[1]Jyoti Yogesh Deshmukh

[2]Dr. Deepika Amol Ajalkar

[3]Anuja Krishna Gaikwad

[4]M. V. Shelke

[5]Ankita Harshad Tidake

[6]Sheetal Phatangare

# Gaussian-Based Dilated 1-D CNN for Classifying B-Cell Epitopes in Zika and Dengue Protein Sequences


***Abstract:*** - The main goal of designing peptide vaccines, conducting immunodiagnosis, and producing antibodies is to accurately identify linear B-cell epitopes. However, experimental analysis to determine these epitopes is costly. This study focuses on developing a Gaussian-based dilated 1-D CNN model for classifying epitopes and non-epitopes in protein sequences related to Zika and Dengue viruses. The Immune Epitope Database (IEDB) was used, containing a total of 1741 and 7020 linear B-cell epitopes for Zika and Dengue viruses, respectively. Physicochemical features of the protein sequences dataset were extracted using the Gaussian distribution to extract optimal features based on feature probability distribution. The proposed model achieved an accuracy score of 83.00% and 85.00%, precision of 87.00%, recall of 83.00% and 85.00%, and an F1-score of 84.00% and 86.00% over the Zika and Dengue datasets. The suggested model outperforms existing methods, demonstrating the potential of deep learning approaches in bioinformatics for enhancing epitope prediction in viruses, with implications for drug discovery and vaccine development.

***Keywords:*** Linear B-Cell, Protein Sequences, Amino Acid Sequences, Epitopes, Non-Epitopes, Deep Learning, Drug Discovery, Vaccine Development

## I. INTRODUCTION

The human the immune system's reaction relies heavily on antibodies, which are essential elements that identify and attach to the proteins of pathogenic organisms like bacteria or viruses [1,51,52,53,54]. An epitope represents the portion of an antigenic material that such antibodies identify. It is possible to identify a linear epitope, that is an ongoing chain of amino acids found within the linear protein sequence, as well as a conformal epitope, which is a group of amino acids that may be divided in the amino acid sequence but are situated strongly in the three-dimensional framework of the protein. In instances, uses like peptide-based vaccine development [2,55,56,57,58], immuno-diagnostic testing [3], and the synthesis of synthesized antibodies [47,48,49,50] depend on the recognition B-cell epitopes (BCEs). Statistical modeling can be crucial in the invention of novel vaccines and medications toward major viruses infections such the hiv , liver disease, or flu viruses, since clinical identification

[1] Marathwada Mitramandal's Institute of Technology, Pune, Maharashtra, India

jyoti1584@gmail.com

[2]G H Raisoni College of Engineering and Management, Pune, Maharashtra, India,

dipikaus@gmail.com

[3]MIT Art Desgin and Technology University's School of Computing, Pune, Maharashtra, India,

anuja.gaikwad@mituniversity.edu.in

[4]AISSMS Institute of Information Technology, Pune, Maharashtra, India

mayura.shelke@gmail.com

[5]Ajeenkya D Y Patil School of Engineering, Pune Maharashtra, India,

ankitatidake@dypic.in

[6]Vishwakarma institute of technology, Pune, Maharashtra, India,

sheetal.phatangare@vit.edu

of BCEs is costly and time-consuming [5]-[6]-[7]. The forecasting of continuous BCEs has drawn a lot of interest [8], despite the fact that conformational BCEs make up almost all of normally existing BCEs [9]. This is because linear BCEs are useful for peptide-based vaccine production, besides additional uses [10]-[11].

Some physiochemical feature of the individual amino acids, such as membrane access [12], fluidity [13], hydrophilic properties [14], or antigenic properties [15], was the single assessed by the early epitope predicting algorithms. Among the techniques which are presently available online include BEPITOPE [16], PEOPLE [17]. Using a folding frame across the search query peptide sequence, these methods determine the typical amino acid probability score for each characteristic [37,38,39,40]. A linear BCE is identified in the corresponding area of the sequence whenever the projected ratios for an ongoing portion of the amino acid are higher than a predetermined cut-off. On the other hand, using a particular amino acid profile or additionally a mixture of traits, an evaluation of 485 likelihood factors showed that these factors are ineffective to identify BCEs and slightly exceeded randomized BCE identification [41,42,43].

Innovative methods that were based on multiple likelihood factors and incorporated previously unincluded amino acid characteristics have been developed in response to the growing accessibility of empirically determined epitopes [19]. These techniques, which differentiate between BCEs and non-BCEs in the arrangement of amino acids using machine learning (ML) techniques, have demonstrated higher performance than individual likelihood scale-based techniques. BCEs are provided as features sets for learning the machine learning models, which originate from various features of amino acids, including the amino acid composition (AAC), the amino acid pair antigenicity level [20]. BepiPred 3.0 [21], ABCPred [22], AAAPred [23], SVMTrip [24], EpitopeVec [18], and EpiDope [25], EpitopeVec [26] represent a few instances of ML-based techniques for BCE modeling. One prominent problem appears to be that none of the previously listed techniques achieve higher performance when used in a cross-testing strategy, because ML training and validation are carried out on separate databases [44,45,46].

**Contribution of the paper**

- To develop the Gaussian based dilated 1-D CNN for classifying epitopes and non-epitopes in protein sequences associated with Zika and Dengue viruses.
- This study's contribution lies in its application of deep learning techniques to bioinformatics, aiming to improve epitope prediction in viruses. Such models have the potential to enhance for understanding of viral proteins, aiding in drug discovery and vaccine development efforts. The research underscores the importance of deep learning in bioinformatics and its potential to impact public health and medical research positively.

**Organization of the paper**

Section 2 presents the brief overview of dataset, Physicochemical feature extraction, Gaussian Distribution function, deep learning model. The suggested methodology and dilated 1D-CNN model are presented in section 3. Section 4 presents the result analysis, comparative analysis of the proposed model. Section 5 discussed the conclusion and future direction of the study.

## II. MATERIALS AND METHODS

Figure 1 shows the architecture of proposed model which consists of data inputs, feature extraction, building Deep learning model, and classification. Input of the model are protein sequence of Zika and Dengue virus dataset that consists of 20 amino acid charecters. The Physicochemical Features of 21-character text sequences are provided as inputs. These features need to be converted into an integer sequence using quantization coding. These features are extracted using Gaussian distribution to extract the optimal features based on the probability of feature distribution. The dilated 1D-CNN model is designed for classifying the epitopes and non-epitopes of protein sequences.

**Datasets**

In this study, extracted the piptide dataset of Zika and Dengue from the IEDB that contains the epitopes and non-epitopes. (http://www.iedb.org/). The IEDB database recorded a total of 1741 and 7020 linear B-cell epitopes of zika and dengue respectively. Among these total epitopes, 1261 positive epitopes and 480 negative epitopes were

recorded from the Zika virus dataset. Likewise, 5008 positive epitopes and 2012 negative epitopes are recorded from the dengue virus dataset.

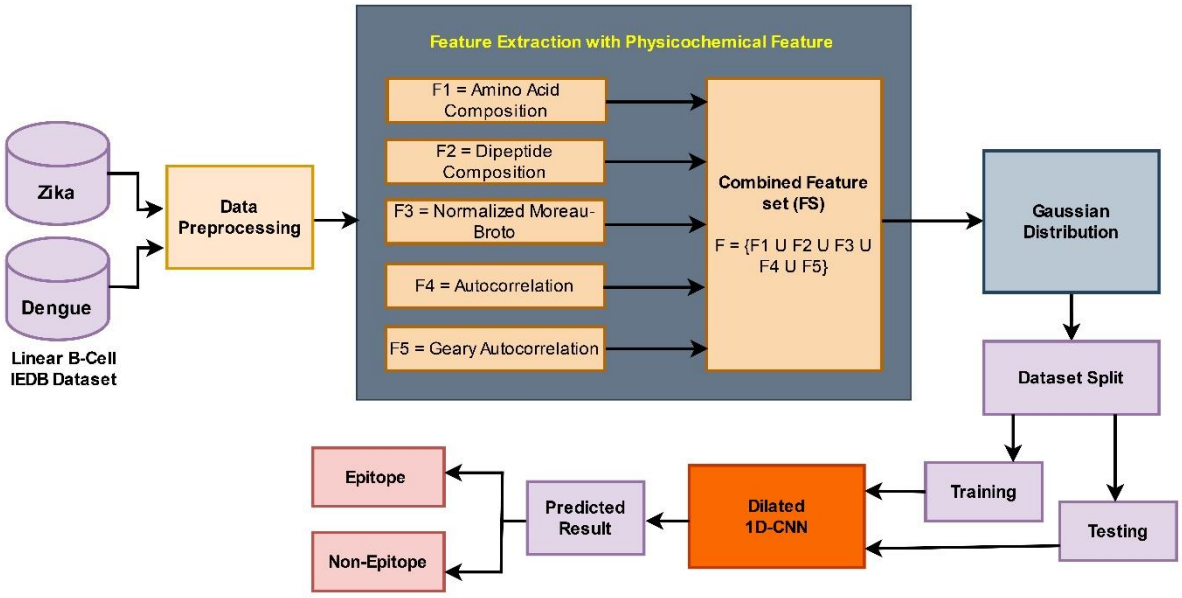


**Figure 1: Architecture of Proposed Model**

**Physicochemical Feature**

**Amino Acid Composition :** A vector indicating the proportional amount of every single amino acid in the protein serves as a representation of the AAC [27]. It could be expressed as:

$$ACC = (f1, f2, \dots \dots f20) \quad (1)$$

Where $f_i = \frac{A_i}{N}$ (i=1,2,3,…..20) shows the type of amino acid $i$, $A_i$ represents the total amount of amino acid sequence, and $N$ represents the lenght of amino acid sequence

**Dipeptide Composition**

A vector that specifies the quantity of dipeptides standardized over every possible dipeptide pairings for a given protein sequence is used to express dipeptide composition (DC). Its features remain constant at 400 in length [28]. It could be expressed as:

$$DC = (f1, f2, \dots \dots f400) \quad (2)$$

Where $f_i = \frac{A_i}{N}$ (i=1,2,3,…..400) shows the type of dipeptide $i$, $A_i$ represents the total amount of dipeptide composition, and $N$ represents the lenght of peptide

**Moreau-Broto Autocorrelation**

The Moreau-Broto autocorrelation feature is a method used to analyze protein sequences. It calculates the correlation between the properties of amino acids in a protein sequence and their positions within the sequence. Mathematically, the Moreau-Broto autocorrelation for a feature $P$of an amino acid sequence is given by:

$$MB - autocorrelation(P) = \sum_{i=1}^{n-1} \left( \frac{P(i) \,.\, P(i+1)}{i} \right) \quad (3)$$

where $n$ is the length of the sequence, and $P(i)$ represents the value of property $P$ for the amino acid at position ii in the sequence. This feature helps in predicting various properties of proteins, such as their function, with other molecules, by considering the spatial relationships between amino acids in the sequence.

**Moran Autocorrelation**

The Moran autocorrelation feature is a method used in bioinformatics to analyze protein sequences [29]. It calculates the correlation between the properties of amino acids in a protein sequence and the properties of neighboring amino acids. Mathematically, the Moran autocorrelation for a property $P$ of an amino acid sequence is given by:

$$Moran - autocorrelation(P) = \frac{n}{\sum_{i=1}^{n}(P(i) - \bar{p})^2} . \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(P(i) - \bar{P}).(P(j) - \bar{P})}{\sum_{i=1}^{n}(P(i) - \bar{p})^2} \quad (4)$$

Where nn is the length of the sequence, $P(i)$ represents the value of property PP for the amino acid at position ii in the sequence, and $\bar{P}$ is the average value of property $P$ across all amino acids in the sequence. This feature helps in predicting various properties of proteins by considering the spatial relationships between amino acids and their properties.

**Geary Autocorrelation**

The Geary autocorrelation feature is a method used to measures the similarity between the properties of amino acids at different positions in the sequence. Mathematically, the Geary autocorrelation for a property $P$ of an amino acid sequence is given by:

$$Geary - autocorrelation(P) = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n}(P(i) - P(j))^2 . w_{ij}}{2.\sum_{i=1}^{n}(P(i) - P(j))^2} \quad (5)$$

Where $n$ is the length of the sequence, $P(i)$ represents the value of property $P$ for the amino acid at position $i$ in the sequence, and $w_{ij}$ is a weight factor that can be defined based on the distance between positions $i$ and $j$ in the sequence. This feature helps in predicting various properties of proteins by considering the spatial relationships between amino acids and their properties.

**Combined Feature**

In this step combine the features can enhance the prediction of model by incorporating diverse information of amino acid sequences over the Zika and Dengue dataset. In this step combining all five physicochemical properties of amino acid sequence that can lead to improved accuracy in predicting epitopes.

$$F = ACC \cup DC \cup MBA \cup MA \cup GA \quad (6)$$

**Gaussian Distribution**

In this study, presented the Gaussian distribution to extract the optimal features for enhancing the weights of the targeted features and its adjacent features [30], so that the proposed deep model can train with optimal features of amino acid. The Gaussian distribution function is:

$$f(a) = \frac{1}{\sqrt{2\pi\delta}} exp\left(-\frac{(a-\mu)^2}{2\delta^2}\right) \quad (7)$$

Gaussian distribution is measured using equations 8

$$f(a) = \int_{-\infty}^{a} f(a)\, dx, \quad (8)$$

Gaussian probability distribution function is defined in equation 9

$$P(a) = F(a) - F(x - w) \quad (9)$$

where $w$ represents the token window, $\delta$ is the standard deviation, $\mu$ is the average of the distribution, and an is an actual value. To denote the length of every token in the tests, defined the token window $w$ to 1. It is also defined the optimum values of $\mu$ and d to 0 and 2.5, respectively. The design set is where these properties are adjusted. This probability is used as a feature in deep learning model for protein sequence prediction.

**Dilated 1D-CNN**

In this study, we designed a dilated 1-D CNN to classify epitopes and non-epitopes from protein sequences related to Zika and Dengue viruses. Before designing the proposed dilated 1D-CNN model, we set the parameters for training the model, which includes 3 channels with an input width of 17, 32 filters, a kernel size of 3, and a

dilation rate of 2. The model was trained with a batch size of 32 over 100 epochs using the Adam optimizer. The model's loss was measured using the binary cross-entropy function.

Given 1D protein sequence features $f: N -> R$ and kernel $k : \{0,1,\ldots n-1\} -> R$, the dilated convolutional function $(f * d\ k): N -> R$ is:

$$(f\ *d\ k)(s) = \sum_{i=0}^{n-1} k(i) * (s - id) \quad (10)$$

Where, $N$ is the real numbers, and n and d represents the kernel size and dilation parameter respectively. If $d = 1$, the neural network operate normal convolutopnal operator. In dilated CNN, residual connection is used to stability of the network [31]. In this network used $1\,x\,1$ convolutional layer to compare the size of input and output. The weight normalization is used in the kernel of dilated layer. To randomized apply the dropout to the output layer. The Relu is used as a activation function . The figure 2 shows the complete architecture of the 1D dialted CNN model.

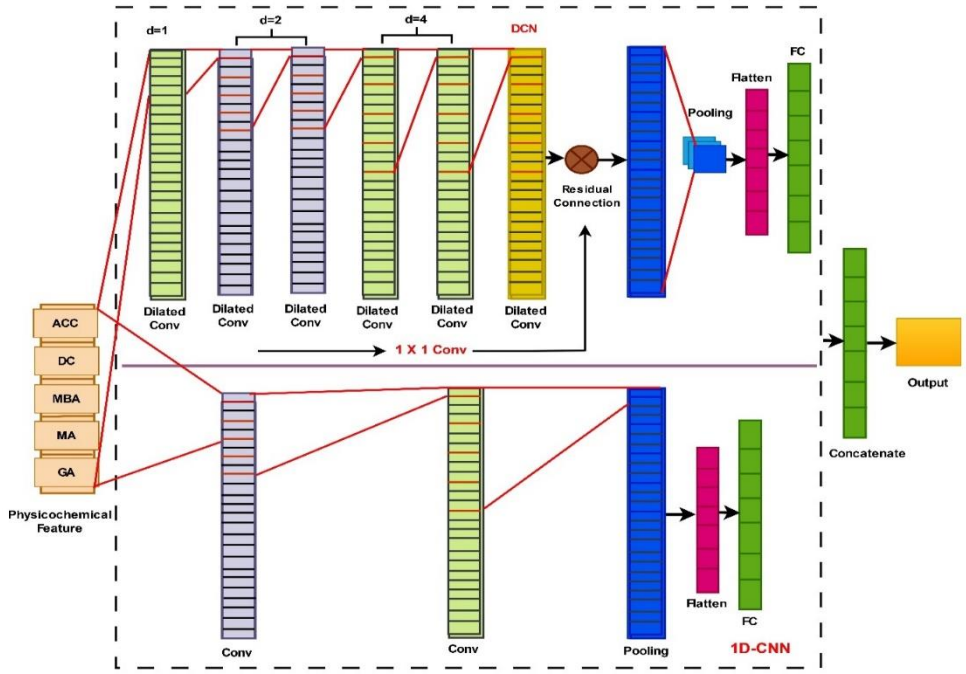


**Figure 2: Architecture of the Dilated 1D CNN Model**

**Dilated Convolutional Operator**

The dilated convolutional function can be defined $(F *_d\ k): N -> C$ as follows

$$(F *_d\ k)\ (s)\ =\ arg\min_{c} \sum_{l=0}^{N-1} k(i) d_C^2(F(s - id), C) \quad (11)$$

Where, C represents the valued operator of $(F *_d\ k)$. the C is used as manifold function which is equal to Euclidean dilated layer.

**Residual Connection**

If $F$ and $Y$ is the input and ouput of neural network. Based on the Euclidean residual connection, apply residual connection in two phases: 1) Concatenate $F$ and $Y(F)$ to obtained the number of input and output channels. 2) $wYC$ is used extract the ouput. Let, $R(F, Y_f)$ is the ouput of the residual connection [32], then $n^{th}$ channel of connection , $R_n(F, Y(F))$ is defined as

$$R_n(F, Y(F))(s) \stackrel{\text{def}}{=} \arg\min_{C} \quad (12)$$

$$\left(\sum_{i=1}^{input} k(i) d_C^2(F(s - id), C) + \sum_{j=1}^{count} k(j + input) d_C^2\,(Y_j(s), C)\right),$$

$$St\ \sum_i k(i) = 1, \forall k(i)\ > 0 \quad (13)$$

Where, $n \in (1,2,\ldots. output)$ and $F_i$ and $Y_j$ represents the $i^{th}$and $j^{th}$ channel of $F$ and $Y$ respectively.

**Loss Function**

The binary cross-entropy loss function is used for classifying epitopes and non-epitopes. It calculates the difference between the predicted probability distribution and the actual distribution of the epitopes and non-epitopes. Mathematically, it is defined as:

$$Binary\ Cross - Entropy\ Loss = -\frac{1}{N}\ \sum_{i=1}^{N}[y_i \log(\widehat{y_i}) + (1 - y_i)\log\ (1 - \widehat{y_i})] \quad (14)$$

Where $N$ is the number of features, $y_i$ is the actual label (epitopes and non-epitopes) for features $i$, and $\widehat{y_i}$ is the predicted probability of the features belonging to epitopes. The loss function penalizes the model more when it makes incorrect predictions with high confidence and less when it is uncertain. The goal is to minimize this loss function during training to improve the model's ability to correctly classify epitopes and non-epitopes.

**Pseudo Code:** Dilated 1D-CNN

**Input:** $F = F(ACC, DC, MBA, MA, GA)$ // Physicochemical Feature

**Output:** $Classify\ Epitope\ and\ Non - Epitope$

Define the Gaussian distribution function based on equation 6

Measure the Gaussian distribution using the cumulative distribution function based on equation 7

Define the Gaussian probability distribution function based on equation 8

**Function 1D-CNN**

Parameters (N, $F_{in}$, $F_{out}$, res, k1, d1, k2, d2, nC, c)

$\boldsymbol{x^{i-1} = Input}(F_{in}, \boldsymbol{N})$

$y1 = Dilated - Conv(x^{i-1}, F_{in,}F_{out}, k1, d1\ )$

$y1 = Dilated - Conv(y1, F_{out}, k2, d2\ )$

$x^i = Residual(x^{i-1}, y1, F_{in,}\ F_{out}, C_{res})$

$y0 = Inv(x^i, nC, c)$

**End Function**

The above pseudo code describes a dilated 1D-CNN for classifying epitope and non-epitope regions in protein sequences based on physicochemical features (ACC, DC, MBA, MA, GA). It begins by defining a Gaussian distribution function and measuring it using a cumulative distribution function. The 1D-CNN function takes parameters such as the number of amino acids (N), input and output features (F_in, F_out), kernel sizes (k1, k2), dilation rates (d1, d2), and number of classes (nC). It then performs dilated convolutions with different kernel sizes and dilation rates, calculates the residual connection between the input and output, and performs an inverse transformation to obtain the classification probabilities for epitope and non-epitope regions.

## III. RESILT ANALYSIS

The proposed Gaussian based Dilated 1D-CNN model was trained using Physicochemical Features of linear B-Cell protein sequence over the Zika and Dengue virus dataset. The implementation was carried out using core Python programming and the scikit-learn library. The experimental setup was conducted on Google Colab, utilizing a high-end GPU and 16 GB of RAM. To measure the performance and effectiveness of proposed Gaussian based Dilated 1D-CNN model for prediction and classification of linear B-cell Epitopes. Following evaluation parameters such as accuracy, precision, Recall and f1 score is used to meets the desired objectives of the study.

$$Accuracy\ = \frac{T\_P + T\_N}{T\_P + T\_N + F\_P + F\_N} \quad (15)$$

$$Precision\ = \frac{T\_P}{T\_P + F\_P} \quad (16)$$

$$Recall = \frac{T\_P}{T\_P + F\_N} \quad (17)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (18)$$

**Table 1: Classification Report of proposed model over the Zika virus dataset**

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.65 | 0.90 | 0.75 | 395 |
| 1 | 0.95 | 0.80 | 0.87 | 997 |
| Accuracy | | | 0.83 | 1392 |
| Micro Avg | 0.80 | 0.85 | 0.81 | 1392 |
| Weighted Avg | 0.87 | 0.83 | 0.84 | 1392 |

Table 1 shows the classification report of a proposed model for classifying epitope and non-epitope of linear B-cell over the Zika virus dataset. For class 0 (epitope), the model achieved a precision of 0.65, recall of 0.90, and F1-score of 0.75. For class 1 (non-epitope), the precision was higher at 0.95, recall was 0.80, and F1-score was 0.87. The accuracy of the model was 0.83. In the micro-average calculation across both classes, the precision was 0.80, recall was 0.85, and F1-score was 0.81, considering all instances. The weighted average, which considers class imbalance, resulted in a precision of 0.87, recall of 0.83, and F1-score of 0.84. This indicates that the model performed well, particularly in classifying class 1 instances.

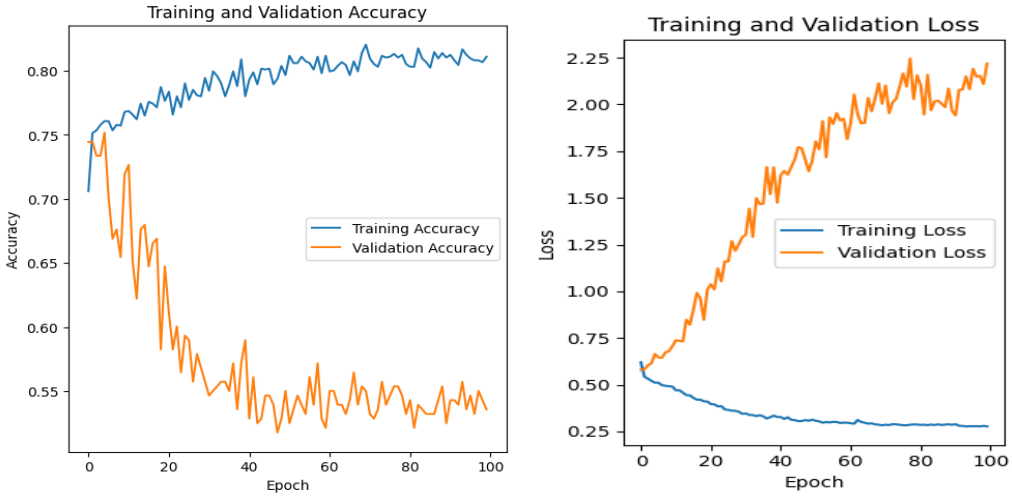


**Figure 3: Training and Validation Accuracy and Loss of model for Zika virus dataset**

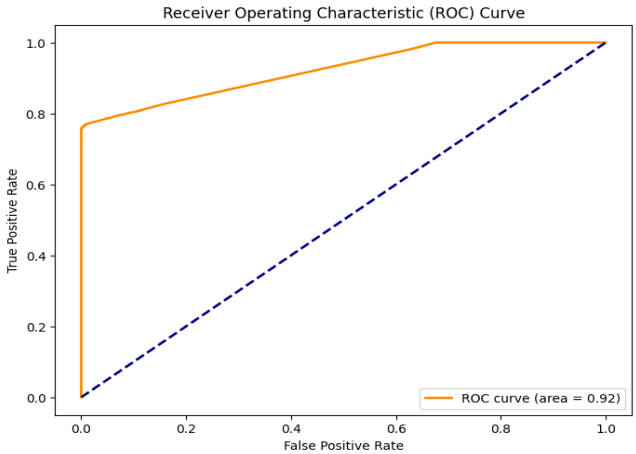


**Figure 4: ROC curve for Zika dataset**

The figure 3 shows the training and validation accuracy and loss of a model trained on the Zika virus dataset. The training accuracy steadily increases over 100 epochs. The training loss, which measures the difference between predicted and actual values during training, decreases over 100 epochs, indicating that the model is improving in its predictions. The figure 3 indicates that insights into how well the model is learning and generalizing from the

Zika data. Figure 4 shows the ROC curve that indicate the performance of a dilated 1D-CNN model trained on the Dengue virus dataset in terms of its true positive rate against the false positive rate across different thresholds. The area under the ROC curve (AUC) is 0.92, indicating that the model has good discriminative ability in distinguishing between positive and negative instances. A higher AUC value suggests that the model is better at classifying Epitopes and Non-Epitopes correctly across various thresholds. The ROC curve and its AUC of 0.92 indicate that the suggested model has a strong performance in classifying Epitopes and Non-Epitopes in the Zika virus dataset.

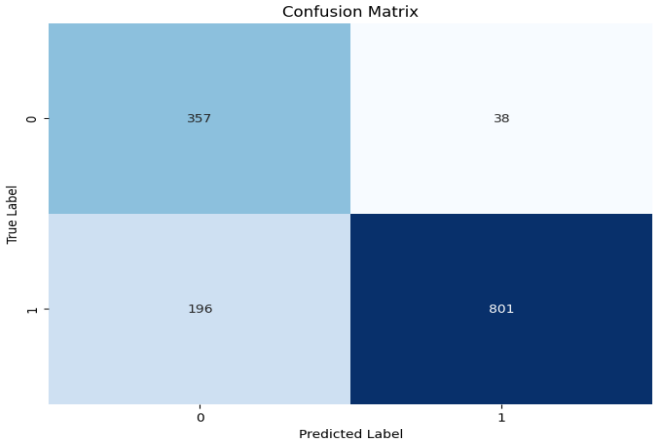


**Figure 5: Confusion Matrix for Zika Dataset**

Figure 5 shows the confusion matrix, indicates that the model correctly identified 196 epitope regions and 38 non-epitope regions, but it misclassified 801 non-epitope regions as epitope and 357 epitope regions as non-epitope over the Zika Dataset.

**Table 2: Classification Report of Proposed Model over the Dengue virus dataset**

| | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 | 0.70 | 0.86 | 0.77 | 1616 |
| 1 | 0.94 | 0.85 | 0.89 | 3999 |
| Accuracy | | | 0.85 | 5615 |
| Micro Avg | 0.82 | 0.85 | 0.83 | 5615 |
| Weighted Avg | 0.87 | 0.85 | 0.86 | 5615 |

Table 2 presents the classification report of a proposed model for classifying epitope and non-epitope of linear B-cell over the Dengue virus dataset. For class 0 (non-epitope), the model achieved a precision of 0.70, recall of 0.86, and F1-score of 0.77. For class 1 (non-epitope), the precision was higher at 0.94, recall was 0.85, and F1-score was 0.89. The overall accuracy of the model was 0.85.

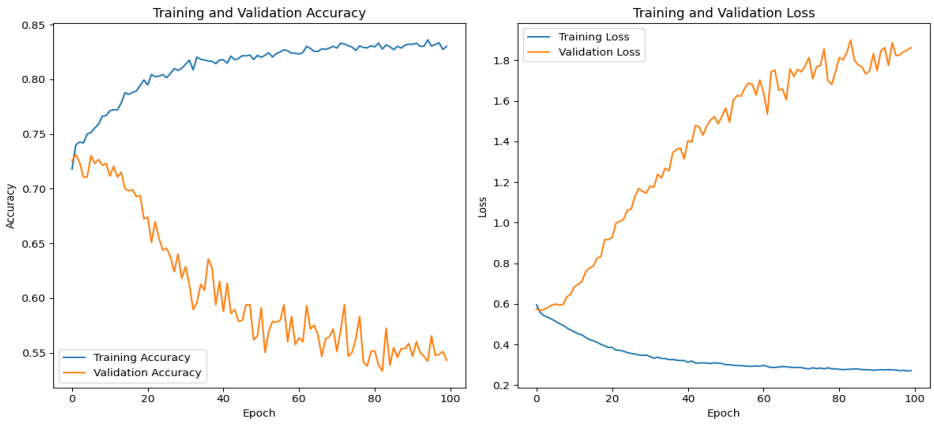


**Figure 6: Training and Validation Accuracy and Loss of model for Dengue virus dataset**

Figure 6 shows the training and validation accuracy and loss of a model trained on the Dengue virus dataset. The training accuracy steadily increases over 100 epochs. The training loss, which measures the difference between predicted and actual values during training, decreases over 100 epochs, indicating that the model is improving in its predictions. The figure 6 indicates that insights into how well the model is learning and generalizing from the Zika data.

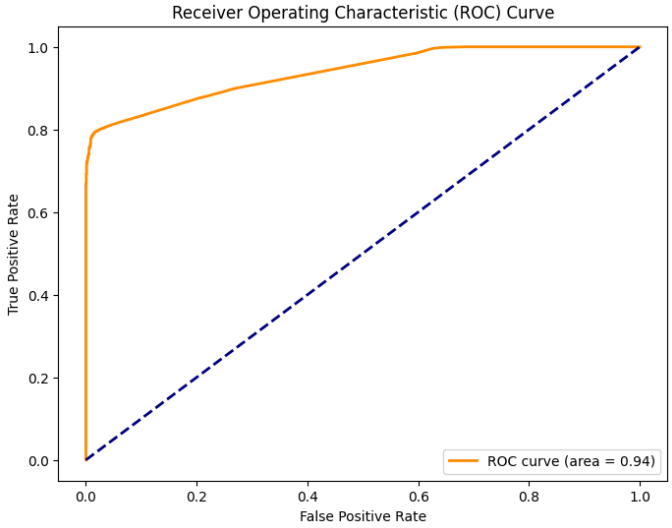


**Figure 7: ROC Curve of Model for Dengue Dataset**

Figure 7 shows the ROC curve that indicate the performance of a dilated 1D-CNN model trained on the Dengue virus dataset in terms of its true positive rate against the false positive rate across different thresholds. The area under the ROC curve (AUC) is 0.94, indicating that the model has good discriminative ability in distinguishing between positive and negative instances. A higher AUC value suggests that the model is better at classifying Epitopes and Non-Epitopes correctly across various thresholds. Overall, the ROC curve and its AUC of 0.94 indicate that the model has a strong performance in classifying Epitopes and Non-Epitopes in the Dengue virus dataset.

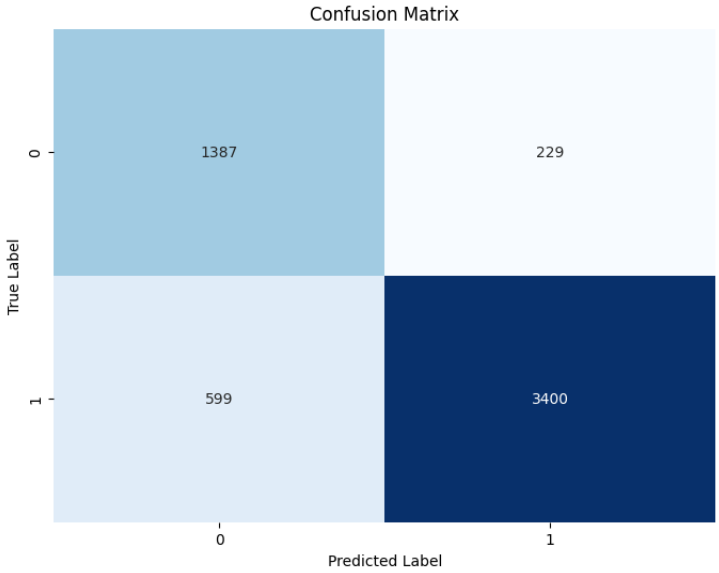


**Figure 8: Confusion Matrix for Dengue Dataset**

Figure 8 shows the confusion matrix, indicates that the model correctly identified 599 epitope regions and 229 non-epitope regions, but it misclassified 1387 non-epitope regions as epitope and 3400 epitope regions as non-epitope over the Dengue Dataset.

**Table 3: Comparative Analysis of Proposed Model with Existing Methods**

| Author | Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Liu F. et. al. (2024) [33] | iBCE-EL | 75.7 | 72.2 | 73.1 | 72.6 |
| Ras-Carmona A et. al. (2022) [34] | BLAST | 72.54 | 81.58 | 63.49 | - |
| Khanna D. et. al. (2020) [35] | APCpred | 76.6 | 70.00 | 82.00 | - |
| Maximilian Collatz et. al. (2020) [25] | Bepipred | 69.00 | 63.5 | 52.00 | - |
| Qi Y. et. al. (2023) [36] | DeepLBCEPred | 67.00 | 63.00 | 71.00 | - |
| **Proposed Dilated 1D-CNN** | **Zika** | **83.00** | **87.00** | **83.00** | **84.00** |
| | **Dengue** | **85.00** | **87.00** | **85.00** | **86.00** |

Table 3 presents a comparative analysis of the proposed model with existing methods for classifying epitope and non-epitope. The proposed Dilated 1D-CNN model outperforms existing methods in terms of accuracy, precision, recall, and F1-score. The proposed model achieves an accuracy of 83%, which is higher than the accuracies reported by Liu F. et al. (75.7%), Ras-Carmona A et. al. (72.54%), Khanna D. et. al. (76.6%), Maximilian Collatz et. al. (69.00%), and Qi Y. et. al. (67.00%). Similarly, the precision, recall, and F1-score of the proposed model are also higher compared to the existing methods, indicating that the proposed Dilated 1D-CNN model is more effective in classifying epitope and non-epitope regions in the dataset.

## IV. CONCLUSION AND FUTURE SCOPE

In this study, we addressed the challenge of accurately identifying linear B-cell epitopes in protein sequences related to Zika and Dengue viruses. We introduced a Gaussian-based dilated 1-D CNN model, aiming to improve epitope prediction, a critical step in peptide vaccine design and immunodiagnosis. Our model, trained on the Immune Epitope Database (IEDB) containing 1741 and 7020 epitopes for Zika and Dengue dataset, respectively, achieved high accuracy score of 83.00% and 85.00%, precision of 87.00%, recall of 83.00% and 85.00%, and an F1-score of 84.00% and 86.00% over the Zika and Dengue datasets. The results indicate the effectiveness of our model in classifying epitopes and non-epitopes, outperforming existing methods. For future research, we plan to enhance the model by exploring advanced deep learning pretrained techniques such as Resnet, LSTM, Mobilenet etc. and implements optimization techniques such as PSO, GA, Jelly Fish optimizer etc. to extract the optimal feature and also incorporating experimental data for validation, thereby further improving epitope prediction accuracy and expanding the model's applicability to other viruses and organisms.

## REFERENCES

[1] Murphy, K., and Weaver, C. (2012). “The induced responses of innate immunity” in Janeway's Immunobiology. 8th ed eds. J. Scobie, E. Lawrence, J. Moldovan, G. Lucas, B. Goatly, and M. Toledo (New York, NY: Garland Science), 75-125.

[2] Sable, N.P., Rathod, V.U. (2023). Rethinking Blockchain and Machine Learning for Resource-Constrained WSN. In: Neustein, A., Mahalle, P.N., Joshi, P., Shinde, G.R. (eds) AI, IoT, Big Data and Cloud Computing for Industry 4.0. Signals and Communication Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-29713-7_17.

[3] Shirai, H., Prades, C., Vita, R., Marcatili, P., Popovic, B., Xu, J., et al. (2014). Antibody informatics for drug discovery. Biochim Biophys Acta 1844, 2002–2015. doi: 10.1016/j.bbapap.2014.07.006.

[4] Nilesh P. Sable, Vijay U. Rathod, Parikshit N. Mahalle, Jayashri Bagade, Rajesh Phursule ; Internet of Things-based Smart Sensing Mechanism for Mining Applications, Industry 4.0 Convergence with AI, IoT, Big Data and Cloud Computing: Fundamentals, Challenges and Applications IoT and Big Data Analytics (2023) 4: 132. https://doi.org/10.2174/9789815179187123040012.

[5] Reta, D. H., Tessema, T. S., Ashenef, A. S., Desta, A. F., Labisso, W. L., Gizaw, S. T., Abay, S. M., Melka, D. S., & Reta, F. A. (2020). Molecular and Immunological Diagnostic Techniques of Medical Viruses. International journal of microbiology, 2020, 8832728. https://doi.org/10.1155/2020/8832728.

[6] V. U. Rathod and S. V. Gumaste, “Role of Deep Learning in Mobile Ad-hoc Networks”, IJRITCC, vol. 10, no. 2s, pp. 237–246, Dec. 2022.

[7] Stech, Marlitt, and Stefan Kubick. 2015. "Cell-Free Synthesis Meets Antibody Production: A Review" Antibodies 4, no. 1: 12-33. https://doi.org/10.3390/antib4010012.

[8] N. P. Sable, V. U. Rathod, P. N. Mahalle, and P. N. Railkar, “An Efficient and Reliable Data Transmission Service using Network Coding Algorithms in Peer-to-Peer Network”, IJRITCC, vol. 10, no. 1s, pp. 144–154, Dec. 2022.

[9] Parker JM, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry. 1986; 25:5425–32.

[10] N. P. Sable, R. Sonkamble, V. U. Rathod, S. Shirke, J. Y. Deshmukh, and G. T. Chavan, “Web3 Chain Authentication and Authorization Security Standard (CAA)”, IJRITCC, vol. 11, no. 5, pp. 70–76, May 2023.

[11] Emini EA, Hughes JV, Perlow DS, Boger J. Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. J. Virol. 1985;55:836–9.

[12] Vijay U. Rathod* & Shyamrao V. Gumaste, “Effect Of Deep Channel To Improve Performance On Mobile Ad-Hoc Networks”, J. Optoelectron. Laser, vol. 41, no. 7, pp. 754–756, Jul. 2022.

[13] Kolaskar AS, Tongaonkar PC. A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Lett. 1990;276:172–4.

[14] Rathod, V.U. and Gumaste, S.V., 2022. Role of Neural Network in Mobile Ad Hoc Networks for Mobility Prediction. International Journal of Communication Networks and Information Security, 14(1s), pp.153-166.

[15] Singh H, Ansari HR, Raghava GPS. Improved method for linear B-cell epitope prediction using Antigen’s primary sequence. PLoS One. 2013;8:e62216.

[16] Y. Mali, Vijay U. Rathod “A Comparative Analysis of Machine Learning Models for Soil Health Prediction and Crop Selection”, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), vol. 11, no. 10s, pp. 811–828, Aug. 2023.

[17] Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017;45:W24–9.

[18] N. P. Sable, V. U. Rathod, M. D. Salunke, H. B. Jadhav, R. S. Tambe, and S. R. Kothavle, “Enhancing Routing Performance in Software-Defined Wireless Sensor Networks through Reinforcement Learning”, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), vol. 11, no. 10s, pp. 73–83, Aug. 2023.

[19] Potocnakova L, Bhide M, Pulzova LB. An introduction to B-cell epitope mapping and in Silico epitope prediction. J Immunol Res. 2016; 2016:6760830.

[20] Zeng, Xincheng, Ganggang Bai, Chuance Sun, and Buyong Ma. 2023. "Recent Progress in Antibody Epitope Prediction" Antibodies 12, no. 3: 52. https://doi.org/10.3390/antib12030052.

[21] Vijay U. Rathod, Yogesh Mali, Nilesh Sable, Deepika Ajalkar, M. Bharathi, and N. Padmaja,” A Network-Centred Optimization Technique for Operative Target Selection”, Journal of Electrical Systems (JEs), vol. 19, no. 2s, pp. 87–96, 2023.

[22] Sato, K., Oide, M. & Nakasako, M. Prediction of hydrophilic and hydrophobic hydration structure of protein by neural network optimized using experimental data. Sci Rep 13, 2183 (2023). https://doi.org/10.1038/s41598-023-29442-x.

[23] Syrlybaeva, R., & Strauch, E. M. (2023). Deep learning of protein sequence design of protein-protein interactions. Bioinformatics (Oxford, England), 39(1), btac733. https://doi.org/10.1093/bioinformatics/btac733.

[24] Wang, Jingjing, Chang Chen, Ge Yao, Junjie Ding, Liangliang Wang, and Hui Jiang. 2023. "Intelligent Protein Design and Molecular Characterization Techniques: A Comprehensive Review" Molecules 28, no. 23: 7865. https://doi.org/10.3390/molecules28237865.

[25] Xia, Y. L., Li, W., Li, Y., Ji, X. L., Fu, Y. X., & Liu, S. Q. (2021). A Deep Learning Approach for Predicting Antigenic Variation of Influenza A H3N2. Computational and mathematical methods in medicine, 2021, 9997669. https://doi.org/10.1155/2021/9997669.

[26] Odorico, M. & Pellequer, J. L. (2003) BEPITOPE: Predicting the location of continuous epitopes and patterns in proteins. J. Mol. Recognit. 16(1), 20-22.

[27] N. P. Sable, V. U. Rathod, R. Sable and G. R. Shinde, "The Secure E-Wallet Powered by Blockchain and Distributed Ledger Technology," 2022 IEEE Pune Section International Conference (PuneCon), Pune, India, 2022, pp. 1-5, doi: 10.1109/PuneCon55413.2022.10014893.

[28] V. U. Rathod and S. V. Gumaste, "Role of Routing Protocol in Mobile Ad-Hoc Network for Performance of Mobility Models," 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), Lonavla, India, 2023, pp. 1-6, doi: 10.1109/I2CT57861.2023.10126390.

[29] Alix, A. J. P. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 18(3-4), 311–314.

[30] Bahai, A., Asgari, E., Mofrad, M. R. K., Kloetgen, A., & McHardy, A. C. (2021). EpitopeVec: linear epitope prediction using deep protein sequence embeddings. Bioinformatics (Oxford, England), 37(23), 4517–4525. https://doi.org/10.1093/bioinformatics/btab467.

[31] Kozlova EEG, Cerf L, Schneider FS, et al. Computational b-cell epitope identification and production of neutralizing murine antibodies against atroxlysin-i. Sci Rep. 2018;8(1):14904.

[32] N. P. Sable, V. U. Rathod, P. N. Mahalle and D. R. Birari, "A Multiple Stage Deep Learning Model for NID in MANETs," 2022 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 2022, pp. 1-6, doi: 10.1109/ESCI53509.2022.9758191.

[33] N. P. Sable, M. D. Salunke, V. U. Rathod and P. Dhotre, "Network for Cross-Disease Attention to the Severity of Diabetic Macular Edema and Joint Retinopathy," 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2022, pp. 1-7, doi: 10.1109/SMARTGENCON56628.2022.10083936.

[34] Han, W., Chen, N., Xu, X. et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning. Nat Commun 14, 3478 (2023). https://doi.org/10.1038/s41467-023-39199-6.

[35] Clifford, J. N., Høie, M. H., Deleuran, S., Peters, B., Nielsen, M., & Marcatili, P. (2022). BepiPred-3.0: Improved B-cell epitope prediction using protein language models. Protein science : a publication of the Protein Society, 31(12), e4497. https://doi.org/10.1002/pro.4497.

[36] Shen, W., Cao, Y., Cha, L., Zhang, X., Ying, X., Zhang, W., Ge, K., Li, W., & Zhong, L. (2015). Predicting linear B-cell epitopes using amino acid anchoring pair composition. BioData mining, 8, 14. https://doi.org/10.1186/s13040-015-0047-3.

[37] Sweredoski, M. J., & Baldi, P. (2009). COBEpro: a novel system for predicting continuous B-cell epitopes. Protein engineering, design & selection : PEDS, 22(3), 113–120. https://doi.org/10.1093/protein/gzn075.

[38] V. U. Rathod, N. P. Sable, N. N. Thorat and S. N. Ajani, "Deep Learning Techniques Using Lightweight Cryptography for IoT Based E-Healthcare System," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-5, doi: 10.1109/CONIT59222.2023.10205808.

[39] V. U. Rathod, Y. Mali, R. Sable, M. D. Salunke, S. Kolpe and D. S. Khemnar, "The Application of CNN Algorithm in COVID-19 Disease Prediction Utilising X-Ray Images," 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), Ravet IN, India, 2023, pp. 1-6, doi: 10.1109/ASIANCON58793.2023.10270221.

[40] Yao B, Zhang L, Liang S, Zhang C (2012) SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity. PLoS ONE 7(9): e45152. https://doi.org/10.1371/journal.pone.0045152.

[41] Maximilian Collatz, Florian Mock, Martin Hölzer, Emanuel Barth, Konrad Sachse, Manja Marz, (2020). EpiDope: A Deep neural network for linear B-cell epitope prediction, bioRxiv 2020.05.12.090019; doi: https://doi.org/10.1101/2020.05.12.090019.

[42] Akash Bahai, Ehsaneddin Asgari, Mohammad R K Mofrad, Andreas Kloetgen, Alice C McHardy, EpitopeVec: linear epitope prediction using deep protein sequence embeddings, Bioinformatics, Volume 37, Issue 23, December 2021, Pages 4517–4525, https://doi.org/10.1093/bioinformatics/btab467.

[43] Janghel RR, Raja R, Cengiz K, Raja H (2022) Next generation healthcare systems using soft computing techniques. CRC Press, New York.

[44] R. A. Mulla, Y. Mali, V. U. Rathod, R. S. Tambe, R. Shirbhate and R. Agnihotri, "Enhancing Query Performance Using Simultaneous Execution and Vertical Query Splitting," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-4, doi: 10.1109/ICCCNT56998.2023.10307920.

[45] Y. Mali, V. U. Rathod, R. S. Tambe, R. Shirbhate, D. Ajalkar and P. Sathawane, "Group-Based Framework for Large Files Downloading," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-4, doi: 10.1109/ICCCNT56998.2023.10308339.

[46] Y. Mali, V. U. Rathod, D. Ajalkar, D. S. Khemnar, S. Kolpe and S. Patil, "Role of Blockchain in Health Application using Blockchain Sharding," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10306760.

[47] Angaitkar, P., Janghel, R.R. & Sahu, T.P. DL-TCNN: Deep Learning-based Temporal Convolutional Neural Network for prediction of conformational B-cell epitopes. 3 Biotech 13, 297 (2023). https://doi.org/10.1007/s13205-023-03716-7.

[48] Angaitkar, P., Janghel, R.R. & Sahu, T.P. gHPCSO: Gaussian Distribution Based Hybrid Particle Cat Swarm Optimization for Linear B-cell Epitope Prediction. Int. j. inf. tecnol. 15, 2805–2818 (2023). https://doi.org/10.1007/s41870-023-01294-8.

[49] Cong Sun, Zhihao Yang, Leilei Su, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, Chemical–protein interaction extraction via Gaussian probability distribution and external biomedical knowledge, Bioinformatics, Volume 36, Issue 15, August 2020, Pages 4323-4330, https://doi.org/10.1093.

[50] Zhen, X., Chakraborty, R., Vogt, N., Bendlin, B. B., & Singh, V. (2019). Dilated Convolutional Neural Networks for Sequential Manifold-valued Data. Proceedings. IEEE International Conference on Computer Vision, 2019, 10620–10630. https://doi.org/10.1109/iccv.2019.01072.

[51] V. U. Rathod, Y. K. Mali, N. P. Sable, R. R. Rathod, M. N. Rathod and N. A. Rathod, "The Use of Blockchain Technology to Verify KYC Documents," 2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), New Raipur, India, 2023, pp. 1-6, doi: 10.1109/ICBDS58040.2023.10346414.

[52] Y. K. Mali, V. U. Rathod, M. D. Salunke, S. B. Satish, P. Dhamdhere and R. R. Rathod, "Role of IoT in Coal Miner Safety Helmets," 2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), Hamburg, Germany, 2023, pp. 221-225, doi: 10.1109/ICCCMLA58983.2023.10346793.

[53] Liu, F., Yuan, C., Chen, H. et al. Prediction of linear B-cell epitopes based on protein sequence features and BERT embeddings. Sci Rep 14, 2464 (2024). https://doi.org/10.1038/s41598-024-53028-w.

[54] Ras-Carmona, A., Lehmann, A.A., Lehmann, P.V. et al. Prediction of B cell epitopes in proteins using a novel sequence similarity-based method. Sci Rep 12, 13739 (2022). https://doi.org/10.1038/s41598-022-18021-1.

[55] M. D. Salunke, V. U. Rathod, Y. K. Mali, R. S. Tambe, A. A. Dange and S. R. Kothavle, "A Prediction and Classification Process for DDoS Attacks Using Machine Learning," 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2023, pp. 1-6, doi: 10.1109/ICCUBEA58933.2023.10392278.

[56] Khanna, D. and Rana, P.S. (2020), Improvement in prediction of antigenic epitopes using stacked generalisation: an ensemble approach. IET Syst. Biol., 14: 1-7. https://doi.org/10.1049/iet-syb.2018.5083.

[57] Qi Y, Zheng P and Huang G (2023) DeepLBCEPred: A Bi-LSTM and multi-scale CNN-based deep learning method for predicting linear B-cell epitopes. Front. Microbiol. 14:1117027. doi: 10.3389/fmicb.2023.1117027.

[58] V. U. Rathod and S. V. Gumaste, "An Effect on Mobile Ad-Hoc Networks for Load Balancing Through Adaptive Congestion Routing," 2023 International Conference on Integration of Computational Intelligent System (ICICIS), Pune, India, 2023, pp. 1-5, doi: 10.1109/ICICIS56802.2023.10430257.