fair comparison of performance with other methods.

The main contribution of this work includes a hybrid deep learning architecture with specialized multi-network-based training to effectively process both local and global interactions between amino acids in making accurate Q3 and Q8 secondary structure predictions using a set of four input features. The use of multi-network-based training techniques for different classes of SS has significantly contributed to improvement in performance. It is evident from the results presented in the work that the proposed Multi-S3P outperformed the

other methods in terms of Q3 and Q8 accuracy and other performance metrics, achieving the highest Q3 accuracy of 87.57% and a Q8 accuracy of 77.56% on TEST2016 test set. These results demonstrate the effectiveness of the proposed method and its potential to enhance secondary structure prediction performance. Our experiment also demonstrates that the combination of different input features significantly impacts performance. Further, the significance of our model's performance lies in its ability to effectively predict SS in the boundary regions, which is a critical and challenging task in protein secondary structure prediction.

The paper is structured as follows: Section II presents the related work in the protein secondary structure prediction area. Section III provides an overview of the basic concepts related to protein structure prediction and protein SS representation. The proposed method, along with the data and features used, is described in Section IV. Section V presents our experimental results and associated discussions. Finally, in Section VI, we summarise our work and provide concluding remarks.

II. RELATED WORKS

In recent years, we found significant progress in protein secondary structure prediction through machine learning and novel deep learning techniques. Many early methods for predicting secondary structures used shallow neural networks [33], [34], Bayesian analysis [35], and information theory [36]. To boost the prediction performance, many researchers employed modified or ensemble neural network architectures. DeepCNF [37] proposed a deep CNN model combined with a Conditional Neural field (CNF). SPIDER3 [31] and NetSurfP-2.0 [38] employed bidirectional recurrent neural networks (BIRNN) to capture long-range amino acids interactions. MUFOLD-SS [19], [39] used deep residual inception networks to learn the global and local interactions between residues. DNSS2 [24] integrated six one-dimensional deep neural networks, such as convolutional, residual, recurrent, fractal memory and inception networks, to predict the Q3 and Q8 secondary structure. SPOT-1D [32] used an ensemble of neural network models consisting of 2D Bidirectional Residual LSTM Networks (2D-BRLSTM) and Residual Convolutional Neural Networks (ResNet). The model SAINT [9] incorporated a self-attention mechanism and Deep3I to enhance the Q8 prediction accuracy. OPUS-TASS [10] recently proposed an architecture based on Transformer along with CNN and LSTM to capture the interactions between the two residues in order to enhance the SS prediction accuracy. A recent method called DLBLS_SS [40] utilized the BILSTM network and temporal convolutional networks to construct the model. In addition, some deep learning-based methods employed huge datasets like ProteinNet [41] to train the model in order to improve the prediction accuracy [42]. However, these models consume a large training time and huge computational power.

Certain predictors, such as PiPred [43], specialized in predicting only the π-helices structures. Some recent predictors, such as NetSurfP-3.0 [44] and SPOT-1D-LM [45] employed the sequence embedding generated by pre-trained protein language models like ESM-1b [46] and ProtTrans [47] to improve the runtime of secondary structure prediction dramatically. However, the prediction performance of these models in terms of accuracy was not improved significantly compared to its predecessor. DML_SS [48] utilized a deep centroid model for protein SS prediction using a lightweight network with multi-branch topology based on deep metric learning. Table 1 shows the summary of the different deep learning architectures used in existing notable SS prediction models.

Overall, the Q3 Protein SS prediction has reached 87-89% accuracy, which is close to its theoretical limit [49], [50]. However, taking current protein structure databases into account, a new study claimed that the theoretical limit of Q3 SS prediction could be extended to 90-92%. The upper limit of Q8 for eight-state measurements is 84-86% [50], [51]. This indicates that there is still a gap in accuracy to be filled.

III. PRELIMINARIES

This section describes the basics of protein structure prediction and protein secondary structure representation for computational PSP.

A. PROTEIN STRUCTURE PREDICTION

Proteins are amino acid sequences. Regular proteins generally have 20 amino acid classes or types. The amino acid types vary in size, structure, charge, shape, hydrogen-bonding capacity, reactivity and hydrophobicity. However, not every protein may include all the 20 types of amino acids. Furthermore, every amino acid can exist in a protein in any position, subject to stoichiometric limitations [52]. Constraints in PSP are also called *restraints*. Different single and 3-letter codes identify amino acid types. As a result, a protein can be characterized primarily by a sequence of those single and 3-letter codes. Table 2 illustrates 1-letter amino acid codes, as well as a protein's amino acid sequence using 1-letter amino acid codes.

Any two consecutive amino acids in the sequence form a *peptide bond* in the given sequence of amino acids in a protein. The consecutive peptide bonds between adjacent amino acids result in the formation of a *polypeptide chain*. This chain is also referred to as the *backbone* or *main chain* of the protein. A polypeptide chain has distinct beginning and ending terminals [53]. One belongs to the amino group, and the other belongs to the carboxyl group, also known as the *N-terminal* and the *C-terminal*, respectively. Nevertheless, once a peptide bond is formed, the remaining portion of an amino acid in the protein's backbone is referred to as an amino acid residue. The length of a protein, denoted by L, is determined by the total number of residues present in the protein.

TABLE 1. Deep learning architectures used in existing notable SS prediction models (CNN: convolutional neural networks; CNF: conditional neural fields; FC: fully connected layers; IN: inception networks; BILSTM: bidirectional long short-term memory; ResNet: residual neural network; RCNN: recurrent convolutional neural network; CRMN: convolutional residual memory networks; FT: fractal networks).

Methods	Neural Network Architectures
DeepCNF [37]	CNF and CNN
SPIDER3 [31]	Eight BILSTM-based models and FC
MUFOLD-SS [19]	Three IN modules followed by a CNN layer and two FC layers.
NetSurfP-2.0 [38]	Two CNN layers followed by two BILSTM layers and an FC layer
SPOT-1D [32]	Ensemble of nine models and an FC layer; 3 LSTM, 3 LSTM-ResNet, and 3 ResNet-LSTM
SAINT [9]	Ensemble of CNN, IN and Self-Attention
OPUS-TASS [10]	Ensemble of CNN, Transformer and BILSTM with an FC layer
DNSS2 [24]	Six types of networks, including CNN, RCNN, ResNet, CRMN, FT and IN
DLBLS_SS [40]	BILSTM and temporal CNN
DML_SS [48]	Embedding networks consisting of IN, CNN, and FC layers

TABLE 2. 1-letter codes of 20 types of amino acid, as well as the amino acid sequence of protein 5*AON* using the amino acids 1-letter codes.

Amino Acid	Codes	Amino Acid	Codes	Amino Acid	Codes	Amino Acid	Codes
Alanine	A	Glutamine	Q	Leucine	L	Serine	S
Arginine	R	Glutamic Acid	E	Lysine	K	Threonine	T
Asparagine	N	Glycine	G	Methionine	M	Tryptophan	W
Aspartic Acid	D	Histidine	H	Phenylalanine	F	Tyrosine	Y
Cysteine	C	Isoleucine	I	Proline	P	Valine	V

5AON = "EREKRVSNAVEFLLDSRVRRTPTSSKVHFLKSKGLSAEEICEAFTKVG"

Certain bonds in the chemical structure of each amino acid are rotatable in 3D space. The bond serves as the axis of rotation for any rotatable bond; the atom at one end of the bond serves as the basis of rotation, and all other atoms within the same amino acid residue or whole atoms in the other amino acid residues in the given protein are rotated. A protein's 3D structure or *conformation* is formed in this way. It should be noted that the peptide bonds formed by successive amino acids are not rotatable. In this regard, the whole 3D structure of a protein is described as the *tertiary structure*, whereas only the chain of amino acids is referred to as the *primary structure*. Therefore, protein structure prediction involves predicting the protein's 3D structure from its sequence of amino acids, i.e. predicting secondary and tertiary structures from primary structure [5].

B. SECONDARY STRUCTURE REPRESENTATION

The local and repetitive patterns of amino acid residues in a protein are referred to as its secondary structures, which adopt specific conformations like α-helices, β-sheets and coils (or loops). These local structures are known as the 3-state *secondary structures*. Considering variants of these main classes, the 3-states (Q3) are expanded to 8-states (Q8) in the new classification determined by the DSSP algorithm, which includes α-helix (H), π-helix (I), 310 helix (G), β-turn (T), β-strand (E), β-bridge (B), bend (S), and loop or others (C), [21] among which the α-helix and β-strand being the two main structural features [54].

Predicting the Q3 or Q8 classes simply refers to the type of secondary structure of a residue, not its position in relation to other secondary structures. Due to the limited range of *phi* and *psi* angles associated with rigid secondary structures such as helices and sheets, it is possible to produce approximate models of such structures using this information.

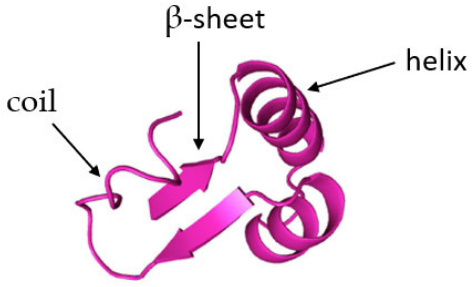


FIGURE 1. Graphical representations of the secondary structures such as helices, sheets, and coils.

Nevertheless, this does not truly aid in the construction of coil-type secondary structures since any angle value is possible. As a graphical Ribbon diagram, Figure 1 shows the various secondary structures, such as β-sheets, helices, and coils.

IV. MATERIALS AND METHODS

The primary goal of this research is to enhance secondary structure prediction by creating more sophisticated learning algorithms and using more informative input features. We have designed and developed a methodological framework to create an efficient deep learning architecture with a novel training approach and to acquire features to enhance secondary structure prediction performance during the process.

A. DATASET

The proposed method uses the same training and validation datasets as the state-of-the-art method SPOT-1D [32] (available at: http://sparks-lab.org/server/spot-1d/). The dataset

was built from PISCES server (CullPDB) [55] curated on February 2017 with the parameters including resolution better than 2.5Å, sequence identity cutoff of 25%, and R-factor <1 [26]. The proteins with over 700 residues and proteins that contain incomplete main-chain atoms were removed to avoid the low-quality structures and the sequence redundancy bias [56]. Finally, we obtained 11,007 proteins, of which 10,024 were used for network training and 983 for model validation.

We conducted our experiments on four independent test sets. Same as SPOT-1D, we used TEST2016 [57] and TEST2018 [32], which contain 1213 proteins and 250 proteins, respectively. In addition, we used two CASP datasets (CASP12 and CASP13) to facilitate a fair comparison to other methods. The CASP12 dataset contains 55 proteins, and the CASP13 dataset consists of 32 proteins [9], [24]. We cleaned the data to remove duplication at a 25% sequence identity cutoff to avoid evaluation bias associated with training data. It is worth noting that all test data contain accurate labels for both 3-state and 8-state secondary structures obtained from the DSSP algorithm [21].

Using the given independent datasets, the performance of the proposed method was tested and compared against various state-of-the-art secondary structure prediction methods, including DeepCNF [37], SPIDER3 [31], RaptorX [11], PSRSM [20], MUFOLD-SS [19], NetSurfP-2.0 [38], Porter 5 [58], SPOT-1D [32], DNSS2 [24], DLBLS-SS [40] and DML_SS [48]. All the approaches were evaluated in terms of Q3, Q8 accuracy and SOV scores on each test dataset.

B. INPUT FEATURES

Homologous proteins, which share a common ancestry and possess comparable sequences of amino acids, generally exhibit analogous secondary structures. This characteristic allows for the classification of homologous proteins as members of the same family by applying a relevant cutoff in MSA analysis [59]. After that, the approximate family structure can be predicted. The MSA appears to provide significantly more information about the structure than a single sequence [60]. Our input features consist of such evolutionary profiles, which include 20 features obtained from the PSSM profile [27], and 30 features obtained from the HMM profile. In addition, we also employed 7PCP of each amino acid and 19 features (PSP19) from OPUS-DOSP [61].

Each PSSM profile was generated by three iterations of PSI-BLAST [26] with default parameters based on UniRef90 database [62] updated in December 2019. We obtained a $L \times 20$ dimension feature vector for each protein from this, where L is the amino acid sequence length. The 20 dimensions indicate the probabilities of residue substitution per sequence residue. Each HMM profile was generated by HHBlits [63] based on Uniclust30 database [64] with default parameters. The HMM profile is an $L \times 30$ dimension feature for each protein, where L is the amino acid sequence length. The 7PCP includes hydrophobicity, Normalized Van der Waal's volume, polarity, polarisability and Normalized frequency of alpha-helix and obtained from [29] and [65].

The PSP19 attribute was obtained from a study by [61], which categorizes 20 amino acid residues into 19 solid-body blocks based on their local structures. Thus, each residue has a 19-dimensional binary vector; if a solid-body block is present in the residue, the corresponding position in the vector is set to 1. Otherwise, it is set to 0. In total, we combined PSSM, HMM, 7PCP and PSP19 to form a final feature vector of $L \times 76$, where L is the length of the primary amino acid sequence.

C. NEURAL NETWORK ARCHITECTURE

The main task of deep learning is to extract either local or non-local interactions from input features using various neural network architectures. According to literature, the Convolutional Neural Network (CNN) effectively captures short-term interaction of features [19], [66]. At the same time, the Recurrent Neural Network (RNN) or Long Short-term Memory (LSTM) [67] can be used for capturing long-range dependencies. To improve the accuracy of protein secondary structure prediction, we developed a novel ensemble approach that utilizes a combination of CNNs, Bidirectional LSTM (BILSTM) networks, and a fully connected (FC) layer component with a self-attention mechanism. This hybrid architecture is designed to capture short-range contextual information and long-range interactions effectively.

1) BASE MODEL

We proposed a base model using CNN, BILSTM and FC components, which was then used in our Multi-S3P model (we provide descriptions of this model in the subsequent section). The framework of the proposed base model is shown in Figure 2. The framework begins with Normalized input feature sets with 0 mean and standard deviation of 1 in the training data to ensure that the values are scaled to be comparable and fall within a similar range. The Normalized input data then send through two parallel network layers. The first parallel network layer is a CNN component with a sequence of five identical convolutional layers, each with five types of kernels: (11,1), (21,1), (31,1), (41,1), and (51,1), with 32 output channels for each kernel type. The 'same' padding is applied in each layer. The shapes of the kernels are primarily established in the sequence direction to acquire information from neighbouring residues, whereas the size along the feature dimension is restricted to one due to GPU memory constraints. The final output of each layer is generated by concatenating the results of each kernel, and Batch Normalization layers are used between every two layers for regularization. The Rectified Linear Unit (ReLU) [68] was used as the activation function in each layer. Finally, the average pooling is utilized in the channel dimension to maintain consistency between the input and output dimensions of the CNN module.

The second parallel network layer is a bidirectional LSTM module with four BILSTM layers, each with 1024 units. As the activation function, Relu was used. To avoid the overfitting problem, a 25% dropout was imposed in each layer [69]. The end of the network consists of a fully connected (FC) layer and a softmax output layer with 11 nodes. Finally, a Softmax activation function was used to convert the output into probabilities. The 11 output nodes corresponded to the Q3 and Q8 secondary structure classes of the protein.

2) MULTI-NETWORK MODEL (MULTI-S3P)

To facilitate the classification of different classes of protein secondary structures (sheets, helices and coils), we trained three similar networks of our base model consisting of a combination of CNN and BILSTM architectures; each specialized in recognizing either sheets, helices, or coils. We employed this concept used in certain specialized predictors, PiPred [43] that only predicts the π-helices structures and SAP4SS [12], a secondary structure specific protein backbone angles predictor. The architecture of the proposed Multi-S3P is shown in Figure 3. Part (a) in the figure is specialized in recognizing helices (H, I and G) SS structures, which has the same architecture as our base model with five identical convolutional layers and four BILSTM layers. Part (b) in the figure is specific for coils types of structures (T, S and C), which has two parallel layers of CNN and BILSTM. The CNN is a sequence of three identical convolutional layers, each with three types of kernels: (11,1), (21,1), and (31,1), with 32 output channels. The BILSTM module is with two identical layers, each with 1024 units. Similarly, part (c) in the figure is specialized in perceiving sheets types of SS structure, which has two parallel modules with two identical convolutional layers and two BILSTM layers.

As shown in the Figure 3 architecture, the three separate networks (part a, part b and part c) each consist of a dense layer, which produces the output for the specific secondary structure class. The output layer uses a softmax activation function, which produces a probability distribution over the possible secondary structure classes (i.e., H for alpha-helices, E for beta-sheets, and C for coils). Each network has its own output layer with a unique name. The outputs of the three separate networks are concatenated together using the Keras concatenate layer.

The concatenated output from the three separate network modules is passed through a self-attention layer. The self-attention layer is a novel layer that has been shown to be effective in a wide range of natural language processing tasks [70]. It allows the model to focus on different parts of the input sequence and weigh the importance of each part based on the context. The model then applies dropout and Normalization layers to improve training stability. The self-attention output is then passed through a separate dense output layer with 11 units, of which three are for Q3 classification, and eight are for Q8 classification. Finally, predicted Q3 and Q8 secondary structures are evaluated.

D. OUTPUTS

Using the proposed deep learning model, we used a multi-network-based prediction technique for predicting Q3 (3-state), Q8 (8-state) classifications using eleven prediction nodes. We employed the definition of DSSP algorithm [21] for assigning SS class labels to the protein sequence for both the Q3 and Q8 secondary structure. The 8-state classification includes 310 helix (G), π-helix (I), α-helix (H), β-bridge (B), β-strand (E), high curvature loop (S), coil (C), and β-turn (T) states. The Q3 essentially divides eight classes into 3-state labels, which are as follows: helix H (G, H, and I in the Q8 definition), strand E (B and E in the Q8 definition), and coil C (C, S and T in the Q8 definition). When each conformation is predicted separately, the accuracy of the predictions is higher compared to when trying to infer the Q3 conformation from the Q8 conformation [32], [71].

E. IMPLEMENTATION AND TRAINING

We used Keras, the Python deep learning API (http://keras.io), along with Tensorflow (v2.8.0) as a back-end to implement and train our model. The model was trained on a GPU node with "Nvidia RTX A5000", having 24 GB of GPU memory. The models in this study are configured to handle a batch size of 4 proteins, with their weights initialized using Glorot uniform initializer. The training process leverages the Adam optimizer [72] to calculate and update the model's parameters with an initial learning rate of 0.001. To enhance model performance, the learning rate is reduced by half when there is a decrease in the validation set accuracy. This process is repeated up to five times before terminating the training process, typically completed in around 36 epochs. On average, each training took 6 hours. The hyperparameters of models were optimized only on the validation set.

F. PERFORMANCE EVALUATION

The accurate classification of protein secondary structure is a challenging task that poses a multi-class problem. Our primary focus was on utilizing two widely recognized measures, namely accuracy and Segment Overlap measure (SOV), as the key metrics for our model analysis. Accuracy serves as a standard performance metric to evaluate the efficacy of secondary structure prediction models. The percentage accuracy measure is utilized for both Q3 and Q8 classifications to quantify the degree of agreement between predicted and observed secondary structure assignments. Let the Q8 secondary structures $S_8 = \{G, I, H, S, T, C, B, E\}$ and the Q3 secondary structures $S_3 = \{C, H, E\}$. The overall Q8 and Q3 accuracy are defined as follows:

$$Q8(Q3) = \frac{\sum n_s}{N_s} \times 100, s \in S_8(s \in S_3) \quad (1)$$

where N_s is the total number of residues that are of state s and n_s is the total number of correctly predicted residues of state s.

SOV measures the agreement between predicted and observed secondary structures at the segment level.

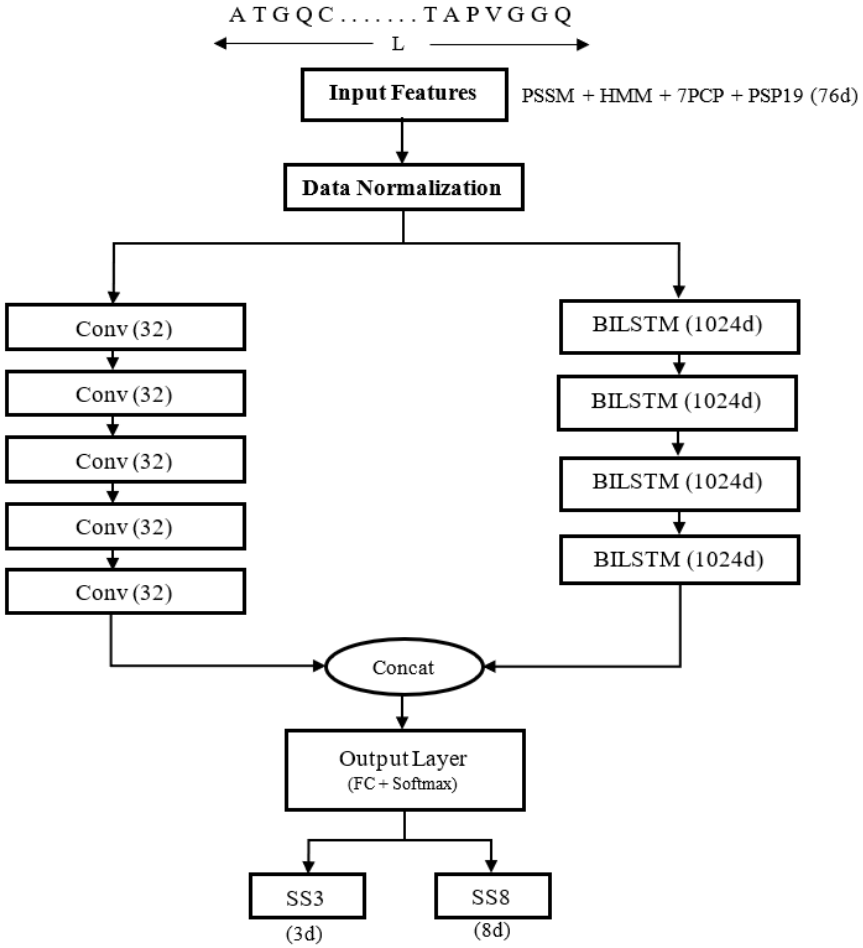


FIGURE 2. Framework of the base-model.

The formula for SOV is defined as shown in Equation (2), at the bottom of the page.

In this formula, SOV represents the segment overlap measure. The calculation involves dividing both the observed and predicted secondary structure sequences into segments that satisfy certain constraints. The variables i and j represent the starting and ending residues of a segment, while s represents the conformational state.

The sets X_s, S_s, T_s, and U_s are defined as follows: X_s is the set of all segments in state s for which there are no overlapping segments in the same state; S_s is the set of all overlapping pairs of observed segments in state s; T_s is the set of all overlapping pairs of predicted segments in state s, and U_s is the union of all sets of overlapping segment pairs for all conformational states derived from the dataset being evaluated.

The lengths of observed and predicted segments are denoted by L_{ij}^{obs} and L_{ij}^{pred}, respectively. The formula calculates the sum of the minimum value between observed and predicted segment lengths for each segment in state s, divided by the sum of the observed and predicted segment lengths, minus the sum of the minimum value between observed and predicted segment lengths for all overlapping segment pairs. The formula can be modified to use a 3-state instead of eight by substituting the appropriate sets and lengths for the 3-state case.

In addition to the model accuracy and SOV, we also looked into the precision, recall, and F1-score to better understand how different techniques performed. In classification problems, high precision indicates that the predicted structures are very accurate and have very few false positives, while high recall indicates that the predicted structures are able to iden-

$$SOV \frac{\sum_{s=1}^{8} \sum_{(i,j)\in X_s} \min(L_{ij}^{\text{obs}}, L_{ij}^{\text{pred}})}{\sum_{s=1}^{8} \sum_{(i,j)\in S_s} L_{ij}^{\text{obs}} + \sum_{s=1}^{8} \sum_{(i,j)\in T_s} L_{ij}^{\text{pred}} - \sum_{s=1}^{8} \sum_{(i,j)\in U_s} \min(L_{ij}^{\text{obs}}, L_{ij}^{\text{pred}})} \quad (2)$$

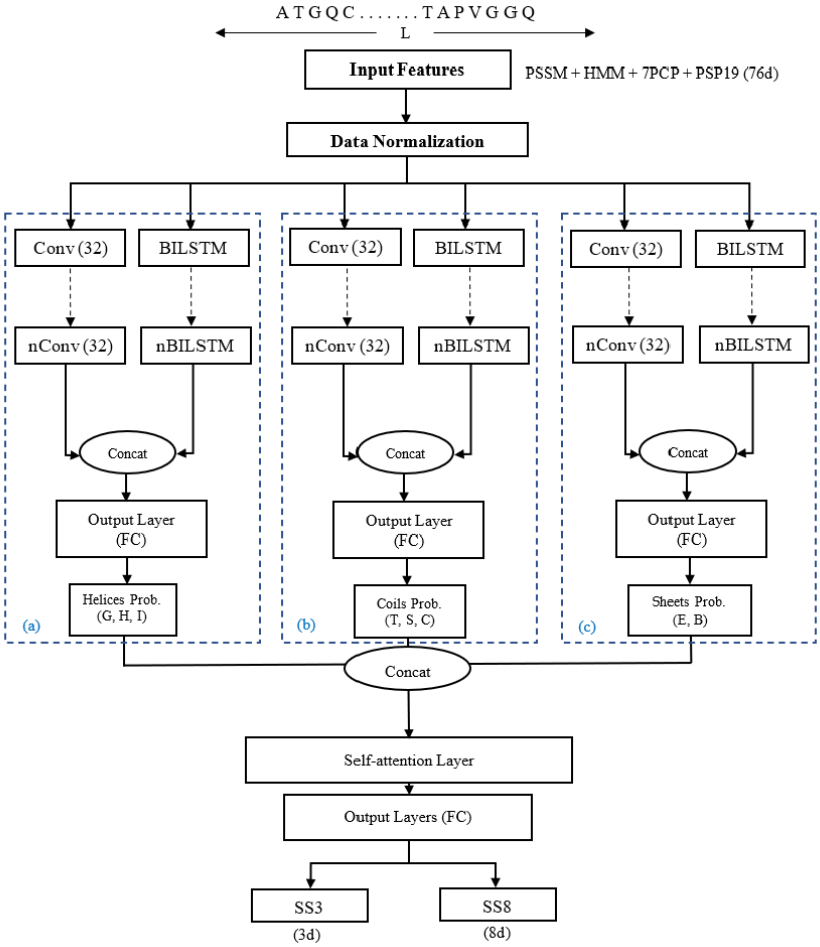


FIGURE 3. The architecture of the proposed Multi-S3P. Part (a) is a specialized network in recognizing helices structures, Part (b) is a specialized network in recognizing coils structures, and Part (c) is a specialized network in classifying sheets structures.

tify most of the true positives. A high F1-score indicates that the predicted structures are both accurate and comprehensive, striking a good balance between precision and recall [73].

V. RESULTS AND DISCUSSION

This section presents detailed experimental findings of the proposed deep neural networks on many commonly used datasets, as well as performance comparisons with existing approaches. After multiple experiments based on a different combination of neural networks and feature sets, we have selected the best model (Multi-S3P). We repeated the training process five times for each model. Then, we calculated the mean prediction accuracy for each model on the validation set. Based on the validation results, we selected our best-performing model (Multi-S3P) to evaluate against benchmark test sets. Table 3 showcases the performance of the proposed two different deep learning models, base-model and Multi-S3P, in predicting Q3 and Q8 secondary structures on the validation set.

Table 4 compares the performance of various secondary structure prediction methods on two datasets, TEST2016 and TEST2018. The methods evaluated include SPIDER3 [31], SPOT-1D [32], PSRSM [20], Porter5 [58], MUFOLD-SS [19], NetSurfP-2.0 [38], SAINT [9], and two methods (Base-model and Multi-S3P) proposed in this study. The table reports the Q3 Accuracy, SOV3, Q8 Accuracy, and SOV8 scores of each method on both datasets.

The results show that Multi-S3P achieves the best performance on both datasets, with the highest Q3 Accuracy, SOV3, Q8 Accuracy, and SOV8 scores. The Q3 accuracy score is 87.57% on TEST2016 and 86.46% on TEST2018. Similarly, the Q8 accuracy score is 77.56% on the TEST2016 and 76.12% on the TEST2018 test set. The performance of the Base-model is comparable to that of SPOT-1D on both datasets. The other methods have varying levels of performance on different aspects of the prediction task.

Table 5 presents a comprehensive comparison of Q3 and Q8 secondary structure prediction performance scores of several protein structure predictors, including state-of-the-art

TABLE 3. The Q3 and Q8 secondary structure prediction performance scores (%) on validation set.

Validation Set				
Methods	**Q3 Accuracy (%)**	**SOV3 (%)**	**Q8 Accuracy (%)**	**SOV8 (%)**
Base-model	87.39	80.53	77.28	75.12
Multi-S3P	87.81	81.03	77.56	75.43

TABLE 4. Comparison of Q3 and Q8 secondary structure prediction performance scores (%) of various predictors on TEST2016 and TEST2018. The superior results are visually emphasized in bold font. The absence of a result is denoted by the symbol "–".

TEST2016				
Methods	**Q3 Accuracy (%)**	**SOV3 (%)**	**Q8 Accuracy (%)**	**SOV8 (%)**
SPIDER3 [31]	84.66	75.62	-	-
MUFOLD-SS [19]	85.97	81.98	76.03	73.67
SPOT-1D [32]	87.16	79.73	77.10	74.98
DML_SS [48]	86.10	**82.72**	76.62	74.06
Base-model	87.18	79.77	77.05	74.65
Multi-S3P (this work)	**87.57**	79.84	**77.56**	**75.29**
TEST2018				
DeepCNF [37]	81.62	66.58	70.43	65.66
PSRSM [20]	81.94	74.22	-	-
SPIDER-3 [31]	83.84	73.89	-	-
Porter5 [58]	84.10	74.04	73.22	70.27
MUFOLD-SS [19]	84.78	77.56	73.66	71.34
NetSurfP-2.0 [38]	85.31	78.58	73.81	71.14
SPOT-1D [32]	86.18	79.00	75.41	73.30
SAINT [9]	-	-	76.09	-
DML_SS [48]	84.83	**80.05**	74.82	72.23
Base-model	86.22	79.07	75.69	73.15
Multi-S3P (this work)	**86.46**	79.48	**76.12**	**73.36**

methods such as DeepCNF [37], SPIDER3 [31], Porter5 [58], MUFOLD-SS [19], SPOT-1D [32], DNSS2 [24], and DLBLS-SS [40], as well as two methods proposed in the paper the Base-model and the Multi-S3P. The table provides a clear representation of the experimental results obtained on the CASP12 and CASP13 datasets, with the Q3, Q8 accuracy and SOV measure for each method. The superior results are visually emphasized in bold font.

It is evident from the results presented in the table that the Multi-S3P outperformed the other methods in terms of Q3 and Q8 accuracy, achieving a Q3 accuracy of 82.54% and a Q8 accuracy of 72.28%. The Base-model, on the other hand, achieved a Q3 accuracy of 82.18% and a Q8 accuracy of 71.95% on the CASP12 test set. Similarly, Multi-S3P achieved a Q3 accuracy of 83.02% and Q8 accuracy of 73.17% on the CASP13 dataset. These results demonstrate the effectiveness of the proposed Multi-S3P and its potential to improve secondary structure prediction performance.

Tables 6 - 8 display the precision, recall, and F1-score of the Q8 class classification obtained by our proposed method as well as three other state-of-the-art methods on the TEST2016 test set, respectively. The outcomes suggest that our Multi-S3P achieved a superior F1-score compared to other methods in five classes (out of Q8 classes), indicating that our proposed method in this study generated more balanced and significant outcomes than other predictors. Specifically, the proposed method performed exceptionally well on the non-ordinary Q8 classes, such as I, G, S, and T [37], outperforming other methods. However, MUFOLD-SS attained a better F1-score for the B class, while SPOT-1D achieved marginally better results for the H and E classes of Q8.

Classes isolated $\beta-$bridges (B), the 310-helix (G), and Bends (S) exhibit poor performance in terms of F1-score compared to other classes. These classes are characterized by greater conformational flexibility, which poses challenges for accurate prediction. Class B are rare and lacks distinct sequence motifs, making it difficult to differentiate them from other structures. The irregular hydrogen bonding pattern of G and their absence of prominent sequence motifs contribute to their challenging prediction. Bends lack well-defined secondary structures. Predicting bends accurately can be challenging due to their diverse structural characteristics and the absence of specific sequence motifs associated with them. Another important reason is that compared to other secondary structure types, these classes are less abundant in protein databases, resulting in relatively fewer instances available for training deep learning models. This limited availability of training data for these classes can impact the model's

TABLE 5. Comparison of Q3 and Q8 secondary structure prediction performance scores (%) of various predictors on CASP12 and CASP13 dataset. The experiment's superior results are visually emphasized in bold font. The absence of a result is denoted by the symbol "–".

Methods	Q3 Accuracy (%)	SOV3 (%)	Q8 Accuracy (%)	SOV8 (%)
CASP12				
DeepCNF [37]	80.35	67.92	67.78	67.53
SPIDER3 [31]	81.84	71.78	-	-
Porter5 [58]	80.58	72.76	68.48	69.91
MUFOLD-SS [19]	79.48	69.26	66.65	65.38
SPOT-1D [32]	82.37	73.36	71.32	69.93
DNSS2 [24]	80.95	71.76	70.82	69.55
DLBLS-SS [40]	81.28	**75.98**	70.54	69.34
Base-model	82.18	73.77	71.95	69.15
Multi-S3P (this work)	**82.54**	74.11	**72.28**	**69.96**
CASP13				
DeepCNF [37]	79.75	68.31	66.49	69.23
SPIDER3 [31]	81.20	73.64	-	-
Porter5 [58]	81.77	73.85	67.71	70.33
MUFOLD-SS [19]	79.88	71.28	66.70	68.06
DNSS2 [24]	81.62	72.19	72.72	72.01
DLBLS-SS [40]	81.39	**77.18**	68.59	69.33
Base-model	82.18	72.64	71.91	71.33
Multi-S3P (this work)	**83.02**	73.90	**73.17**	**72.19**

ability to learn and generalize well, leading to poorer performance [9], [43].

Addressing these challenges requires the development of more sophisticated prediction methods that can capture the nuanced characteristics of these classes. Incorporating additional informative input features can help improve the prediction. Our ongoing research efforts aim to enhance the accuracy and reliability of prediction through advancements in deep learning architectures and the integration of diverse data sources and features.

A. PERFORMANCE IN BOUNDARY REGIONS

Protein secondary structure prediction can be challenging in predicting the boundary regions between different types of SS where one type of SS ends and another type begins. Predictors often assign incorrect secondary structures to residues located in the boundary regions. This is because these boundary regions often exhibit structural ambiguity and can be difficult to distinguish from other secondary structures [2], [22], [74].

We conducted an experiment to evaluate the performance of predictors in these boundary regions of 8-state prediction. Figure 4 shows the process of extracting secondary structure boundaries. For this purpose, we loop through each label character in the actual secondary structures (actual labels) string and remove consecutive duplicates. For each character, we compare them to the next character. If they are different, it means that the boundary between two different secondary structures has been crossed. We then add both characters to the actual list and the corresponding characters from the predicted secondary structures (predicted labels) string to the predicted list. This process effectively extracts the actual and predicted secondary structures in the boundary regions.

Actual_SS = "C E E E E C C C C C T T T T C C C "
Predicted_SS = "C C C E E C C C T T C C C E E E C"

Actual_ Boundary = "C E E C C T T C"
Predicted_ Boundary = "C C E C T C E E"

FIGURE 4. The extraction of secondary structure boundaries from actual and predicted secondary structures.

The extracted actual and predicted secondary structures were used to evaluate the accuracy of the predictor in the boundary regions. The boundary regions are underlined in Figure 4. Table 9 shows the performance of Multi-S3P and two other state-of-the-art methods in the boundary regions on the CASP12 dataset. Our model Multi-S3P shows significant performance in effectively predicting SS in the boundary regions.

B. ABLATION STUDY AND INPUT FEATURE ANALYSIS

Input feature representation plays a major role in protein secondary structure prediction. The input features in this work are represented by a concatenation of PSSM, HMM, 7PCP and PSP19 profiles, all of which transmit evolutionary information of amino acids in protein sequences. After determining the optimum deep learning architecture, it was used to investigate the impact of various combinations of input features. Table 10 shows the Q3 accuracy of different combinations of input features on our validation set. The PSSM profile has greater predictive performance than the HMM profile. When PSSM was paired with HMM, the prediction accuracy increased by about 1.51% for Q3, demonstrating

TABLE 6. Precision of each Q8 class classification obtained by proposed method and other three state-of-the-art predictors on TEST2016 test set. The experiment's superior results are visually emphasized in bold font.

8-state Class	Multi-S3P	MUFOLD-SS	NetSurf-2.0	SPOT-1D
H	0.877	0.868	**0.885**	0.884
B	**0.731**	0.608	0.650	0.671
E	0.844	0.850	0.822	**0.852**
G	**0.578**	0.519	0.536	0.547
I	0.965	0.857	0.044	**1.000**
T	**0.663**	0.631	0.615	0.641
S	**0.631**	0.689	0.579	0.624
C	**0.647**	0.608	0.613	0.631

TABLE 7. Recall of each Q8 class classification obtained by Multi-S3P and other three state-of-the-art predictors on TEST2016 test set. The experiment's superior results are visually emphasized in bold font.

8-state Class	Multi-S3P	MUFOLD-SS	NetSurf-2.0	SPOT-1D
H	**0.946**	0.943	0.933	0.941
B	0.107	**0.115**	0.071	0.097
E	0.885	0.842	**0.903**	0.878
G	**0.387**	0.348	0.334	0.375
I	**0.431**	0.383	0.426	0.128
T	**0.612**	0.586	0.585	0.612
S	**0.357**	0.313	0.278	0.337
C	0.729	0.727	0.704	**0.741**

TABLE 8. F1-score of each Q8 class classification obtained by Multi-S3P and other three state-of-the-art predictors on TEST2016 test set. The experiment's superior results are visually emphasized in bold font.

8-state Class	Multi-S3P	MUFOLD-SS	NetSurf-2.0	SPOT-1D
H	0.910	0.904	0.908	**0.911**
B	0.187	**0.193**	0.126	0.170
E	0.864	0.846	0.861	**0.865**
G	**0.464**	0.417	0.412	0.445
I	**0.596**	0.529	0.079	0.226
T	**0.636**	0.608	0.599	0.626
S	**0.456**	0.409	0.376	0.437
C	**0.685**	0.662	0.655	0.682

TABLE 9. The performance of Multi-S3P and other two state-of-the-arts methods in the boundary regions on the CASP12 dataset.

Methods	Q8 Accuracy (%)
NetSurf-2.0	52.10
SPOT-1D	57.85
Multi-S3P (this work)	**59.90**

TABLE 10. Q3 prediction accuracy (%) with different combinations of input features on our validation set.

Feature Combinations	Q3 Accuracy (%)
PSSM	85.40
HMM	83.78
PSSM + HMM	86.91
PSSM + HMM + 7PCP	87.67
PSSM + HMM + 7PCP + PSP19	**87.81**

that the HMM feature was complementary to the PSSM, which was consistent with the findings of other state-of-the-arts predictors [19], [24], [32]. When the PSSM profile was combined with the other three input features (HMM, 7PCP and PSP19), the accuracy increased by 2.41%. However, when the 7PCP and PSP19 were combined with PSSM and HMM, the accuracy improved by 0.9%.

An ablation study is carried out using the best-performing model to show the benefits of the CNN and LSTM network

TABLE 11. Q3 prediction accuracy (%) of individual model performance on our validation set as compared to the ensemble performance.

Network Combinations	Q3 Accuracy (%)
Multi-S3P	**87.81**
Without CNN	87.10
Without LSTM	86.89
Without Self-Attention	87.38

we proposed. We performed this study by removing the CNN components and LSTM components separately. In addition, we also tested our model by removing the self-attention layer in the output layer. Table 11 shows the individual model performance on our validation set as compared to the ensemble model (Multi-S3P) performance.

The novelty of our model lies in the combination of three specialized deep neural models for the different types of SS classes (sheets, coils, and helices) with a self-attention layer to predict protein SS. We found that training separate networks for each type and combining their outputs through a self-attention layer led to improved performance. This approach allows the model to capture better the unique characteristics of each type of secondary structure and to learn how to combine this information more effectively. Additionally, using a self-attention layer allows the model to focus on the most important features for predicting secondary

structure, further improving performance. Our approach is, therefore, a novel and effective way of predicting protein secondary structure, with potential applications in a range of bioinformatics and related fields. Moreover, we introduced a new criterion to assess the model performance in the SS boundary regions, which is a challenging task in protein secondary structure prediction. Overall, the results demonstrate that the proposed multi-network-based method is highly effective for secondary structure prediction and outperforms other state-of-the-art methods.

VI. CONCLUSION

In conclusion, proteins are essential for living organisms due to their diverse functions in cells. The native 3D structure of a protein determines its function, and misfolded proteins can cause acute illnesses in living organisms. The computational prediction of protein structures has been developed as an alternative to experimental techniques. The AlphaFold2 deep-learning approach has shown its full potential in predicting protein structures with high accuracy. Although the accuracy of protein structure prediction has been increased by deep learning techniques, the complexity of protein 3D structures still poses a challenge for computational prediction algorithms. Therefore, researchers generally divide this problem into manageable sub-problems, such as secondary structure prediction. Secondary structure prediction, which is critical for predicting numerous structural features, offers information about protein activity, functions, and relationships. The integration of sequence evolutionary profile features derived from multiple sequence alignments, such as PSSM and the HMM feature, have played a significant role in protein secondary structure prediction over the last few decades.

In this paper, a multi-network-based deep learning model (Multi-S3P) is designed to classify different classes of protein secondary structures, including sheets, helices, and coils. The Multi-S3P consists of three separate networks, each specialized in recognizing a specific type of structure. The first network is dedicated to identifying helices, the second to recognizing coils, and the third to perceiving sheets. The architecture used CNN and BILSTM networks with a self-attention mechanism to learn short-range and long-range interactions. Attention mechanisms can be useful for combining information from multiple sources and learning to focus on the most relevant parts of the input. The four input features used were PSSM, HMM, 7PCP, and PSP19. It is evident from the results presented in the work that the proposed Multi-S3P outperformed the state-of-the-art methods in terms of Q3 and Q8 accuracy and segment overlap measure (SOV), achieving the highest Q3 accuracy of 87.57% and a Q8 accuracy of 77.56% on the TEST2016 test set. In addition to the model accuracy and SOV, we employed the precision, recall, and F1-score to better understand how different techniques performed. The results demonstrate the effectiveness of the proposed method and its potential to improve secondary structure prediction performance. The combination of different input features was shown to impact performance significantly. The PSSM profile has greater predictive performance than the other profile. Our proposed model, Multi-S3P, shows high performance in effectively predicting SS in the boundary regions, which is a challenging task in protein secondary structure prediction. Overall, the proposed method provides a promising approach for accurate secondary structure prediction of proteins. However, it is important to note that training and tuning multiple separate models with an attention-based mechanism can be computationally expensive. Therefore, we intended to use the lightweight language model embedding generated by pre-trained protein language models as input features to overcome the computation cost in our future work.

DATA AVAILABILITY

The original contributions presented in the study are included in the article and the data and code used in this study are available at https://github.com/mufassirin/Multi-S3P; further inquiries can be directed to the corresponding author.

REFERENCES

[1] N. D. Jana, S. Das, and J. Sil, *A Meta-Heuristic Approach to Protein Structure Prediction*. Cham, Switzerland: Springer, 2018.
[2] M. M. M. Mufassirin, M. A. H. Newton, and A. Sattar, "Artificial intelligence for template-free protein structure prediction: A comprehensive review," *Artif. Intell. Rev.*, pp. 1–68, Dec. 2022.
[3] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.
[4] M. AlQuraishi, "End-to-end differentiable learning of protein structure," *Cell Syst.*, vol. 8, no. 4, pp. 292.e3–301.e3, Apr. 2019.
[5] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
[6] J. Pereira, A. J. Simpkin, M. D. Hartmann, D. J. Rigden, R. M. Keegan, and A. N. Lupas, "High-accuracy protein structure prediction in CASP14," *Proteins, Struct., Function, Bioinf.*, vol. 89, no. 12, pp. 1687–1699, 2021.
[7] M. Varadi et al., "AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Res.*, vol. 50, no. D1, pp. D439–D444, Jan. 2022.
[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[9] M. R. Uddin, S. Mahbub, M. S. Rahman, and M. S. Bayzid, "SAINT: Self-attention augmented inception-inside-inception network improves protein secondary structure prediction," *Bioinformatics*, vol. 36, no. 17, pp. 4599–4608, Nov. 2020.
[10] G. Xu, Q. Wang, and J. Ma, "OPUS-TASS: A protein backbone torsion angles and secondary structure predictor based on ensemble neural networks," *Bioinformatics*, vol. 36, no. 20, pp. 5021–5026, Dec. 2020.
[11] Y. Gao, S. Wang, M. Deng, and J. Xu, "RaptorX-angle: Real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning," *BMC Bioinf.*, vol. 19, no. S4, pp. 73–84, May 2018.
[12] M. A. H. Newton, F. Mataeimoghadam, R. Zaman, and A. Sattar, "Secondary structure specific simpler prediction models for protein backbone angles," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–14, Dec. 2022.
[13] F. Mataeimoghadam, M. A. H. Newton, A. Dehzangi, A. Karim, B. Jayaram, S. Ranganathan, and A. Sattar, "Enhancing protein backbone angle prediction by using simpler models of deep neural networks," *Sci. Rep.*, vol. 10, no. 1, p. 19430, Nov. 2020.
[14] J. Rahman, M. A. H. Newton, M. A. M. Hasan, and A. Sattar, "Real-to-bin conversion for protein residue distances," *Comput. Biol. Chem.*, vol. 104, Jun. 2023, Art. no. 107834.
[15] J. Rahman, M. A. H. Newton, M. A. M. Hasan, and A. Sattar, "A stacked meta-ensemble for protein inter-residue distance prediction," *Comput. Biol. Med.*, vol. 148, Sep. 2022, Art. no. 105824.

[16] M. A. H. Newton, J. Rahman, R. Zaman, and A. Sattar, "Enhancing protein contact map prediction accuracy via ensembles of inter-residue distance predictors," *Comput. Biol. Chem.*, vol. 99, Aug. 2022, Art. no. 107700.

[17] P. Di Lena, K. Nagata, and P. Baldi, "Deep architectures for protein contact map prediction," *Bioinformatics*, vol. 28, no. 19, pp. 2449–2457, Oct. 2012.

[18] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain," *Proc. Nat. Acad. Sci. USA*, vol. 37, no. 4, pp. 205–211, Apr. 1951.

[19] C. Fang, Y. Shang, and D. Xu, "MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction," *Proteins, Struct., Function, Bioinf.*, vol. 86, no. 5, pp. 592–598, May 2018.

[20] Y. Ma, Y. Liu, and J. Cheng, "Protein secondary structure prediction based on data partition and semi-random subspace method," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Jun. 2018.

[21] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers, Original Res. Biomolecules*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

[22] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, and Y. Zhou, "Sixty-five years of the long March in protein secondary structure prediction: The final stretch?" *Briefings Bioinf.*, vol. 19, no. 3, pp. 482–494, 2018.

[23] 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, pp. 1–8, Aug. 2019.

[24] Z. Guo, J. Hou, and J. Cheng, "DNSS2: Improved ab initio protein secondary structure prediction using advanced deep learning architectures," *Proteins, Struct., Function, Bioinf.*, vol. 89, no. 2, pp. 207–217, 2021.

[25] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," *Biochemistry*, vol. 13, no. 2, pp. 222–245, 1974.

[26] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[27] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.

[28] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, Feb. 2012.

[29] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *Mol. Model. Annu.*, vol. 7, no. 9, pp. 360–369, 2001.

[30] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, and Y. Zhou, "Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning," *Sci. Rep.*, vol. 5, no. 1, pp. 1–11, Jun. 2015.

[31] R. Heffernan, Y. Yang, K. Paliwal, and Y. Zhou, "Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility," *Bioinformatics*, vol. 33, no. 18, pp. 2842–2849, Sep. 2017.

[32] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks," *Bioinformatics*, vol. 35, no. 14, pp. 2403–2410, Jul. 2019.

[33] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *J. Mol. Biol.*, vol. 202, no. 4, pp. 865–884, Aug. 1988.

[34] L. H. Holley and M. Karplus, "Protein secondary structure prediction with a neural network," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 1, pp. 152–156, 1989.

[35] S. C. Schmidler, J. S. Liu, and D. L. Brutlag, "Bayesian segmentation of protein secondary structure," *J. Comput. Biol.*, vol. 7, nos. 1–2, pp. 233–248, Feb. 2000.

[36] J.-F. Gibrat, J. Garnier, and B. Robson, "Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs," *J. Mol. Biol.*, vol. 198, no. 3, pp. 425–443, 1987.

[37] S. Wang, J. Peng, J. Ma, and J. Xu, "Protein secondary structure prediction using deep convolutional neural fields," *Sci. Rep.*, vol. 6, no. 1, pp. 1–11, Jan. 2016.

[38] M. S. Klausen, M. C. Jespersen, H. Nielsen, K. K. Jensen, V. I. Jurtz, C. K. Sønderby, M. O. A. Sommer, O. Winther, M. Nielsen, B. Petersen, and P. Marcatili, "NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning," *Proteins, Struct., Function, Bioinf.*, vol. 87, no. 6, pp. 520–527, Jun. 2019.

[39] C. Fang, Z. Li, D. Xu, and Y. Shang, "MUFold-SSW: A new web server for predicting protein secondary structures, torsion angles and turns," *Bioinformatics*, vol. 36, no. 4, pp. 1293–1295, Feb. 2020.

[40] L. Yuan, X. Hu, Y. Ma, and Y. Liu, "DLBLS_SS: Protein secondary structure prediction using deep learning and broad learning system," *RSC Adv.*, vol. 12, no. 52, pp. 33479–33487, 2022.

[41] M. AlQuraishi, "ProteinNet: A standardized data set for machine learning of protein structure," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–10, Dec. 2019.

[42] J. Singh, J. Singh, K. Paliwal, A. Busch, and Y. Zhou, "SPOT-1D2: Improving protein secondary structure prediction using high sequence identity training set and an ensemble of recurrent and residual-convolutional neural networks," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2021, pp. 1–7.

[43] J. Ludwiczak, A. Winski, A. M. D. S. Neto, K. Szczepaniak, V. Alva, and S. Dunin-Horkawicz, "PiPred—A deep-learning method for prediction of π-helices in protein sequences," *Sci. Reports.*, vol. 9, no. 1, p. 6888, 2019.

[44] M. H. Høie, E. N. Kiehl, B. Petersen, M. Nielsen, O. Winther, H. Nielsen, J. Hallgren, and P. Marcatili, "NetSurfP-3.0: Accurate and fast prediction of protein structural features by protein language models and deep learning," *Nucleic Acids Res.*, vol. 50, no. W1, pp. W510–W515, Jul. 2022.

[45] J. Singh, K. Paliwal, T. Litfin, J. Singh, and Y. Zhou, "Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment," *Sci. Rep.*, vol. 12, no. 1, p. 7607, May 2022.

[46] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 15, Apr. 2021, Apr. no. e2016239118.

[47] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing," 2020, *arXiv:2007.06225*.

[48] W. Yang, Y. Liu, and C. Xiao, "Deep metric learning for accurate protein secondary structure prediction," *Knowl.-Based Syst.*, vol. 242, Apr. 2022, Art. no. 108356.

[49] B. Rost, "Protein secondary structure prediction continues to rise," *J. Struct. Biol.*, vol. 134, nos. 2–3, pp. 204–218, 2001.

[50] D. P. Ismi, R. Pulungan, and Afiahayati, "Deep learning for protein secondary structure prediction: Pre and post-AlphaFold," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 6271–6286, Nov. 2022.

[51] C.-T. Ho, Y.-W. Huang, T.-R. Chen, C.-H. Lo, and W.-C. Lo, "Discovering the ultimate limits of protein secondary structure prediction," *Biomolecules*, vol. 11, no. 11, p. 1627, Nov. 2021.

[52] A. Mittal, B. Jayaram, S. Shenoy, and T. S. Bawa, "A stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding?" *J. Biomolecular Struct. Dyn.*, vol. 28, no. 2, pp. 133–142, 2010.

[53] 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, Sep. 2017.

[54] Z. Lyu, Z. Wang, F. Luo, J. Shuai, and Y. Huang, "Protein secondary structure prediction with a reductive deep learning method," *Frontiers Bioeng. Biotechnol.*, vol. 9, p. 404, Jun. 2021.

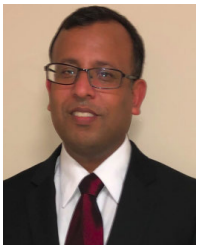
[55] G. Wang and R. L. Dunbrack Jr., "PISCES: A protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, Aug. 2003.

[56] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, Sep. 2014.

[57] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4039–4045, Dec. 2018.

[58] M. Torrisi, M. Kaleel, and G. Pollastri, "Porter 5: Fast, state-of-the-art *ab initio* prediction of protein secondary structure in 3 and 8 classes," *BioRxiv*, p. 289033, Mar. 2018.
[59] C. Sander and R. Schneider, "Database of homology-derived protein structures and the structural meaning of sequence alignment," *Proteins, Struct., Function, Genet.*, vol. 9, no. 1, pp. 56–68, Jan. 1991.
[60] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 90, no. 16, pp. 7558–7562, Aug. 1993.
[61] G. Xu, T. Ma, T. Zang, W. Sun, Q. Wang, and J. Ma, "OPUS-DOSP: A distance- and orientation-dependent all-atom potential derived from side-chain packing," *J. Mol. Biol.*, vol. 429, no. 20, pp. 3113–3120, Oct. 2017.
[62] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu, "UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, Mar. 2015.
[63] M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, and J. Söding, "HH-suite3 for fast remote homology detection and deep protein annotation," *BMC Bioinformatics*, vol. 20, p. 473, Sep. 2019.
[64] M. Mirdita, L. von den Driesch, C. Galiez, M. J. Martin, J. Söding, and M. Steinegger, "Uniclust databases of clustered and deeply annotated protein sequences and alignments," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D170–D176, Jan. 2017.
[65] J.-L. Fauchère, M. Charton, L. B. Kier, A. Verloop, and V. Pliska, "Amino acid side chain parameters for correlation studies in biology and pharmacology," *Int. J. Peptide Protein Res.*, vol. 32, no. 4, pp. 269–278, Jan. 2009.
[66] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[67] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[68] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[69] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[70] B. Yang, L. Wang, D. F. Wong, S. Shi, and Z. Tu, "Context-aware self-attention networks for natural language processing," *Neurocomputing*, vol. 458, pp. 157–169, Oct. 2021.
[71] R. Heffernan, K. Paliwal, J. Lyons, J. Singh, Y. Yang, and Y. Zhou, "Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning," *J. Comput. Chem.*, vol. 39, no. 26, pp. 2210–2216, Oct. 2018.
[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014, *arXiv:1412.6980*.
[73] Y. Sasaki. (2007). *The Truth of the F-Measure*. Accessed: May 26, 2021. [Online]. Available: https://www.cs.odu.edu/mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf
[74] Y. Zhang, "Progress and challenges in protein structure prediction," *Current Opinion Struct. Biol.*, vol. 18, no. 3, pp. 342–348, Jun. 2008.

M. M. MOHAMED MUFASSIRIN received the B.Sc. degree (Hons.) in computer science from the South Eastern University of Sri Lanka and the M.Sc. degree in computer science from the University of Peradeniya. He is currently pursuing the Ph.D. degree with the School of Information and Communication Technology, Griffith University, Australia. He is also a Lecturer of computer science with the South Eastern University of Sri Lanka. His research interests include artificial intelligence, bioinformatics, machine learning, and protein design.



M. A. HAKIM NEWTON received the B.Sc.Eng. and M.Sc.Eng. degrees from the Bangladesh University of Engineering and Technology (BUET), and the Ph.D. degree from Strathclyde University, U.K. He is a Lecturer with the School of Information and Physical Sciences, The University of Newcastle, Australia. He is also an Adjunct Senior Research Fellow with the Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Australia. He was a Research Engineer with the National ICT Australia (NICTA). His research interests include artificial intelligence, intelligent search, machine learning, and bioinformatics.

JULIA RAHMAN is currently pursuing the Ph.D. degree with the School of Information and Communication Technology, Griffith University, Brisbane, Australia. She is an Assistant Professor and a Founding Member of the Machine Learning Research Group, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. Her research interests include bioinformatics, machine learning, and artificial intelligence.

ABDUL SATTAR is a Professor with the School of Information and Communication Technology, Griffith University, Australia. He was the Founding Director of the Institute for Integrated and Intelligent Systems, at Griffith University. He was also the Education Director with the Queensland Research Laboratory (QRL), National ICT Australia (NICTA). He won a number of ARC discovery grants and international awards for his work in artificial intelligence.