JCBI
Journal of Computing & Biomedical Informatics Volume 03 Issue 02
ISSN: 2710 - 1606 2022

Research Article
https://doi.org/10.56979/302/2022/81

Anticancer Peptides Prediction: A Deep Learning Approach

Hassan Kaleem[1,*], Sundas Rukhsar[1] and Muhammad Noman Khalid[2]

[1]SQL Consultancy Ltd, 9 Frances Street Crewe, England.
[2]Medicine and Surgery, Allama Iqbal Medical College, Lahore, Pakistan.
[*]Corresponding Author: Hassan Kaleem. Email: hassanrao.hr@gmail.com

Received: July 28, 2022 **Accepted:** September 13, 2022 **Published:** September 27, 2022

Abstract: Anticancer peptides play a vital role in the treatment of cancer, due to that it has gained a lot of attention. Several machine learning and deep learning algorithms were developed for the prediction of anticancer peptides. Machine learning algorithms involves features extraction from the dataset and then model is trained to make predictions. In machine learning algorithms features extraction and the training of the model takes a lot of time and efforts, this is a complex process for biologists and biochemists. On the other hand deep learning algorithms require a large amount of dataset for training and accurate predictions. This study has proposed a deep learning algorithm which can be trained on smaller dataset because it uses hyperparameter optimization framework for the accurate predictions of anticancer peptides. The deep learning model has outperformed all the other algorithms and achieved the optimal 99% Acc and 0.982 MCC on Main dataset, 98% Acc and 0.972 MCC on Alternative dataset. The code is available at Github for validation purposes [33].

Keywords: Deep Learning; Anticancer Peptides; KerasTuner; Hyperparameter Optimization.

1. Introduction

Anticancer peptides (ACPs) are small groups of amino Acids (10-60 mostly) joined by peptide bonds showing discriminating and lethal properties towards cancer cells. Due to their innate properties like high perforation, high selectivity and ease of moderation and low manufacturing costs synthetic peptide based medicines and immunogens represent an encouraging class/group of therapeutic agents which shows decreased drug resistance and also suppress angiogenesis of cancer cells. One such example is their increasing use in treatment of hepatocellular carcinoma (HCC). ACPs that are specially formulated can improve affinity, selectivity, and with consistency for more tumor cell eradication. In order to perform cell permeability, the influence of amino acid residues on the anti-cancer action of ACPs is based on cationic, hydrophobic, and amphiphilic qualities with double stranded structure. Particularly, negatively charged amino acid residues (such as glutamic and aspartic acids) perform anti-accretion activity against tumor/cancer cells, whereas positively charged amino acid residues (such as lysine, arginine, and histidine) can destroy and enter cancer cell membrane to cause cytotoxicity. Additionally, hydrophobic amino acid residues, such as tryptophan, phenylalanine, and tyrosine, exert their impact on the cytotoxic action of malignancy. Additionally, the secondary structure of ACPs, which are composed of both positively charged and hydrophobic amino acids, is crucial for the interlinkage of peptides with cancer cell membranes. [1-5]. Artificial intelligence is when machines can learn and do tasks that are not possible without human involvement. Machine learning is when machines can obtain skills and learn things. Deep learning is a branch of machine learning which uses Artificial Neural Network (ANN) to obtain skills from dataset. In ML, features should be extracted from the data so it can be passed to machine and can obtain skills. In deep learning, we have to pass raw data and it generates features by itself. Neural network is a kind of network which is inspired

by human brain [6]. In deep learning model is trained to perform specific tasks by taking raw inputs from labeled data. The dataset could contain sound, text, sequences or images. The results that are achieved by deep learning were never possible before. These trained model of deep learning can even exceed human level performance due to ultra-high accuracy. It is because deep learning models uses multiple layered neural network architecture and large labeled dataset [7]. Following Figure shows the architecture of deep learning model.

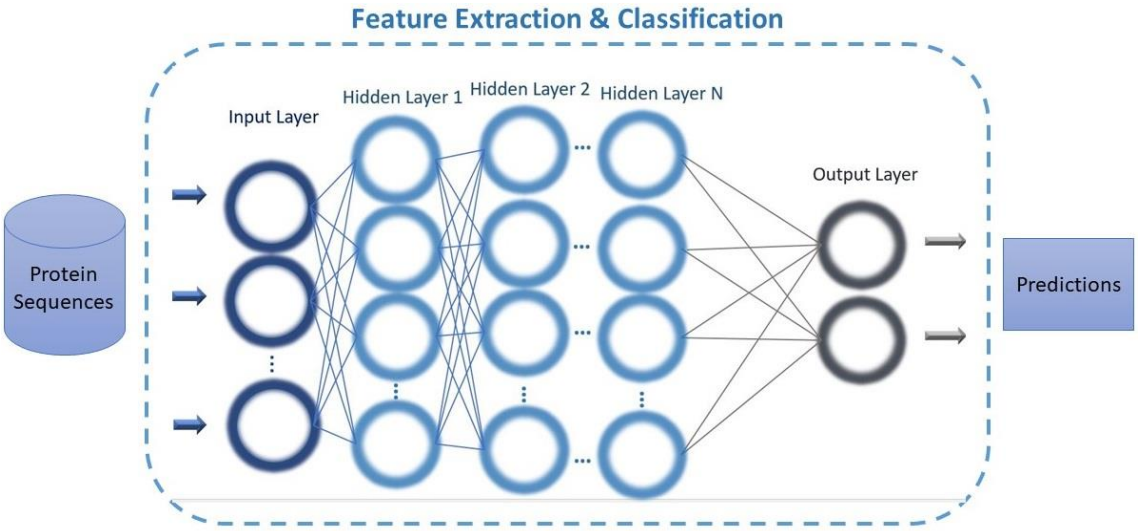

Figure 1. General Architecture of Deep Learning Model.

Several machine learning and deep learning algorithms are developed to classify ACP from non-ACP. Ankur et al conducted an experiment on prediction of cell penetrating peptides [8]. Support vector machine (SVM) algorithm is used for the prediction of cell penetrating peptides. The dataset used in the study is consist of 708 peptides. During the experiment several motifs are identified in cell penetrating peptides, based on this a hybrid model is developed for the prediction. The model achieved 81.31% accuracy and 0.63 MCC (Mathew's Correlation Coefficient). Yu et al conducted an experiment on prediction of therapeutic peptides by using novel features encoding and adaptive feature learning [9]. The dataset which is used in the experiment is consist on eight therapeutic peptides. RandomForest classifier is used for the prediction. The model achieved 98% accuracy for ABP therapeutic peptide because its training samples are 1600 and both the classes are in balance. Shaherin et al conducted a study on Machine intelligence in peptide therapeutics [10]. The study analyzed the existing prediction algorithms by using well-constructed dataset. Their results are compared their accuracy and prediction scores. The study provides a brief explanation on how to build an accurate algorithms for anticancer peptides prediction. Atul et al conducted an experiment on anticancer peptides by using In Silico model (AntiCP) [11]. The dataset which is used in the experiment was derived from SwisProt. Multiple feature extraction method were used in the experiment like Amino Acid Composition, dipeptides composition and binary profile. Support vector machine (SVM) algorithm is used for the prediction. Binary profile based SVM model achieved maximum 91.44% accuracy and 0.83 MCC. Zohre et al conducted a study on the prediction of anticancer peptides by using Chou's amino acid composition [12]. Support vector machine (SVM) algorithm is used in the experiment. The model achieved maximum accuracy of 89.7% based on local alignment kernel method. Saravanan et al conducted a study on the prediction of anticancer peptides (ACPP) [13]. Multiple datasets were used in the study including a newly developed dataset. SVM classifier was used for the prediction, the model achieved 96% accuracy and 0.97 MCC. Wei et al conducted an experiment on identifying the anticancer peptides by using a sequence based tool (iACP) [14]. The dataset which is used in the experiment consist on 150 positive and 150 negative samples for anticancer peptides. G-gap dipeptide was used for features extraction and SVM algorithm was used for the prediction. The model achieved 92.67% accuracy and 0.85 MCC on independent dataset. Feng et al conducted an experiment on identifying the anticancer peptides by using improved hybrid composition [15]. Amino acid composition (AAC), average chemical shifts (acACS) and reduced amino acid composition (RAAC) were used to extract the features, after that SVM algorithm was used for

the prediction. The jackknief testing achieved 93.61% accuracy. To reach that accuracy the all the extracted features were fused together. Shahid et al conducted an experiment on anticancer peptides prediction by using hybrid feature space (iACP-GAEnsC) [16]. Amino acid composition (AAC), dipeptides composition (DPC) and reduced amino acid composition (RAAC) were used to extract features. The features were tested separately as well as fused by using genetic algorithm based ensemble classification. The model achieved 96.45\% accuracy. Muhammad et al conducted an experiment on the discrimination of anticancer peptides by incorporating sequential and evolutionary profiles information (TargetACP) [17]. SVM algorithm was used in the experiment. After jackknife cross-validation on the benchmark dataset, the model achieved 98.78% accuracy. The model achieved 94.66% accuracy on independent dataset. Nalini et al conducted an experiment on prediction and analysis of anticancer peptides by using a computational tool (ACPred) [18]. SVM and RandomForest classifiers were used in the experiment. After jackknife cross-validation testing, the model achieved 95.61% accuracy on identifying the anticancer peptides. Balachandran et al conducted an experiment on anticancer peptides by using machine learning algorithms (MLACP) [19]. Amino acid composition (AAC), dipeptide composition (DPC), atomic composition (AC), and physicochemical properties were used for the extraction of the features. SVM and RandomForest classifiers were used for the prediction. The model achieved 88.7% accuracy and 0.78 MCC. Chuanyan et al conducted an experiment on the prediction of therapeutic peptides by using deep learning and word2vec (PTPD) [20]. Two datasets were used in the experiment, independent dataset was consist of 138 ACPs and 206 non-ACPs while virulent protein dataset was consist of 225 ACPs and 2250 random protein from the SwisProt. The model achieved 96% accuracy on independent anticancer peptide dataset and 94% on virulent protein dataset. Leyi et al conducted an experiment on the prediction of anticancer peptides by using sequence based predictor (ACPred-FL) [21]. SVM algorithm was used in the experiment and a new dataset was constructed from identifying the ACPs. The model achieved significantly higher accuracy on both 10-fold cross-validation testing and independent testing. Leyi et al conducted another experiment on prediction of therapeutic peptides by using adaptive feature learning (PEPred-Suite) [22]. Eight benchmark datasets were collected and used in the study. RandomForest classifier was used for the prediction, eight models of RandomForest classifier were trained by using the learnt representative features. The model achieved best AUC of 93.6% on AVP dataset. Hai et al conducted an experiment on anticancer peptides using long short-term memory (LSTM) deep learning based classifier (ACP-DL) [23]. Two dataset ACP740 and ACP240 were used in the experiment. On 5-Fold cross-validation testing the LSTM model achieved 0.69 MCC for ACP740 and 0.71 for ACP240. Bing et al conducted an experiment on the prediction of anticancer peptides by using fusing multi-view information (ACPred-Fuse) [24]. Machine learning algorithm was used for the prediction. Results of multiview features and existing feature descriptors were compared together, the multiview can better discriminate the characteristics of ACPs. Piyush et al conducted an experiment on anticancer peptides prediction using updated model (AntiCP 2.0) [25]. Various features were used for the machine learning models. Two datasets were used in the experiment, Main and alternative datasets. For Main dataset. Dipeptide composition (DPC) achieved maximum 0.51 MCC for ExtraTree classifier and for alternative dataset, amino acid composition (AAC) based ExtraTree classifier achieved 0.80 MCC. Phasit et al conducted a study on anticancer peptides by using novel flexing scoring card method iACP-FSCM [26]. Benchmark dataset which was used in AntiCP 2.0 [25], was used in the experiment. Three features extraction method were used in the experiment, amino acid composition (AAC), dipeptide composition and composition of terminal region. After features extraction the model was trained to make prediction on anticancer peptides. On independent testing, the classifier achieved the accuracy of 0.825% on Main dataset and 0.910% on Alternative dataset. Our study has proposed a deep learning based approach which also outperformed STALLION [27]. STALLION is model that is used to predict the Kace sites.

2. Materials and Methods

The benchmark dataset is collected from AntiCP 2.0 [25]. The benchmark dataset is used for fair comparison of the proposed model. The dataset consist on two parts, main dataset (861 positive samples and 861 negative samples) is used to develop the model and alternative dataset (970 positive sequences and 970 negative sequences) is used to evaluate the model's performance. Each dataset is consist on two parts, training and testing. Main dataset contains 1378 (689 positive sequences and 689 negative sequences) sequences for training and 344 (172 positive sequences and 172 negative sequences) sequences for testing

and alternative dataset contains 1552 (776 positive sequences and 776 negative sequences) sequences for training and 388 (194 positive sequences and 194 negative sequences) sequences for testing. The dataset is available at (https://webs.iiitd.edu.in/raghava/anticp2/download.php).

In first experiment, machine learning algorithm were used for anticancer peptides prediction. Machine learning algorithms require features extraction from the dataset. IFeature [28] live server is used for features extraction from protein sequences. Multiple features extraction methods were used to extract the features from the dataset. After that Lazypredict [29] was used for prediction. Lazypredict is an open source python library which uses 40 machine learning algorithms for classification. Top two algorithms were selected from Lazypredict, the experiment was conducted with separate features as well as fused features. Data fusion is a process in which all the extracted features are combined together and passed to the Machine Learning classifier. After data fusion the model performed really well but failed to outperform iACP-FCSM. After that deep learning model was used in the experiment, deep learning model outperformed all the other predictors in anticancer peptides prediction.

KerasTuner [30] is used to tune the model and which uses TensorFlow at the back-end. KerasTuner is hyper-parameter optimization framework which finds the best hyperparameter values for your model. In KerasTuner multiple units are passed to the model, so it can train and test the model according to the given units. Deep learning models takes raw input sequences, calculates the features by itself and then it train itself to make predictions. Following Figure 2 shows the proposed deep learning model's architecture.

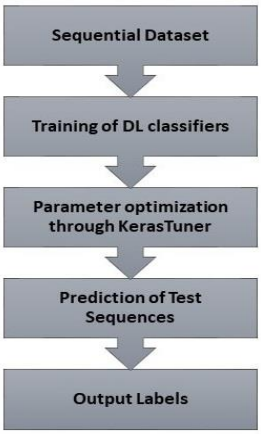

Figure 2. Proposed Deep Learning Model Architecture.

2.1. Computational Environment

Google Colaboratory is used in the development of the suggested strategy. It is an online Jupyter Notebook platform that uses the Python programming language and is cloud-based. Because python libraries are already available on the server and we don't need to manually install them, we utilized Google Colaboratory. The graphics processing unit (GPU) offered by Google Colaboratory also helps deep learning algorithms train and test very quickly.2.2. Architecture of Proposed Deep Learning model, than best performance is selected based on those given units. Table 1 contain parameters used in KerasTuner.

Table 1. Parameters used in KerasTuner

Maximum Features	**500**
Activation Function Hidden Layer	Relu
Activation Function Output Layer	Sigmoid
Units	2, 4, 8, 16, 32, 64, 128
Epochs	20
Batch size	32
Type	Random Search
Objective	Validation Accuracy
Max Trails	10

2, 4, 8, 16, 32, 64 and 128 units were passed to KerasTuner and it is configured to find 10 random tuples of the hidden units and activation function. For each trail and execution, KerasTuner will fit the model

with 20 epochs as configured in the script. Following Figure 3 shows the training accuracy of the proposed deep learning model. The model's training accuracy changes as the units are changed.

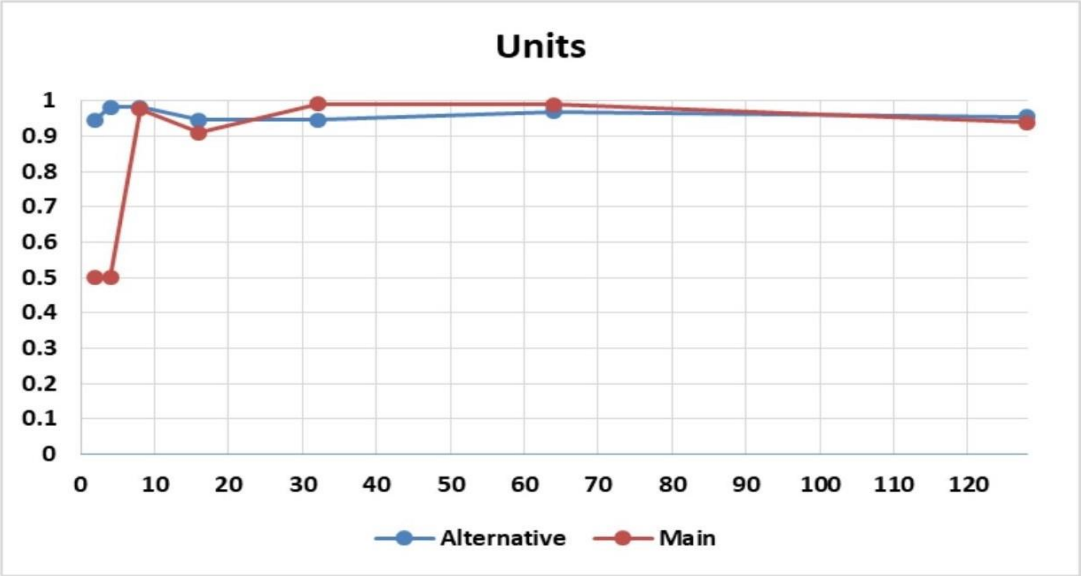

Figure 3. Training Accuracy of KerasTuner.

2.2. Architecture of Proposed Deep Learning model

Five performance measurements that were widely employed in other studies [31] [32] were applied to evaluate the model. These measurements includes following, Matthew's correlation coefficient (MCC), Area under the curve (AUC), Sensitivity (Sn), Specificity (Sp) and Accuracy (Acc).

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

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

$$AUC = Sn\,(1 - Sp)$$

True Positives are denoted by TP, True Negatives by TN, False Positives by FP and False Negative by FN.

3. Results

In the first part, machine learning algorithms were used for the prediction of anticancer peptides. The fused features were passed to LGBM and ExtraTree classifier. On independent testing, LGBM classifier achieved 0.64 MCC on Main dataset and 0.67 MCC on Alternative dataset. After hyper-parameter optimization of machine learning classifier, LGBM classifier achieved 0.65 MCC on Main dataset and 0.69 MCC on Alternative dataset. In the second experiment, the datasets are split into training and testing samples for Main and Alternative datasets. 5 fold cross-validation testing is applied on the datasets to make sure every portion of the datasets is passed to the deep learning model for training and testing. After that mean is calculated.

The proposed deep learning model outperformed iACP-FSCM on both main and alternative datasets. The following Table 2 shows the comparison between KerasTuner and iACP-FSCM.3.1.

Table 2. Comparison of iACP-FSCM and KerasTuner on 5 fold cross validation testing for main and alternative dataset.

Methods	Main Dataset				Alternative Dataset			
	Acc	**Sn**	**Sp**	**MCC**	**Acc**	**Sn**	**Sp**	**MCC**
iACP-FSCM	0.825	0.726	0.903	0.646	0.889	0.876	0.902	0.779
KerasTuner	1.00	1.00	1.00	1.00	0.99	1.00	0.981	0.981

3.1. Computational Environment

In this section, same experimental setting is used to identify ACPs from non-ACPs. KerasTuner is used to tune the deep learning model, it is a hyperparamter optimization framework which finds the best parameters from the data to predict accurately. It uses Hyperband, Random Search and Bayesian Optimization algorithms, Random Search is used for the prediction of anticancer peptides. The proposed deep learning model is trained on the training dataset and tested on the independent dataset. The model outperformed iACP-FSCM and Anti-CP 2.0 on Main and Alternative datasets. The model achieved the MCC 0.982 for Main and 0.972 for Alternative dataset. The following Figure 4 shows the comparison of KerasTuner, iACP-FSCM and Anti-CP 2.0 on independent datasets.

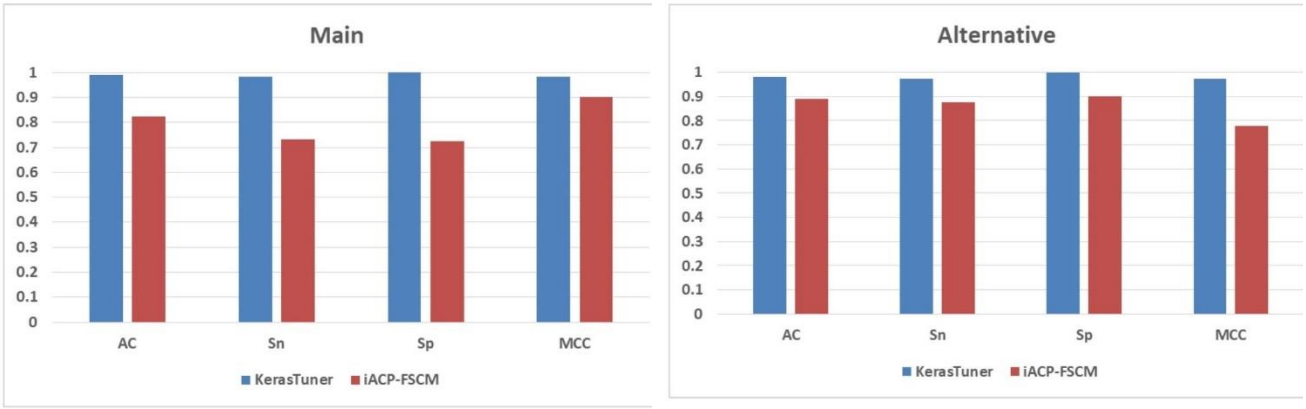

Figure 4. Performance comparison between KerasTuner, iACP-FSCM and AntiCP_2.0 in classifying ACP from non-ACP during independent testing.

4. Conclusion

In this study, a deep learning based hyperparameter optimization framework is proposed for the accurate predictions of anticancer peptides. Previously, several machine learning and deep learning model were used for the prediction of anticancer peptides. Machine learning algorithms require features extraction from the dataset which is a complex and time taking process, multiple features extraction methods were used in this study. After extracting the features, Lazypredict was used to check the highest accuracy classifier. LGBM and ExtraTree classifiers achieved the maximum accuracy but they both failed to outperform iACP-FCSM. In this study, we found out that machine learning approaches requires features extraction from the protein sequences, data fusion and model training, this is a complex and time taking process also the results were not satisfactory. On the other hand deep learning algorithms require a large amount of dataset for training of the classifier but currently large datasets are not available on anticancer peptides. KerasTuner was used to solve this problem, KerasTuner is a hyperparamter optimization algorithm which uses TensorFlow at its backend. KerasTuner outperformed all the other predictors that were developed for anticancer peptides prediction. The deep learning model has outperformed all the other predictors developed for the prediction of anticancer peptides.

References

1. Chiangjong, W., Chutipongtanate, S. & Hongeng, S. Anticancer peptide: Physicochemical property, functional aspect and trend in clinical application. Int. J. Oncol. 57, 678–696 (2020).
2. Shoombuatong, W., Schaduangrat, N. & Nantasenamat, C. Unraveling the bioactivity of anticancer peptides as deduced from machine learning. EXCLI journal 17, 734 (2018).
3. Marquez-Rios, E. & Del-Toro-Sanchez, C. L. Antioxidant peptides from terrestrial and aquatic plants against cancer. Curr. Protein Pept. Sci. 19, 368–379 (2018).
4. Klaunig, J. E. Oxidative stress and cancer. Curr. pharmaceutical design 24, 4771–4778 (2018).
5. Cardell Jr, R. R. Subcellular alterations in rat liver following hypophysectomy. Biochimica et Biophys. Acta (BBA)-General Subj. 148, 539–552 (1967).
6. Schmidhuber, J. Deep learning in neural networks: An overview. Neural networks 61, 85–117 (2015).
7. Javaid, A., Niyaz, Q., Sun, W. & Alam, M. A deep learning approach for network intrusion detection system. Eai Endorsed Transactions on Secur. Saf. 3, e2 (2016).
8. Gautam, A. et al. In silico approaches for designing highly effective cell penetrating peptides. J. translational medicine 11, 1–12 (2013).
9. Zhang, Y. P. & Zou, Q. Pptpp: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. Bioinformatics 36, 3982–3987 (2020).
10. Basith, S., Manavalan, B., Hwan Shin, T. & Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Medicinal research reviews 40, 1276–1314 (2020).
11. Tyagi, A. et al. In silico models for designing and discovering novel anticancer peptides. Sci. reports 3, 1–8 (2013).
12. 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 ames test. J. Theor. Biol. 341, 34–40 (2014).
13. Vijayakumar, S. & Ptv, L. Acpp: a web server for prediction and design of anti-cancer peptides. Int. J. Pept. Res. Ther. 21, 99–106 (2015).
14. Chen, W., Ding, H., Feng, P., Lin, H. & Chou, K.-C. iacp: a sequence-based tool for identifying anticancer peptides. Oncotarget 7, 16895 (2016).
15. Li, F.-M. & Wang, X.-Q. Identifying anticancer peptides by using improved hybrid compositions. Sci. reports 6, 1–6 (2016).
16. 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. intelligence medicine 79, 62–70 (2017).
17. 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).
18. Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V. & Shoombuatong, W. Acpred: a computational tool for the prediction and analysis of anticancer peptides. Molecules 24, 1973 (2019).
19. Manavalan, B. et al. Mlacp: machine-learning-based prediction of anticancer peptides. Oncotarget 8, 77121 (2017).
20. Wu, C., Gao, R., Zhang, Y. & De Marinis, Y. Ptpd: predicting therapeutic peptides by deep learning and word2vec. BMC bioinformatics 20, 1–8 (2019).
21. Wei, L., Zhou, C., Chen, H., Song, J. & Su, R. Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. Bioinformatics 34, 4007–4016 (2018).
22. Wei, L., Zhou, C., Su, R. & Zou, Q. Pepred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. Bioinformatics 35, 4272–4280 (2019).
23. 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. Ther. Acids 17, 1–9 (2019).
24. Rao, B., Zhou, C., Zhang, G., Su, R. & Wei, L. Acpred-fuse: fusing multi-view information improves the prediction of anticancer peptides. Briefings Bioinforma. 21, 1846–1855 (2020).
25. Agrawal, P., Bhagat, D., Mahalwal, M., Sharma, N. & Raghava, G. P. Anticp 2.0: an updated model for predicting anticancer peptides. Briefings bioinformatics 22, bbaa153 (2021).
26. Charoenkwan, P. et al. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci. reports 11, 1–13 (2021).
27. Basith, S., Lee, G. & Manavalan, B. Stallion: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. Briefings bioinformatics 23, bbab376 (2022).
28. Chen, Z. et al. ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34, 2499–2502 (2018).
29. Pandala, S. Lazypredict (2020).
30. O'Malley, T. et al. Kerastuner. https://github.com/keras-team/keras-tuner (2019).
31. Su, R., Hu, J., Zou, Q., Manavalan, B. & Wei, L. Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. Briefings Bioinforma. 21, 408–420, DOI: 10.1093/bib/bby124 (2019).
32. Basith, S., Manavalan, B., Hwan Shin, T. & Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. Medicinal Res. Rev. 40, 1276–1314, DOI: 10.1002/med.21658 (2020).

33. Git Hub Source Code = https://github.com/RaoHassanKaleem/Deep-Learning-algorithm-for-anticancer-peptides-prediction/blob/main/KerasTuner.ipynb).