Volume 4, Issue 5, 2026

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

BT-TRANS ACP: A BINARY TREE-GROWTH TRANSFORMER WITH HYBRID DEEP LEARNING FOR ANTICANCER PEPTIDE PREDICTION

Ali Ghulam[*1], Sikander Rahu[2], Dostdar Hussain[3], Arif Ahmed[4], Tarique Ali Brohi[5], Syed Saqib Hussain[6], Fatima-Tuz-Zahra[7] Zar Nawab Khan Swati[8], Waseem Ayaz[9], Mujeebu Rehman[10]

[*1]*Information Technology Center, Sindh Agriculture University Tandojam, Sindh Pakistan*
[2]*Department of Computer Science and Software Engineering, Jinnah University for Women, Karachi, Pakistan*
[3,6,8]*Department of Computer Science, Karakoram International University, Gilgit, Pakistan.*
[4]*Department of Computer Science, University of Sindh, Jamshoro, Sindh, Pakistan*
[5]*Department of Computer Science, SZABIST, University Hyderabad, Campus*
[7]*Comsats University Islamabad*
[9]*School of Electronic Engineering, Beijing University of Posts and Telecommunication, Beijing, China*
[10]*School of Computer Science and Engineering, Guilin University of Electronic Technology, China*

[*1]garahu@sau.edu.pk

DOI: https://doi.org/10.5281/zenodo.20034944

Keywords
Anticancer peptide prediction, Transformer-based encoding, Physicochemical descriptors

Article History
Received: 10 March 2026
Accepted: 20 April 2026
Published: 05 May 2026

Corresponding Author: *
Ali Ghulam

Abstract

*The issue of cancer remains a significant health challenge in the world because it causes serious mortalities. Despite the different traditional therapies and wet-laboratory technologies, which can be employed to treat cancerous cells, these can be faced with several setbacks, including extreme costs and severe side effects. Peptides have become a growing focus of interest over the past few years due to the high selectivity, reliable targeting ability, and low toxicity of peptides. After developing the existing computational tools, we offered a reasonable and functional framework, **BT-TransACP**, to the appropriate forecast of anticancer peptides. In this framework handles mathematically coded peptide sequences through the aid of an consideration-based ProtBERT-BFD encoder to encode semantics data with the addition of CTDT-derived structure features. A suboptimal set of properties of the integrated representation is then chosen based on binary tree growth (BTG) algorithm, which is a k-nearest neighbor algorithm. Our suggested framework is the **BT-TransACP**, a Binary Tree-Growth Transformer framework with a hybrid deep learning model to predict anticancer peptides. The fined feature vector is applied and a hybrid deep learning architecture that consists of CNN and LSTM layers are trained. The training accuracy of the model was 95.33% and the AUC was 0.97 with **BT-TransACP** model. The model was also tested on three separate datasets with the accuracies of 94.92, 92.26 and 91.16, respectively, to test its generalization capability. These findings indicate the high efficacy and effectiveness of BT-TransACP framework, which justifies the importance of the framework in research by scholars and in the development of pharmaceutical drugs.*

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

1. Introduction

Although cancer research and treatment have made significant progress in the past decades, cancer remains to be among the major causes of death in the world. The World Health Organization (WHO) estimates the prevalence of cancer as the cause of one-sixth of all reported mortality, and that in 2018, 9.6 million cancer-related deaths were reported [1,2]. The most common cancers that are highly fatal include the lung, colorectal, stomach, liver, and breast cancer with prostate and skin cancer being commonly diagnosed. The main characteristic feature of any type of cancer is the uncontrolled multiplication of cancer cells, which acquire mass, divide, and spread to healthy tissues and organs. Therefore, the major goal of cancer management is to inhibit cell proliferation of tumors and avoid metastases [3]. Today, chemotherapy and radiotherapy are the most common therapy treatments on the management of cancer [4]. This has brought about the demand to have new forms of therapy. Over the last ten years, anticancer peptides (ACPs) have been of great interest as therapeutic agents due to their high specificity to cancer cells and relatively reduced toxicity [5]. In spite of the promise, detection of ACPs by experimental means is still expensive, technically difficult and time-consuming. This has necessitated the need to establish effective and reliable approaches to the detection of ACP in order to enhance more effective cancer treatment procedures in the future.

Various machine-learning models were proposed during the past decade in predicting anticancer peptides (ACPs). Tyagi et al. have come up with AntiCP[6], which is a SVM-based method that employs both sequential and binary feature representations. After that, ACP sequences were encoded by Haji Sharifi et al. using local alignment information and PseAAC-derived features [7]. The concept of sequence-based predictors of ACP was then introduced by Chen et al. with iACP being the first sequence-based predictor [8]. Based on this advance, Akbar et al. implemented iACP-GAEnsC, an ensemble classifier that was optimized using a genetic algorithm [9], In the last ten years, hundreds of peptide-based therapeutic approaches to various cancer forms have been explored, with a number of them now in different stages of preclinical and clinical trials [9]. These developments underscore the increased importance of coming up with new anticancer peptides (ACPs) as alternative or supplementary to traditional cancer treatments. Nonetheless, it is usually expensive, cumbersome, and time consuming to conduct experimental discoveries and validate new ACPs. Therefore, computational methods that are based on sequences have proven necessary in order to enable high throughput screening and selection of promising ACP candidates before experimental synthesis. To fulfill this requirement, various computational prediction models, including the AntiCP, iACP and the approach suggested by Hajisharifi et al. (2014) have been presented to identify ACP [10-13]. Such methods are generally based on properties of primary peptide sequences, such as amino acid composition (AAC), binary profiles, dipeptide composition (DPC), and Chous pseudo-amino acid composition (PseAAC) which are then used to train support vector machine (SVM)-based classifiers. It is important to note that the majority of the existing approaches use a similar machine learning algorithm, and some of them, such as iACP or the model created by Hajisharifi et al., are trained on the same benchmark datasets. Although they have shown promising predictive accuracy, iACP and AntiCP are the two publicly available to assist in ACP-related studies [14-16]. In spite of the fact that the existing computational models provide important insights into the prediction of ACP, the predictive accuracy and predictive robustness could be enhanced considerably. In our paper, we suggest improved machine learning-based models with support vector machines (SVM) and random forests (RF) models, which we will call SVMacp and RFacp, respectively, and all of them, machine learning accentuated processes. The models incorporate various features of sequences such as AAC, DPC, atomic composition (ATC), and physicochemical properties (PCP). Critical tests on benchmark datasets indicate that the suggested methods are more effective than the existing methods in ACP prediction [17-18]. In addition, we also designed an online web-based tool to enable the larger

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

scientific community involved in the discovery of anticancer peptides and biomedical research.

Methods that store sequence-order information and can assist the extraction of the local intrinsic features are to be applied in order to enhance discriminative ability. In the past decade, deep learning has gained more and more power in ACP studies due to its ability to handle large datasets and the possibility to learn feature representations on its own[19]. Wu et al. created a deep learning model on the basis of word2vec sequence encoding [20, [21]. In as much as such embedding strategies can effectively encode the features of peptide sequences and still maintain the original data attributes, in most cases, they overlook previous biological information on amino acids [22]. Ahmed et al. developed ACP-MHCNN that was an integrative method that involved binary features, physicochemical descriptors, and sequential information [23]. The model suggested by Sun et al. combines the Bidirectional LSTM architecture with BERT transformers, this is ACP-BC[24]. Zhu et al. used a Bi-LSTM network in ACP-Check that was supplemented with five hand-crafted features and an ACP identification fully connected module [25, 26, 27], have also been created. Azim et al. proposed an ensemble random forest model called iACP-RF that uses binary profile features and frequency distributions of amino acids[28]. The ACP-ML model was based on a majority-voting ensemble model[29]. It produced peptide characteristics based on five sequence formulation tactics, class balancing based on SMOTENN and SMOTE Tomek, and two-level feature selection. Ten machine-learning classifiers were trained on the features that were selected and the final predictions were received via an ensemble voting mechanism. Karim et al. suggested ANNprob-ACPs, an integration of features obtained via the use of a probability-based combination of nine successive encoding schemes that were assessed on six machine-learning models[30]. The MA-PEP framework applied two encoding modules where the first one was on sequential information and the other on chemical descriptors and combined them with various attention mechanisms then trained a multilayer perception [31]. An overview of existing computational ACP models suggests that most of them are largely based on sequence-based residue encodings in addition to the fact that they seldom consider sequence-order structure in much detail. Moreover, most models have not sufficiently represented the contextual relation between peptide sequences and as such cannot learn embedded structural patterns. The selection of features is also another issue since most current methods rely on traditional filter-based methods which might miss a lot of informative features resulting in poor model performance. Moreover, a number of methods still rely on conventional learning algorithms, which can be characterized by significant computational complexity and lack of scalability. These findings highlight the necessity of better methodologies on a number of fronts: encoding, feature selection, model training, interpretability, generalization, computational efficiency, and predictive accuracy in general.

Over the last few years, one can find a variety of deep learning-based computational engines such as 2L-piRNADNN, iPredCNC [32], Deep-m5U [33], m6Aword2vec [34], and PSSM-Sumo [35,36] that have demonstrated high predictive performance even with comparatively small datasets. All these studies indicate the increasing power and performance of deep learning techniques in the field of bioinformatics. The general workflow of the proposed BT-TransACP is depicted in Figure.

We present a new feature selection algorithm, which is on a Binary Tree Growth (BTG) algorithm being operated by kNN to reduce the number of computations and isolate the most informative features.

- Numerous machine learning and deep learning models were tested, and the hybrid CNN+LSTM architecture recorded the highest predictive score of all the models.
- In order to evaluate the strength and extrapolation of the predictor we also tested BT-TransACP on three independent test sets, which validated the efficiency of the proposed method.

Volume 4, Issue 5, 2026

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

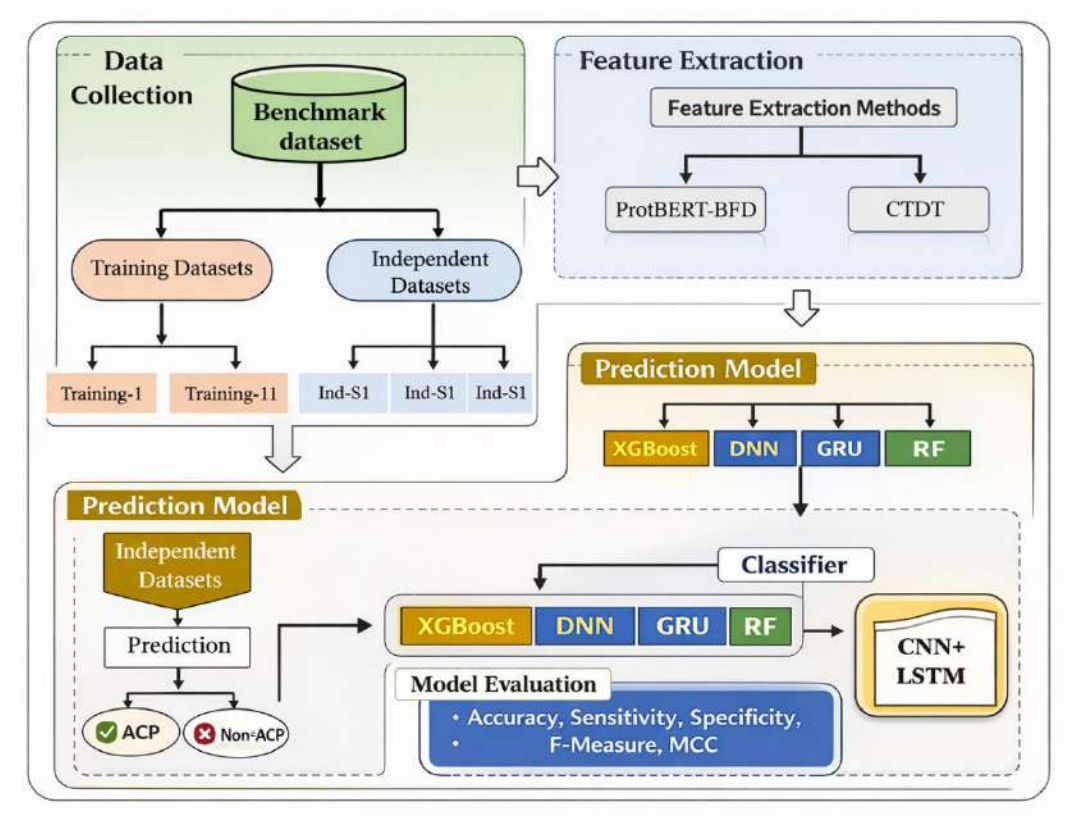

Figure 1. Schematic representation of the BT-TransACP computational framework

2. Materials and methods

2.1. Benchmark dataset

Selecting a proper benchmark dataset is a very important aspect of constructing any deep learning model in bioinformatics. In this work performance of our suggested method was evaluated with two sets of balanced training, based on which Training-I and Training-II have been prepared. The positive class (experimentally validated ACPs) and the negative class (antimicrobial peptides (AMPs)) make up the Training-I dataset. Existing models like ACPred-FL, iACP, ACPP and AntiCP, which were published before, were used to derive the AMPs and ensure that not one of the AMPs that were chosen had anticancer activity.

Following the elimination of peptides with anticancer and antimicrobial properties, there were 861 sequences in each of the classes. Training-I dataset was, therefore, constituted by 861 experimentally validated anticancer peptides (ACPs), and 861 non-ACPs. Training-II data set was developed based on ACPs associated with randomly selected peptides sequences as a negative sample. These non-ACP peptides have been obtained within the Swiss-Prot database. This created a balanced dataset comprising of 970 verified ACPs and 970 non-ACPs picked randomly. Ind-S3 dataset was based on Wei et al. and is comprised of 164 samples, 82 positive (ACP), and 82 negative sequences.

2.2. Feature extraction

Transformer-based protein Bidirectional Encoder Representations (ProtBERT-BFD)

Transformer-encoded encoders have in recent years proven to be extremely effective in bioinformatics, especially, in the intricate computational modeling of biological sequences. Bidirectional Encoder Representations coded by Transformer (BERT) architecture was initially created to encode the contextual and semantic

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

relationships amongst the tokens in sequences [48]. We adopt the successful experience of BERT-based models in the study of protein sequences, and instead of basing our study on the models from the literature, we deploy the ProtBERT-BFD model that employs self-attention based mechanisms to create highly informative and context-sensitive representations of the peptide sequences. In the present case, ProtBERT-BFD with pre-trained generic model was utilized in order to encode the peptide sequences into word-embedding vectors. The peptides were seen as sentences and tokenized into 200 units the initial token being a special character of CLS, which sums up the embedding data of all the tokens in the sequence. Sequences that were shorter than 200 tokens were padded with a special token called PAD to have a constant 200-token representation. Separate peptide samples were separated using a SEP token. Global average pooling was then used to convert every peptide, quantified by 200 tokens, to a 1024-dimensional feature vector. The training of the deep learning model was then fed with these vectors as input.

2.3. Composition/transition/distribution

In machine learning based classification problems, peptide sequences will need to be transformed into fixed-length feature vectors. The CTD method converts variable length sequences into fixed length counterparts, yet it retains biologically meaningful information. The features of CTD are Composition (CTDC), Transition (CTDT) and Distribution (CTDD). We have concentrated in this work on the Transition (T) descriptors that are measures of the frequency of the adjacent residues of various groups. As an example, to compute the effect of hydrophobicity property of a peptide of length L, the frequency of polar-neutral and neutral pairs of consecutive pairs is computed. The rest of the transitions, such as polar and neutral-hydrophobic, are calculated in the following way:

$$\boldsymbol{T(p,n) = \frac{N(p,n)+N(n,p)}{L-1}} \qquad (1)$$

N(p,n) and ***N(n,p)*** is the number of such pairs of residues within the peptide. The CTDT feature vectors are used to represent each peptide with ***13x3=39*** values.

Classification algorithms

In bioinformatics, credible training models are significant in evaluation of computation methods. In order to identify the most effective learning strategy to use in our framework, we compared a number of the classification methods.

2.4. Convolutional neural networks (CNN)

Convolutional neural networks are often applied to image recognition, object detection, video processing, etc. CNNs are able to acquire spatial feature hierarchies automatically on the input. All neurons in a given layer are only connected to the surrounding neurons in the prior layer and that way, CNNs are able to extract local features at a lower computational cost due to their weight sharing. This building assists them in learning aspects that do not change with translation.

The convolution operation is formulated based on the biological vision and it can be expressed as: Because the input is unchanged, the first component equally applies to both forward and backward propagation. Since the input value remains the same, the former component is applicable to forward and backward propagation alike.

$$\text{Convolution}(x)_i^k = \text{ReLU}(\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} W_{mn}^k x_{i+m,n}) \qquad (2)$$

In this case, $\boldsymbol{X}$ is the input matrix, $\boldsymbol{i}$ denotes the output location and $\boldsymbol{k}$ is the index of a filter. $\boldsymbol{W^k = (W_{mn}^k)_{M\times N}}$ commonly denotes the weights of the ***k-th*** kernel of size $\boldsymbol{M}$ x $\boldsymbol{N}$.

$$\text{ReLU}(x) = \max(0, x) \qquad (3)$$

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Reducing the dimensions of the data is done by pooling layers, which are beneficial because they are used to capture more general patterns of features and preserve translation invariance. A pooling operation is defined as:

$$\text{pooling}(X)_i^k = \max\left(x_{iM,k}, x_{iM+1,k}, \ldots, x_{iM+M-1,k}\right) \quad (4)$$

This is accomplished after pooling where CNNs take the resultant feature maps into fully connected layers. In the case of binary classification, the output is generated with the help of the sigmoid activation:

$$\textbf{sigmoid}(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-x}} \quad (5)$$

2.5. Long short-term memory (LSTM)

A specialized variant of recurrent neural network that attempts to overcome the shortcomings of traditional RNNs in the modeling of long sequences in bioinformatics is called Long Short-Term Memory networks. Protein sequences do frequently have critical dependencies which may be very distant, and LSTMs can remember these dependencies with their gated structure73,74. This architecture assists them in managing the amount of past data that is retained, updated or forgotten as the sequence is being handled.

LSTMs have input, forget and output gates to control the flow of information, instead of simple hidden-state updates. During every time step, the hidden state is changed depending on the old state and the new input:

Instead of simple hidden-state updates, LSTMs use input, forget, and output gates to regulate the flow of information.

$$\boldsymbol{h_t} = \boldsymbol{\partial}(\boldsymbol{h_{t-1}W_{hh}} + \boldsymbol{x_tW_{hx}} + \boldsymbol{b_{sh}}) \quad (6)$$

In this case, the hidden state at time t, x is the input feature at that time, Wx represents the weights of input features to the hidden layer, bsh is the bias term, and the activation function of the hidden layer.

The results of each time step are obtained by:

$$\boldsymbol{O_t} = \text{softmax}(\boldsymbol{W_{oh}h_t} + \boldsymbol{b_{sO}}) \quad (7)$$

Through this structure, LSTMs can effectively capture sequential and contextual relationships within peptide data, making them long-range interactions essential.

Table 1. Example-Based Demonstration of Mask Operation Methods.

Step	Vector Values
Original tree	1 0 1 0 1 0 1
Random finest tree	0 1 1 1 0 1 0
Mask operator applied	1 0 1 1 0 0 0
New tree after masking	0 0 1 1 1 0 1

is the output of the network at time t. The weights in the matrix W_{oh} are those that transform the hidden state to the output layer and b_{sO} is the bias term. The loss function is considered as.

$$\text{Loss} = -\frac{1}{S}\sum_{i=1}^{S}\sum_{t=1}^{T}\log\left(O_t^i\right) \quad (8)$$

In this case, S denotes the training samples, T denotes the sequence length, and $\boldsymbol{O_t^i}$ denotes the probability which is predicted to take the correct class of the i-th sequence at time t.

Volume 4, Issue 5, 2026

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

2.6. Proposed model architecture

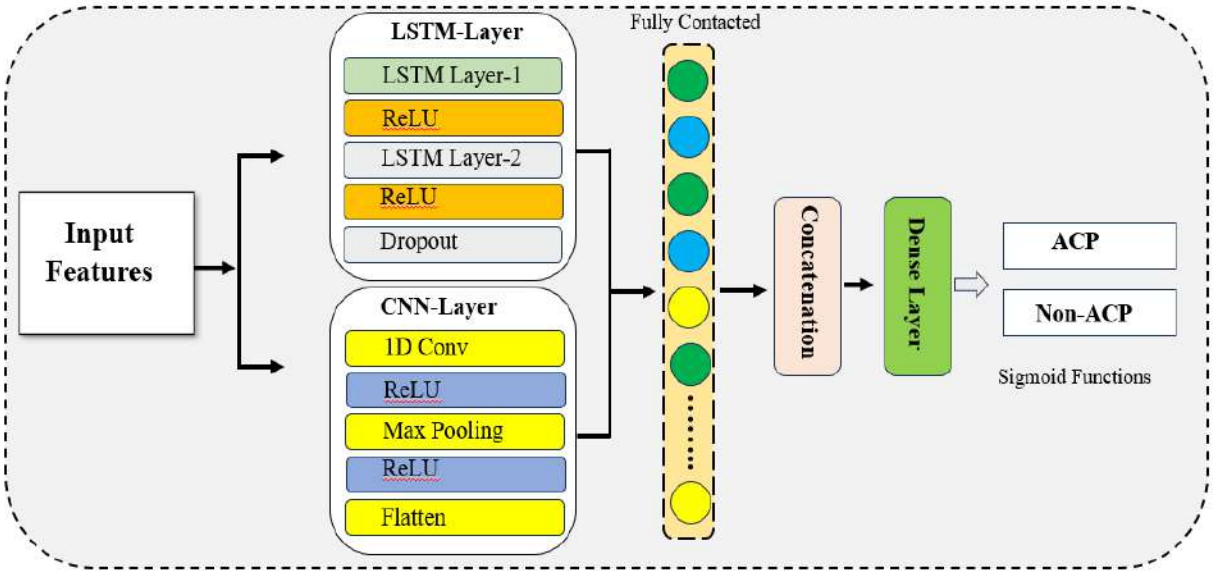

Figure 2. Overview of the BT-TransACP Model Architecture

Figure 2 shows the general design of the recommended framework. We developed a hybrid CNN-LSTM network to be able to capture local structure patterns of sequences and long-range dependencies in peptides. The model starts with an input layer which is fed with the pre-extracted feature representations. The sequence features are then learned using a one-dimensional convolutional branch with **32** filters whose size is **3 x 3** and the **ReLU activation** function. Simultaneously, an LSTM line is used to learn sequential and temporal relationships that are present in peptide sequences. The representations of features produced by the two branches are then combined and subjected to a fully connected layer. The mapped fused features are then projected to a final output layer of a single neuron that has the sigmoid activation function that yields the probability score during classification. Peptides whose output values are lower than 0.5 are termed as non-ACPs, whilst those with an equal number of 0.5 or more are termed as ACPs[37]. The convergence of the models was observed on the validation set and an early stopping criterion was used where applicable to avoid overfitting. The learning rate was set to 0.001, and Adam used as the optimization algorithm, and categorical cross-entropy as the loss function.

2.7. Evaluation metrics

In order to evaluate the performance of deep learning models, suitable evaluation measures are required. A confusion matrix in binary classification is a summary of the results based on the true positives (TP), the true negatives (TN), the false positives (FP), and the false negatives (FN). Along with the accuracy, we also tested the model based on sensitivity, specificity, Matthews Correlation Coefficient (MCC) and the Area Under the Receiver Operating Characteristic Curve (AUC) to give a total assessment on predictive performance. These evaluation metrics have mathematical formulations as shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (10)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (11)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (12)$$

In this case, TP equals true positives, TN equals

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

true negatives, FP equals false positives and FN equals false negatives. The MCC value is between -1 and +1 with +1 showing the perfect prediction, 0 showing the random prediction and -1 showing the complete disagreement of the prediction and the actual results.

3. Results and discussion

They evaluated the extrapolative capability of anticancer peptide (ACP) samples based on one and 10-fold cross-validation (CV) of single classifiers and two-way deep learning peptide models. In order to have strong and consistent results, the mean CV values were repeated 100 times with the stratified cross-validation process and in this case the training data were randomly assigned to the folds. At first, peptide samples were described with ProtBert-BFD-based transformer encoding with CTDT-generated composite physicochemical features. The ProtBert-BFD and CTDT features obtained were then combined to create a multi-perspective feature vector. In order to simplify the calculation of the multi-perspective vector, the Binary Tree Growth (BTG) feature selection method was used to select the highly relevant features. The proposed training models were used to evaluate all extracted vectors comprising of individual, hybrid, and selected feature sets. The results of the prediction based on the training and independent data are discussed in the following subsections.

Table 2. Configured Hyperparameters for the Proposed Hybrid Mode

Parameter	Values
Number of neurons in convolution layers	32, 64, 128
Number of neurons in LSTM layers	32, 64, 128
Activation function (Convolution and LSTM layers)	ReLU
Activation function (Dense layer)	Sigmoid
Number of epochs	40, 60
Batch size	32, 64
Dropout rate	0.1, 0.2, 0.3, 0.4
Learning rate	0.01, 0.001
Regularization	0.001
Optimizer	Adam

3.1. Predictive accuracy of various models using hybrid feature representation

The table 3 is a summary of the predictive capability of different classifiers on two training sets (Training-I and Training-II) with the proposed hybrid feature vector. The hybrid feature vector using CNN+LSTM gave the most overall accuracy (96.37%), AUC (0.98), and a significantly low sensitivity (21.66%), which may represent an imbalance in the detection of positive samples in Training-I. The XGBoost [38] and DNN classifiers showed the best performance with an accuracy of 92.81 and 92.83 respectively and with equal sensitivity and specificity values whereas the GRU and the Random Forest (RF) exhibited moderate performance with the accuracy of 80.37 and 82.71 respectively.

CNN+LSTM also demonstrated high accuracy (94.88) and MCC (0.93), backed by high sensitivity (97.44) and specificity (91.22) and has a strong performance in all the dataset on Training-II. XGBoost is also effective as it gave an accuracy of 93.44 and a high MCC of 0.91. DNN, GRU, and RF classifiers revealed a competitive performance with marginally smaller accuracy and MCC values, which suggests that the hybrid feature vector can serve numerous classification models and CNN+LSTM always performs better than the rest of the models in depicting sequence-contextual patterns.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Table 3. Performance Comparison of Different Classifiers on Hybrid Feature Vectors

Dataset	Encoding Method	Classifier	Acc(%)	Sens(%)	Spec (%)	MCC	AUC
Training-I	Hybrid Vector	Xgboost	92.81	94.21	91.32	0.85	0.97
		DNN	92.83	93.22	90.81	0.82	0.94
		GRU	80.37	87.45	77.37	0.54	0.89
		RF	82.71	81.73	83.56	0.65	0.92
		CNN+LSTM	**96.37**	**21.66**	**98.12**	**0.89**	**0.98**
Training-II	Hybrid Vector	Xgboost	93.44	91.41	98.55	0.91	0.95
		DNN	92.08	90.13	94.45	0.85	0.96
		GRU	88.65	86.67	94.67	0.89	0.94
		RF	92.22	91.22	92.26	0.83	0.95
		CNN+LSTM	**94.88**	**97.44**	**91.22**	**0.93**	**0.97**

3.2. Behavior training model 80 epochs.

Figure 3 presents the progress of the model in terms of its accuracy and loss during training and validation in 80 epochs. The accuracy curves can be explained by the fact that the model trained fast at the initial stages, and accuracy both training and validation increased at a steep rate, within the first ten epochs. The curves stabilize after this, which demonstrates that the model has an overall and stable performance level.

The trend in the loss curves is the same. The rate of both training and validation loss reduction is high in the initial few epochs, the hallmark of successful optimization. The training loss gets close to zero, whereas the validation loss reaches a stable point, which means that the model is generalizable and does not experience any obvious overfitting.

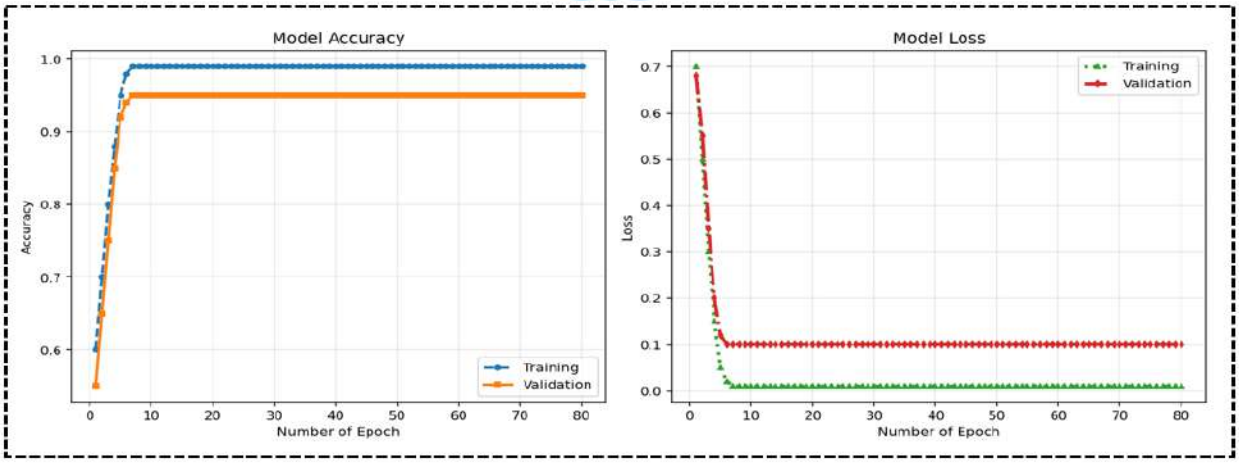

Figure 3. Evolution of Accuracy and Loss over Epochs for Training-I Dataset

3.3. Classification models analysis on training-I and training-II Datasets

Figure 4 shows the ROC curves of five machine-learning and deep-learning models on the Training-I and Training-II dataset. The ROC curve in both panels shows the quality of every classifier to distinguish anticancer peptides and non-anticancer peptides using various decision thresholds.

The CNN+LSTM model has the most effective performance in the Training-I data (left panel), which is the highest curve with an AUC of 0.98.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

XGBoost and DNN model are close to the AUC with values of 0.97 and 0.94. GRU and RF demonstrate relatively worse performance, and AUC values are 0.89 and 0.92. These findings suggest that the hybrid deep-learning architecture is more adequate to capture the underlying sequence patterns than the individual neural networks and the traditional machine-learning models.

The same trend is justified by Training-II dataset (right panel). CNN+LSTM is again improved with the best AUC of 0.97 and then XGBoost (0.95), DNN (0.96), GRU (0.94) and RF (0.95). In both datasets, the CNN+LSTM model has shown the highest steepest ROC curve and AUC, which validates the reliability and robustness of the model in the classification of anticancer peptides. The findings indicate that the combination of convolutional and recurrent layers contributes to better features learning and higher predictive power in comparison to the other considered approaches.

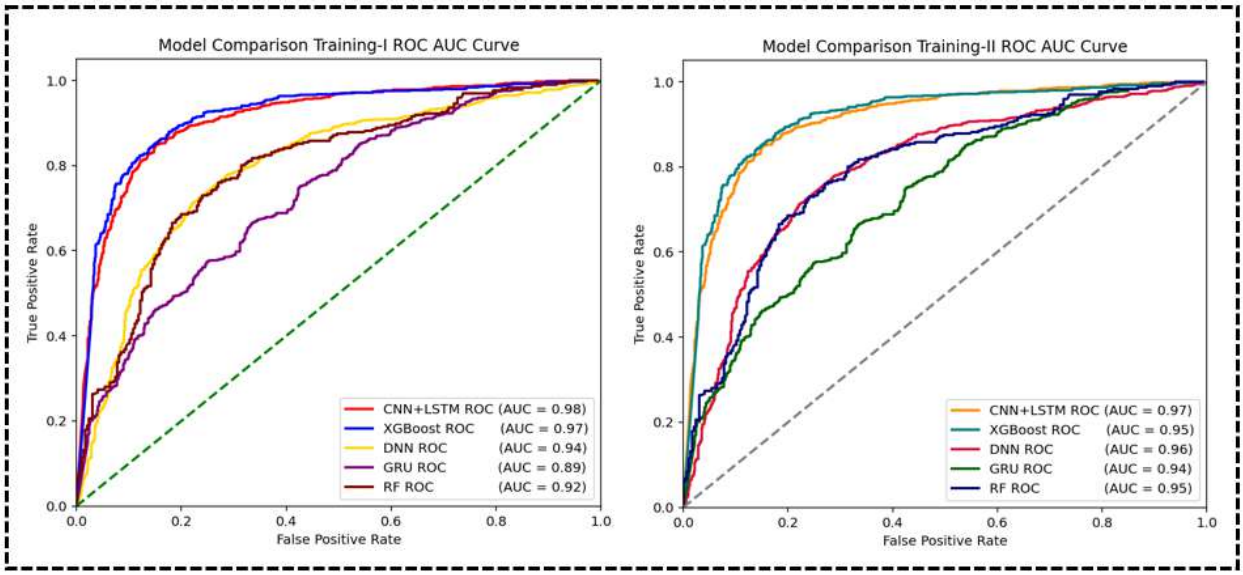

Figure 4. Comparative ROC Curves of the BT-TransACP Model for Training-I and Training-II Datasets

4. Conclusion

This paper presented a powerful and effective hybrid model, BT-TransACP to classify anticancer peptides (ACPs) and non-ACPs. The classical feature-encoding methods have significant shortcomings - such as the extraction of evolutionary features through large databases is time-consuming and many of the traditional sequential descriptors cannot retain information about sequence-order. In order to solve these problems, we used ProtBERTBFD, an attention based transformer to learn rich contextual and semantic representations and enhanced it with CTDT-derived structural and physicochemical descriptors in order to obtain important structural information. The optimized set of features was tested on a variety of training structures and the addition of a deep hybrid CNN + RNN structure significantly improved predictive capability. The suggested BT-TransACP model delivered 95.33% and 93.87% or more accuracies on the training-I and training-II datasets respectively. Its strength and formalizability was established by independent dataset validation, with 94.92, 92.26, and 91.16 and accuracies. These results indicate that BT-TransACP is more superior in terms of its performance compared to the current computer-based ACP prediction models. This solution is a promising instrument in drug discovery and academia research, which will help to create peptide-based medicines.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Authors Contributions.

AG contributed to the design of the methodology. SR and WZ performed the results analysis. DH and MR carried out the conceptual analysis of the study. AA was responsible for the comparative evaluation of the results. TA conducted the literature review and write-up. SSH contributed to the manuscript writing. FTZ and ZNKS jointly worked on manuscript preparation, writing, and English language editing.

REFERENCE

Lee, P. Y., Low, T. Y. & Jamal, R. in Advances in Clinical Chemistry Vol. 88 (ed Gregory S. Makowski) 67–89 (Elsevier, 2019).

Chhikara, B. S. & Parang, K. Global Cancer Statistics 2022: the trends projection analysis. Chem. Bio. Lett. 10, 451–451 (2023).

Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. Ca Cancer J. Clin. 73, 17–48 (2023).

Kuroda, K., Okumura, K., Isogai, H. & Isogai, E. The human cathelicidin antimicrobial peptide LL-37 and mimics are potential anticancer drugs. Front. Oncol. 5, 144 (2015).

Ghulam, A., Ali, F., Sikander, R., Ahmad, A., Ahmed, A., & Patil, S. (2022). ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemometrics and Intelligent Laboratory Systems*, *226*, 104589.

Tyagi, A. et al. In silico models for designing and discovering novel anticancer peptides. Sci. Rep. 3, 1–8 (2013).

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M. & Mohabatkar, H. Predicting anticancer peptides with Chou′ s pseudo amino acid composition and investigating their mutagenicity via Ame's test. J. Theor. Bio. 341, 34–40 (2014).

Chen, W., Ding, H., Feng, P., Lin, H. & Chou, K.-C. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 7, 16895 (2016).

Akbar, S., Hayat, M., Iqbal, M. & Jan, M. A. iACP-GAEnsC: Evolutionary genetic algorithm-based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif. Intell. Med. 79, 62–70 (2017).

Manavalan, B. et al. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget 8, 77121 (2017).

Xu, L., Liang, G., Wang, L. & Liao, C. A novel hybrid sequence-based model for identifying anticancer peptides. Genes 9, 158 (2018).

Kabir, M. et al. Intelligent computational method for discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information. Chemom. Intell. Lab. Syst. 182, 158–165 (2018).

Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V. & Shoombuatong, W. ACPred: a computational tool for the prediction and analysis of anticancer peptides. Molecules 24, 1973 (2019).

Ghulam, A., Sikander, R., Ali, F., Swati, Z. N. K., Unar, A., & Talpur, D. B. (2022). Accurate prediction of immunoglobulin proteins using machine learning model. *Informatics in Medicine Unlocked*, *29*, 100885.

Charoenkwan, P. et al. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci. Rep. 11, 3017 (2021).

Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N. & Raghava, G. P. AntiCP 2.0: an updated model for predicting anticancer peptides. Brief. Bioinform. 22(3), 153 (2021).

Akbar, S., Hayat, M., Tahir, M. & Chong, K. T. cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. IEEE Access 8, 131939–131948 (2020).

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Ghulam, A., Swati, Z. N. K., Ali, F., Tunio, S., Jabeen, N., & Iqbal, N. (2023). DeepImmuno-PSSM: Identification of Immunoglobulin based on Deep learning and PSSM-Pro les.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. J. Chem. Inf. Model. 55, 263–274 (2015).

Vazhayil, A. & KP, S. DeepProteomics: protein family classification using Shallow and Deep Networks. arXiv preprint arXiv:1809.04461 (2018).

Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R. & Atalay, V. DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. Sci. Rep. 9, 7344 (2019).

Shahid, Hayat, M., Alghamdi, W., Akbar, S., Raza, A., Kadir, R. A., & Sarker, M. R. (2025). pACP-HybDeep: predicting anticancer peptides using binary tree growth based transformer and structural feature encoding with deep-hybrid learning. *Scientific Reports*, *15*(1), 565.

Yi, H.-C. et al. ACP-DL: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation. Mol. Therapy-Nucleic Acids 17, 1–9 (2019).

Sun, M., Hu, H., Pang, W. & Zhou, Y. ACP-BC: A model for accurate identification of anticancer peptides based on fusion features of bidirectional long Short-Term memory and chemically derived information. Int. J. Mol. Sci. 24, 15447 (2023)

Zhu, L., Ye, C., Hu, X., Yang, S. & Zhu, C. ACP-check: An anticancer peptide prediction model based on bidirectional long shortterm memory and multi-features fusion strategy. Comput. Biol. Med. 148, 105868 (2022). 29.

Wu, X., Zeng, W., Lin, F., Xu, P. & Li, X. Anticancer peptide prediction via multi-kernel CNN and attention model. Front. Genet. 13, 887894 (2022).

Han, B., Zhao, N., Zeng, C., Mu, Z. & Gong, X. ACPred-BMF: bidirectional LSTM with multiple feature representations for explainable anticancer peptide prediction. Sci. Rep. 12, 21915 (2022)

Azim, S. M. et al. Accurately predicting anticancer peptide using an ensemble of heterogeneously trained classifiers. Inf. Med. Unlocked 42, 101348. https://doi.org/10.1016/j.imu.2023.101348 (2023).

Bian, J. et al. ACP-ML: A sequence-based method for anticancer peptide prediction. Comput. Biol. Med. 170, 108063. https://doi.org/10.1016/j.compbiomed.2024.108063 (2024).

Karim, T., Shaon, M. S. H., Sultan, M. F., Hasan, M. Z. & Kafy, A. A. ANNprob-ACPs: A novel anticancer peptide identifier based on probabilistic feature fusion approach. Comput. Biol. Med. 169, 107915. https://doi.org/10.1016/j.compbiomed.2023.107915 (2024).

Liang, X., Zhao, H. & Wang, J. MA-PEP: A novel anticancer peptide prediction framework with multimodal feature fusion based on attention mechanism. Prot. Sci. 33, e4966 (2024).

Khan, S. et al. A Two-Level computation model based on deep learning algorithm for identification of piRNA and their functions via chou's 5-steps rule. Int. J. Pept. Res. Therapeutics 26, 795–809. https://doi.org/10.1007/s10989-019-09887-3 (2020).

Khan, Z. U., Ali, F., Ahmad, I., Hayat, M. & Pi, D. iPredCNC: computational prediction model for cancerlectins and noncancerlectins using novel cascade features subset selection. Chem. Int. Lab. Syst. 195, 103876 (2019).

Noor, S. et al. Deep-m5U: a deep learning-based approach for RNA 5-methyluridine modification prediction using optimized feature integration. BMC Bioinform. 25, 360. https://doi.org/10.1186/s12859-024-05978-1 (2024).

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Tahir, M., Hayat, M. & Chong, K. T. Prediction of N6-methyladenosine sites using convolution neural network model based on distributed feature representations. Neural Netw. 129, 385–391 (2020).

Khan, S., Al Qahtani, S. A., Noor, S. & Ahmad, N. PSSM-Sumo: deep learning based intelligent model for prediction of sumoylation sites using discriminative features. BMC Bioinform. 25, 284. https://doi.org/10.1186/s12859-024-05917-0 (2024).

Sikander, R., Ghulam, A., Hassan, J., Rehman, L., Jabeen, N., & Iqbal, N. (2023). Identification of cancerlectin proteins using hyperparameter optimization in deep learning and DDE profiles. *Mehran University Research Journal Of Engineering & Technology*, *42*(4), 28-40.

Rahu, S., Ghulam, A., Farman, A., Talpur, D. B., Talpur, M. S. H., Saba, E., ... & Tunio, S. (2022). 'UBI-XGB: IDENTIFICATION OF UBIQUITIN PROTEINS USING MACHINE LEARNING MODEL'. *Journal of Mountain Area Research*, *8*.