Chapter 4

A resource-efficient deep learning model for the discovery of antiviral peptides

In the recent past, the COVID-19 pandemic caused a substantial number of deaths and disabilities, which has prompted the scientific community to question our preparedness against such future viral outbreaks. This has led to a strong collaboration between biomedical scientists and artificial intelligence (AI) experts for the development of numerous *in silico* techniques for aiding and accelerating the discovery of wide spectrum therapeutic antivirals that can be derived from antiviral peptides (AVPs). The majority of living organisms synthesize these molecules as a crucial component of their innate immune response to counteract viral infections. The current investigation presents the Deep-AVPiden (DS) model and its corresponding online application, which is accessible at `https://deep-avpiden.anvil.app`. This research endeavor aims to identify previously unknown AVPs in the proteomes of various organisms. It is a computationally efficient temporal convolutional network (TCN)-based architecture incorporating depth-wise separable convolutions. The Deep-AVPiden (DS) model demonstrates an

accuracy of 88% and exhibits a precision level of 90%. Upon conducting a comparative analysis between the proposed model and the state-of-the-art classifiers, it was seen that the proposed model exhibits much superior performance in comparison to them. In order to evaluate the efficacy of the model, antiviral proteins present in the immune systems of various organisms, including plants, mammals, and fishes, were found and analyzed. The identified AVPs exhibited significant sequence similarity with antimicrobial peptides that laboratory-based tests have confirmed empirically.

4.1 Introduction

The discovery of novel antimicrobial drugs that kill or inhibit life-threatening pathogens is attracting much attention due to the incapacity and inefficiency of conventional antibiotics. However, it is pertinent that the new class of therapeutics must have high efficacy, broad-spectrum activity, and few or no side effects on human health. In this direction, medications can be developed using antimicrobial peptides (AMPs), which form an integral part of living organisms' natural first line of defense. Nowadays, analyzing and modeling AMPs using machine/deep learning has caught momentum [44, 76, 28, 5, 3, 77, 78]. Deep learning-based sequence modeling techniques such as recurrent neural networks (RNNs), long-short term memory (LSTM) networks, temporal convolutional networks (TCNs) [51, 52, 72], etc., can be effectively used to develop robust models to classify and discover novel therapeutic peptides in proteomes of various life-forms. Note that sequence modeling is a technique that inputs and outputs sequential data, which can be text, audio, video, etc. For this purpose, RNN was developed as a deep learning architecture for capturing dependencies between the units of a given sequence to make predictions. However, it fails to capture long-distance dependencies between these units due to the vanishing gradient problem. LSTMs were proposed as an improvement over RNNs in that they overcome this problem by using a gating mechanism (input, output, and forget gates) to remember the correlation among the

units over a long distance. However, LSTMs require more memory than RNNs to store partial results. Additionally, RNN and LSTM-based models work sequentially, so the elements of a given sequence (also known as timesteps) cannot be processed in parallel. However, such shortcomings are not present in TCNs. The computations performed by this deep learning architecture can be easily distributed and parallelized on multi-core processing systems, and it also does not consume much operational memory.

Antiviral peptides (AVPs) are AMPs that protect the host organism against viral infections by either directly attacking the viruses themselves or by targeting the host cells in order to impede viral reproduction. According to [79], certain antiviral peptides exhibit virucidal properties by either obstructing viral proteins outside the host cell or by competing for binding sites on the host cell membrane. On the other hand, several AVPs disrupt several stages of the viral life cycle, including but not limited to viral gene expression and replication. It is noteworthy that a considerable number of AVPs may be found within the proteomes of mammals, plants, and fishes. Cyclotides, a group of antiviral peptides present in several plant species, have been seen to exhibit inhibitory effects against a diverse range of human viruses, including human immunodeficiency virus (HIV) [80], H1N1 [81], and dengue [82]. These cyclotides have been discovered to impede the binding process between the viruses and the cell membrane of the host. Cecropin-A, sourced from a moth, exhibits antiviral properties against HIV through the inhibition of its genetic expression. In a similar vein, it has been observed that a group of AVPs called dermaseptins, which are present in the frogs, exhibit the ability to inhibit the replication of HIV-1 [83]. When considering antiviral peptides derived from marine species, it is worth noting that a specific group of peptides called clavanins has been found to effectively decrease the pathogenicity of herpes simplex virus (HSV), rotavirus, and adenovirus [84].

Several deep learning-based methods have been developed for the purpose of classifying the AVPs. The Deep-AVPpred model uses convolutional neural networks (CNNs)

to predict and identify antiviral peptides [85]. On the other hand, DeepAVP utilizes both bidirectional LSTM and CNN to achieve the same objective [86]. The authors of [49] conducted a study in which they employed bi-LSTM, CNN, and support vector machine (SVM) based models to perform multi-label classification. The objective was to predict several functional activities demonstrated by a peptide, such as antiviral, antibacterial, antifungal, etc. In yet another research, Kurata et al. [87] conducted experiments involving multiple machine and deep learning algorithms, including Transformers, CNNs, bi-LSTM, Random Forests (RFs), and SVM. Their findings indicated that the RF model achieved the highest performance in predicting anti-coronavirus peptides. CNNs are unable to effectively capture the long-distance dependencies that exist between the units of an AVP sequence. The bi-LSTM-based models are able to overcome this limitation. However, they still face difficulties while dealing with very long sequences. Moreover, training and optimizing a bi-LSTM model is time-consuming, mostly because of its sequential mode of execution and inherent non-parallelizability. Additionally, it requires a substantial amount of memory resources [72]. In conclusion, a notable challenge encountered by deep neural networks is the resource and computation-intensive nature of both training and operation processes. Also, the size of the models poses significant challenges in terms of training and deployment, particularly on devices with limited resources.

In addition to deep learning algorithms, researchers have utilized quantifiable aspects of peptides, including their physicochemical, compositional, and structural characteristics, in conjunction with machine learning algorithms such as SVM and RFs to construct classifiers for antiviral peptides. In the study conducted by Thakur et al. [8], a range of manually designed features were employed to classify AVPs based on peptide sequences. These features included motifs, amino acid composition, and certain physicochemical properties. The AntiVPP 1.0 model employs an RF model, which utilizes compositional and physicochemical properties in order to make predictions on antiviral

peptides. The AVPiden model, presented by Pang et al. [9], utilizes RFs for a two-stage classification process. The initial phase involves the classification of peptides into two groups: AVPs and non-AVPs. Subsequently, the second phase entails the assessment of the antiviral potential of AVPs against eight specific types and six distinct families of viruses. Qureshi et al. [88] utilized four machine learning methods for the purpose of AVP classification. The ENNAVIA model performs the classification of AVPs and non-AVPs by utilizing a deep neural network architecture and incorporates physico-chemical and compositional properties [13]. In a research conducted by Schaduangrat et al. [89], a total of six machine learning methods were employed to do AVP classification. In yet another study, the PreTP-Stack model is constructed utilizing a set of ten distinct features and employing four distinct machine learning techniques [15]. Lastly, the FIRM-AVP model was proposed in [14] to investigate the efficacy of three distinct machine learning strategies in constructing an AVP classifier. The authors determined that the SVM-based model exhibited superior performance compared to the other approaches. One of the primary limitations associated with the utilization of machine learning-based models is the increased requirement for the collection and refinement of manually crafted features that function as their input. Moreover, it should be noted that machine learning models exhibit inferior performance compared to their deep learning equivalents when presented with extensive datasets. Another additional limitation of most of these studies is the lack of specialized web servers to assist wet lab researchers in the discovery of AVPs [90].

In order to address the limitations described above, a novel approach called Deep-AVPiden (DS) has been proposed, which leverages TCNs [76, 51, 72]. Deep-AVPiden (DS) is designed as a lightweight model to classify and discover AVPs using depth-wise separable convolutions (DwSCs), significantly reducing the number of training parameters. The suggested model has been trained to detect AVPs in proteins across a range of taxa, including mammals, plants, amphibians, fishes, arthropods, and oth-

ers. The performance of the model has been evaluated in comparison to recent classifiers such as AVPIden, ENNAVIA, iAMP-CA2L, Meta-iAVP, PreTP-Stack, iACVP, and DeepAVP. The findings demonstrate that the proposed model outperforms these classifiers. A web application based on the proposed model has been deployed at `https://deep-avpiden.anvil.app/`. In addition to its ability to classify AVPs, this application can also identify AVPs within proteins. To demonstrate the functionality of this application, several antiviral proteins present in various mammalian, plant, and fish species have been found. The primary contributions of this study are outlined as follows:

1. A novel lightweight deep learning model, Deep-AVPiden (DS), has been proposed based on the depth-wise separable TCNs to discriminate between AVPs and non-AVPs.
2. The proposed model has been subjected to a comparative analysis with existing state-of-the-art classifiers, demonstrating its superior performance in comparison to them.
3. An online application has been developed and implemented at https://deep-avpiden.anvil.app/. This application serves the purpose of assisting wet-lab researchers in identifying AVPs within protein sequences.
4. The web application facilitated the identification of fifteen AVPs within proteins found in various organisms.

The rest of this chapter is organized as follows. Section 4.2 presents the tools, techniques, and dataset used in this work. The proposed work is elucidated in section 4.3. The results and outcomes have been elaborated in section 4.4. The concluding remarks and future works are discussed in section 4.5.

4.2 Data and preliminaries

4.2.1 Dataset

The proposed model takes peptides as data points, which are essentially sequences of letters representing standard amino acids (AAs). The AVPs were obtained from multiple sources, including AVPdb [91], HIPdb [92], the starPep database [34, 57, 58], DRAMP [32], and the SATPdb [93]. The non-AVPs were obtained from the Swiss-Prot database [35] and AVPdb. Following the acquisition of 10,500 AVPs and 9000 non-AVPs, data cleaning was conducted. Peptides comprising non-standard AAs (B, J, O, U, X, and Z) and possessing an AA count below five or above fifty were excluded. Subsequently, the CD-HIT program [94, 95, 96] was employed individually on the AVPs and non-AVPs datasets, with a threshold value of 0.9 to eliminate highly similar sequences from each dataset. In order to mitigate any potential bias in performance resulting from the unequal distribution of instances across classes, a total of 699 non-AVPs were randomly eliminated. The final dataset comprised 5414 peptides, with 2707 AVPs and 2707 non-AVPs. These peptides were further divided into three subsets: training (70%), test (15%), and validation (15%) sets, which were used to build and evaluate the proposed model.

4.2.2 Word embeddings

Word embedding techniques are utilized to transform individual words, namely the numerical representations of AA residues, into vectors of a fixed length. One-hot encoding (OHE) and word2vec are widely employed techniques for achieving this objective. Two widely utilized word2vec algorithms exist, namely the skip-gram and continuous bag of words (CBoW). These algorithms aim to transform individual words, represented by OHE vectors, into feature vectors of fixed length. This transformation is achieved by considering the context of each word, which pertains to the words that appear in

proximity to the given word within the training set's data points. Consequently, words that are semantically comparable are assigned similar feature vector representations.

4.2.3 Temporal Convolutional Networks

Temporal Convolutional Networks (TCNs) are composed of one or more blocks comprising one-dimensional convolutional (1D-CONV) layers. Within these hierarchical levels, the filter taps have the potential to be implemented on the input units or timesteps in a non-consecutive manner. To simplify, dilated convolutions are employed, wherein it is not mandatory for the filter taps to be applied to consecutive timesteps within a given layer. The extent of the dilation factor (d) influences the size of the receptive field. A larger dilation factor increases the size of the receptive field, which allows for an effective capture of dependencies between distant timesteps.

TCNs can be categorized into two types: acausal and causal. The causal TCNs employ a convolutional layer that solely relies on the preceding timesteps (ranging from 1 to t-1) to compute the output at a given timestep t". Conversely, acausal TCNs incorporate both past and future time steps in order to accomplish this task. In this study, causal TCNs have been employed to construct the proposed model.

4.2.4 Depth-wise separable convolutions

The emergence of depth-wise separable convolutions (DwSCs) might be attributed to the growing motivation within the academic community to develop compact and efficient models. Prior to the conceptualization of this notion, either the pre-existing models underwent compression or the underlying networks had to be made less complex. Therefore, as a substitute, DwSCs were proposed by Sifre et al. [97] and subsequently applied with success by Ioffe et al. [98] and Howard et al. [99]. The conventional convolution process is decomposed into two distinct components, namely depth-wise and point-wise convolutions. These components can be defined as follows.

1. **Depth-wise Convolutions**: During this stage, a filter is individually applied to each input channel. In the context of a standard convolution operation, where N filters of dimensions 1 X f_k X n_c (f_k denoting the specified filter size and n_c representing the number of channels) are applied to a 2-D matrix of dimension 1 X M X n_c. But in this case, only a single filter of dimensions 1 X f_k X 1 is individually applied to each of the n_c channels. Consequently, this process yields an output of dimension 1 X M' X n_c.
2. **Point-wise Convolutions**: The output of the depth-wise convolutions is subjected to a 1 X 1 convolutional layer consisting of N filters of dimensions 1 X 1 X n_c. The output is a matrix of dimensions 1 x M' x N.

This type of factorization greatly decreases the number of training parameters in the network architecture, significantly decreasing the number of calculations performed during the model training phase. Thus, the resulting model can be trained in a short duration, has fewer storage requirements, and may be effectively trained and deployed on platforms with limited resources, such as mobile devices.

4.3 Proposed Work

As depicted in Figure 4.1, the Deep-AVPiden (DS) model comprises multiple layers, which are elucidated as follows.

1. **Embedding layer**: The skip-gram algorithm has been employed in this study to generate a word embedding matrix for a set of 20 standard AAs. The training set provided the data points for this purpose. The function of this particular layer is to transform the numerical string (which was generated using a one-to-one function that mapped each letter to a number) that represents a peptide into a feature matrix with dimensions (50,512) (the first element represents the size of each numerical string, while the second element represents the size of the fixed-length feature vector that represents each AA).

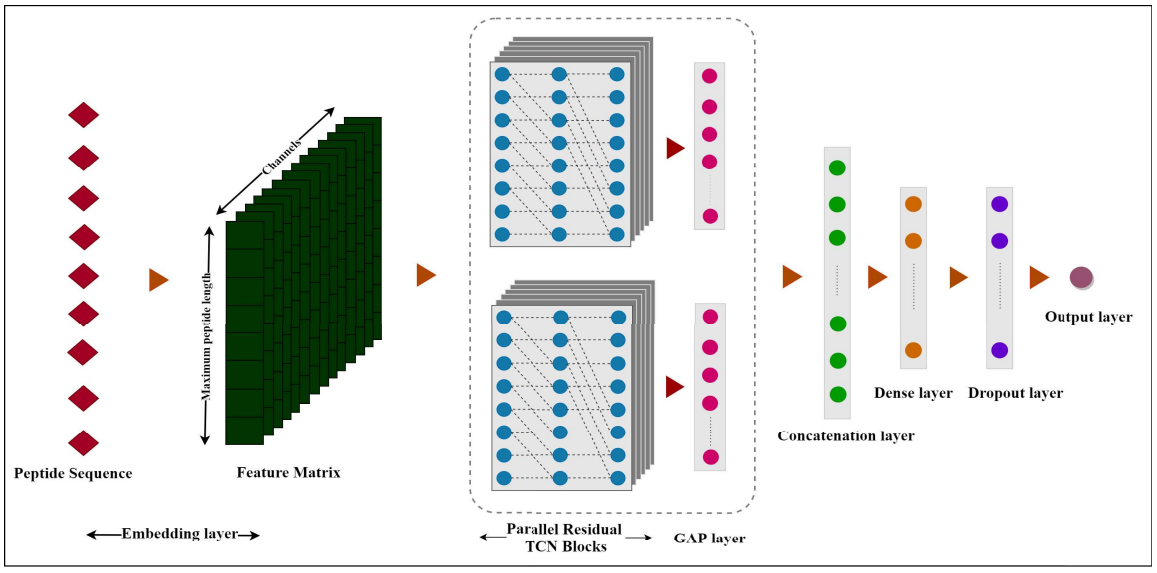

Figure 4.1: The Deep-AVPiden (DS) architecture

2. **Spatial Dropout layer**: This layer regularizes the input by removing some columns (frames) from the feature matrix, as opposed to removing individual elements. The utilization of this layer is favored in comparison to the conventional dropout layer in scenarios where a significant connection exists among the frames. The Deep-AVPiden (DS) model utilizes a one-dimensional spatial dropout layer subsequent to the embedding layer.
3. **TCN blocks**: The architecture of this model is based on a dilated causal TCN. It comprises two parallel TCN blocks, each composed of 1D-convolutional (CONV) layers with skip connections. The distinction between the two blocks is the varying dimensions of the filters employed within them. The activation function utilized by these layers is the rectified linear unit (ReLU). Batch normalization layers have been employed in between to enhance the stability of the learning process. The dilation factors (d) of 1, 2, and 4 were employed to build the TCN blocks.
4. **Global Average Pooling (GAP) layer**: A one-dimensional global average pooling (1D-GAP) layer is employed following each TCN block to compute the average of the feature map derived from it.

5. **Concatenation layer**: The output of the two GAP layers is combined by this layer for subsequent processing.
6. **Dense layer with dropout**: A dense layer has been added following the concatenation layer. It comprises a total of 64 units. Following the inclusion of this layer, a dropout layer is introduced in order to mitigate the issue of overfitting.
7. **Output layer**: This layer comprises a neuron that employs the sigmoid function as its activation. The resulting value produced by this neuron is a continuous real number that falls within the closed interval from 0 to 1. A peptide is classified as an AVP if the output value equals or exceeds 0.5.

4.4 Experiments, Results, and Discussions

Extensive experimentations have been conducted on the proposed model, Deep-AVPiden (DS), which was coded using the Python programming language. The trials were conducted on an NVIDIA V100 GPU core. The compute node also included 16 GB RAM. To execute the deep learning model, the Keras framework with Tensorflow was employed as the underlying platform [40].

4.4.1 Evaluation Criteria

Several evaluation metrics, like accuracy, recall, precision, and area under the receiver operating characteristic curve (AUC), were used to evaluate the proposed model. These metrics are as described in Eqs. 5.11-5.16.

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$Precision\ (Pr) = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall\ (Rec)\ (\text{or}\ True\ Positive\ Rate\ (TPR)) = \frac{TN}{TN + FP} \tag{4.3}$$

Table 4.1: Comparison of Deep-AVPiden with existing models on test set

Model	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)
Deep-AVPiden (DS)	88.47±0.13	88.49 ± 0.40	88.98 ± 0.38	94.90 ± 0.05
iACVP	65.83	77.33	46.59	75.49
AVPIden	59.98	57.20	73.74	68.81
Meta-iAVP	57.63	58.75	58.75	58.29
DeepAVP	53.08	53.94	58.99	52.77
iAMP-CA2L	52.36	88.89	6.23	52.72
PreTP-Stack	52.09	54.73	38.85	52.46
ENNAVIA	51.27	55.79	51.51	48.99

$$\textit{F1-score } (Fs) = \frac{2 \times Pr \times Rec}{Pr + Rec} \tag{4.4}$$

$$\textit{False Positive Rate } (FPR) = 1 - \frac{TN}{FP + TN} \tag{4.5}$$

$$AUC = \int TPR.\ \mathrm{d}(FPR) \tag{4.6}$$

4.4.2 Performance Evaluation

The proposed model was evaluated by conducting experiments on a test set, where it was compared against several state-of-the-art classifiers like DeepAVP [86], AVPIden [9], iAMP-CA2L [49], ENNAVIA [13], Meta-iAVP [89], PreTP-Stack [15], and iACVP [87]. It should be noted that the comparison was conducted exclusively among models that have eliminated identical and homologous sequences from their dataset. This step is crucial in order to mitigate any potential bias in the performance evaluation of a model. In addition, it should be noted that ENNAVIA and AVPIden exclusively categorize peptides that fall within the length ranges of [7,40] and [8,50], respectively. Additionally, the iACVP model categorizes sequences that consist of more than five AA residues. Hence, the test set was curated in accordance with their respective specifications. Furthermore, upon analysis of the iAMP-CA2L model, it was noted that it occasionally fails to classify the functional type of the peptide, that is, determining if an AMP is antiviral. In order to mitigate any potential misunderstanding, occurrences of such kind were excluded from the test set while presenting the findings for iAMP-CA2L.

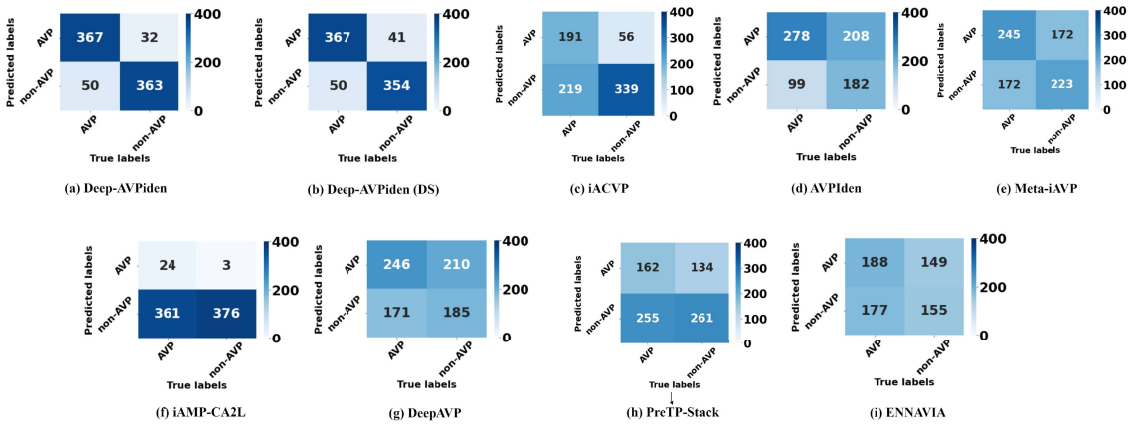

Figure 4.2: Confusion matrices obtained for various models including Deep-AVPiden on the test set

The performance outcomes of the state-of-the-art models, along with Deep-AVPiden (DS), have been displayed in Table 4.1. The superiority of the proposed model is clearly demonstrated as it exhibits significantly better performance across all metrics when compared to other models. The confusion matrices for various models are presented in Figure 4.2. It is apparent that the proposed model exhibits a higher number of true positives (TPs) and true negatives (TNs) while demonstrating a lower number of false positives (FPs) and false negatives (FNs) compared to other models.

4.4.3 Predicting AVPs using the web application

An online application based on Deep-AVPiden (DS), which is publicly available, has been set up at `https://deep-avpiden.anvil.app`. To demonstrate its working, the app was used to identify AVPs in the antiviral proteins of several animals, plants, and fishes. The antiviral proteins under consideration are classified into many families, such as ribosome-inactivating protein (RIP), RNA-binding protein (RBP), and Dicer-like protein (DCL), among others. The protein sequences listed in Table 4.2 were input into the deployed web application along with the provided parameters.

1. Probability Score: 0.90

2. Minimum length of desired AVPs : 10
3. Maximum length of desired AVPs : 30

Table 4.2: The AVPs discovered in the proteins of mammals, fish, and plants, with a probability score $\geq 90\%$ and showing some sequence similarity with the AMPs existing in public databases.

S.no	Accession number	Protein name	Protein Length	Discovered AVPs
1.	AAS77872.1	PAP	313	SDPFETNKCRYHI
2.	AAD32679.1	PIP	315	FAPASTWAASPNPI
3.	NP_197532.3	DCL4	1702	LSCILNNLELLRSWK
4.	AAB31048.1	Trichosanthin	289	FISNLRKALPNERKLYDIPLL
5.	NP_001319600.1	APUM5	913	EELVKQLAGQMVSLSLQMYGCR
6.	AAI12003.1	IFN-alpha-1	189	ICSLGCDLPQTHSLAHT
7.	ABD52364.1	IFN-alpha-2	187	FCTEPSSAAWNRTL
8.	AAI19352.1	IFN-alpha-3	186	FTSKDLSATWNATLLDSF
9.	EAW58615.1	IFN-alpha-4	187	VLNCKSICSLGCDLPQ
10.	AAM78026.1	IFN-alpha-5	189	CNSVCSLGCDLPQTHGLL
11.	ATI15613.1	TRIM-8	568	LCPFCCISHCT
12.	KAG1939425.1	Ubl	379	RRSWPEPVIHPEPV
13.	AAO37934.1	Mx	626	PENIGEQIKRLIRKFI
14.	NP_001187107.1	IFN	162	FLNILNTRQLTELT
15.	TSK18011.1	PRDX1	417	FVILEKMLMEICVIFSCV

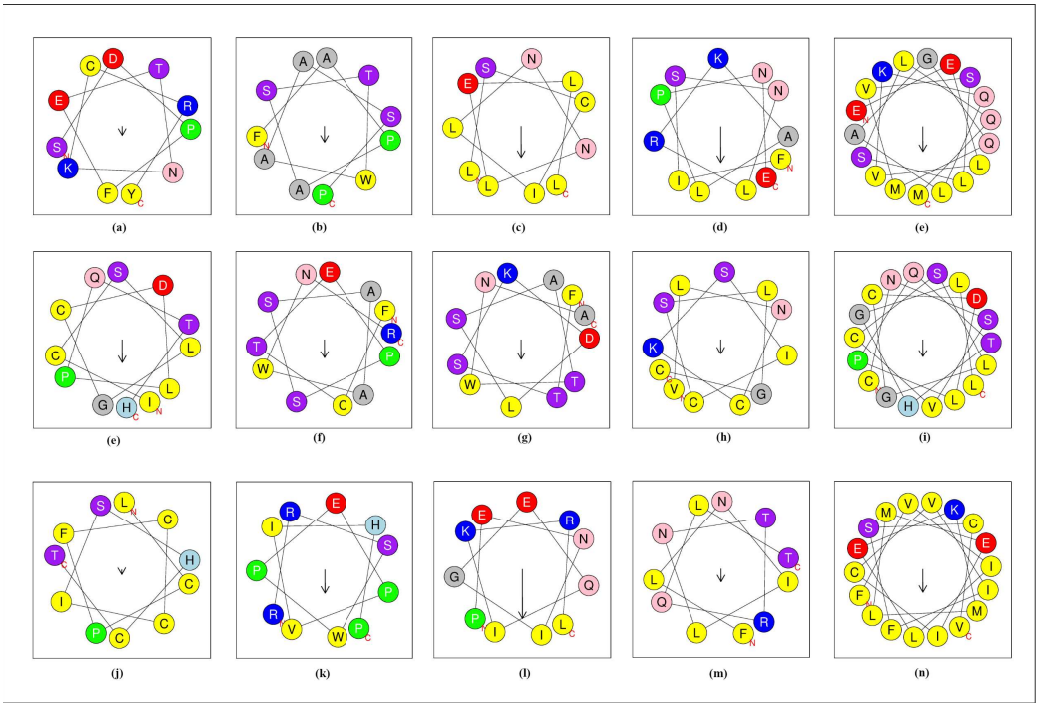

Figure 4.3: Alpha-helical representations of AVPs discovered in the plant, mammal, and fish proteins

Table 4.3: The AVPs discovered using Deep-AVPiden (DS) were subjected to BLAST analysis. Column 5 shows the method used to validate the AMPs similar to them as antimicrobial and/or antiviral (as mentioned in column 6). Column 7 consists of the similar AAs of discovered peptides and the ones found by BLAST analysis.

AVP discovered		Similar annotated AMP				
Sequence	No. of AAs	Sequence	No. of AAs	Validation Method	Nature	Similar AA positions with annotated AMPs
SDPFETNKCRYHI	13	VNT...QTT	262	X-Ray Diffraction	Antiviral	SDPFETNKCRYHI
FAPASTWAASPNPI	14	MET...GWF	224	Predicted (based on signature)	Antimicrobial	-AP-ST-A-SP-P
LSCILNNLELLRSWK	15	NWY...GIA	69	Predicted	Antimicrobial	L-CIL-N———
FISNLRKALPNERKLYDIPLL	21	DVS...NMA	247	X-Ray Diffraction	Antiviral	FISNLRKALPNERKLYDIPLL
EELVKQLAGQMVSLSLQMYGCR	22	DDG...GSC	42	Predicted (based on signature)	Antimicrobial	—-K-LAGQM————
ICSLGCDLPQTHSLAHT	17	CDL...SKE	165	Solution NMR	Antiviral	——CDLPQTHSL—
FCTEPSSAAWNRTL	14	MAF...NSP	195	Experimentally validated	Antiviral	F-TE-SSAAW-TL
FTSKDLSATWNATLLDSF	18	CDL...SKE	165	Solution NMR	Antiviral	F–KD-SA-W-TLLD–
VLNCKSICSLGCDLPQ	16	GSV...TKD	31	Experimentally validated	Antiviral	VLNC—C-LG——
CNSVCSLGCDLPQTHGLL	18	CDL...SKE	165	Solution NMR	Antiviral	———CDLPQTH-L-
LCPFCCISHCT	11	QSH...CKF	25	Predicted	Antimicrobial	LC-FCC——
RRSWPEPVIHPEPV	14	RRL...KPL	36	Predicted	Antimicrobial	-R-WP-P—P-P-
PENIGEQIKRLIRKFI	16	ELN...VEP	42	Predicted	Antimicrobial	-EN-GE-IK——-
FLNILNTRQLTELT	14	ATC...KGT	67	Predicted (based on signature)	Antimicrobial	——TRQLT-L-
FVILEKMLMEICVIFSCV	18	MHS...QNY	97	Predicted (based on signature)	Antimicrobial	F—E–L-E-C—SC-

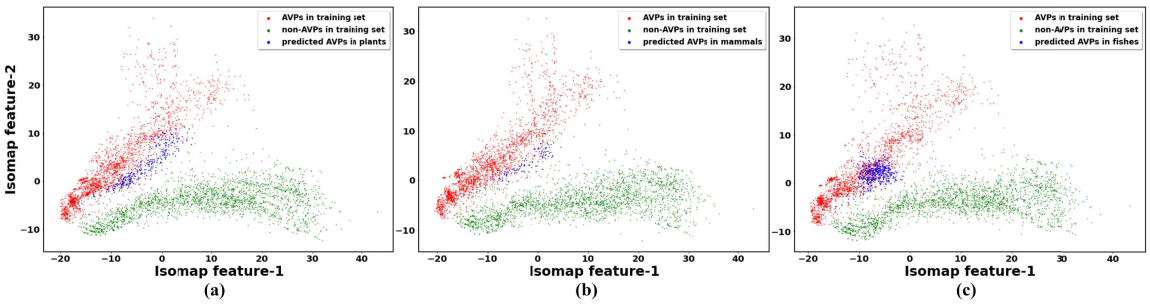

Figure 4.4: Scatter plot showing the distribution of AVPs predicted in the plant, mammal, and fish antiviral proteins, along with the AVPs and non-AVPs in the training set

The BLAST tool [100] has been utilized to verify the potential sequence similarity between the identified AVPs and the experimentally confirmed antimicrobial peptides (AMPs) available in public databases. Subsequently, the AVPs that exhibited high similarity with the annotated AMPs were identified and documented in Table 4.2 and their alpha-helical representations are shown in Figure 4.3. In addition to this, the methodology employed for confirming these AMPs and the corresponding amino acid positions with the identified AVPs are mentioned in Table 4.3. The antiviral potential of these AVPs is highly probable, and this hypothesis can be substantiated through laboratory-based chemical synthesis. The length of the identified AVPs exhibits a no-

table reduction in comparison to that of its corresponding parent protein. Therefore, this method effectively discerns the fundamental antiviral region in a specific protein, which is accountable for its antiviral efficacy. Further, an attempt was made to visually represent the distribution of these AVPs in relation to the AVPs present in the training set. In order to achieve this objective, the isometric mapping approach has been employed. The two-dimensional representation of the data points is depicted in Figure 4.4, revealing a noticeable similarity in the distributions of the predicted AVPs and the experimentally verified AVPs. Therefore, the identified antiviral peptides are believed to exhibit significant antiviral efficacy.

4.5 Conclusion

This study introduces a resource-efficient model called Deep-AVPiden (DS). The purpose of this model is to efficiently identify antiviral peptides inside protein sequences, thereby enhancing the pace of antiviral drug discovery. The proposed model is a deep learning architecture that utilizes depth-wise separable TCNs to make predictions regarding the antiviral properties of a particular peptide. As a result of the smaller number of parameters, this model can be implemented on devices with limited resources. The input for this model consists of peptides, which are represented as alphabetical strings. These peptides are transformed into feature matrices and fed into the model to calculate a probability score, which is utilized to assess the antiviral capabilities of the peptides. The model exhibits an accuracy of 89%, surpassing the performance of state-of-the-art classifiers.

In addition, a web application has been implemented and made accessible at `https://deep-avpiden.anvil.app/`. It allows users to input various proteins and discover AVPs, demonstrating promising potential in combating viral infections. After selecting fifteen antiviral proteins sourced from diverse mammals, plants, and fishes, a set of AVPs that are expected to exhibit promising antiviral properties have been identified.

These AVPs can be experimentally validated for their antiviral potential. In the future, advanced sequence modeling methodologies such as BERT may be used to construct a robust classification model. Additionally, the development of a multi-label classifier for AVPs can be considered which would classify an AVP based on the corresponding target virus family.