# A computer-aided drug design system to predict anti-cancer peptides

Louai Zaiter[1]
[1]Aberystwyth University, United Kingdom

## Abstract

Anti-cancer peptides are a series of short peptides that inhibit tumor cell proliferation. This study introduces a computer-aided drug design system to predict potential anti-cancer peptides. We extract features from peptide sequences such as QSAR descriptors, sequence profiles, and physiochemical and biological properties. Then, we select the best set of features using the recursive feature elimination that has a random forest as an estimator. The selected features are fed into deep learning and machine learning classifiers i.e. the artificial neural network, the support vector machine, and the k-nearest neighbors. We use two publicly available datasets i.e. CancerPPD and AntiCP 2.0. The proposed artificial neural network has outperformed all other models while scoring a precision of 92%.

Keywords— Anti-cancer peptides, computer-aided drug design, artificial neural networks, k-nearest neighbors, support vector machines, recursive feature elimination

## 1 Introduction

Cancer is one of the leading causes of death around the world. According to [9], there were 375,400 new cases of cancer in the United Kingdom between 2016 and 2018. There exist several treatments for cancer including chemotherapy, radiation therapy, and surgery. The issue with those procedures is that they have a lot of side effects and even if the treatment is successful there is a high chance of reoccurrence. Peptides are molecules that contain two or more amino acids. In recent years, peptide-based therapy has been considered an efficient strategy to treat cancer.

This study proposes a machine-learning (ML) algorithm that can predict anti-cancer peptides (ACP) with high precision. The pipeline is composed of four steps; (1) feature extraction, (2) preprocessing, (3) feature selection, and (4) classification. We extract different sets of features such as physiochemical and biological properties, QSAR descriptors [5], and sequence profiles. The best set of features is selected using the recursive feature elimination (RFE) [3] algorithm and, as binary classifiers, we used the artificial neural network (ANN) [2], the support vector machine (SVM) [4], and the k-nearest neighbors (KNN) [6].

The rest of the paper is divided as follows; the second section presents the related works, the third section is about the datasets, the fourth section presents the methodology, the fifth section shows the results and findings, and the last section is a conclusion.

## 2 Related Works

Yan et al. [13] introduced a deep-learning algorithm to detect ACP. They used handcrafted features along with ordinal encoding with positional information. As a deep learning model, they used short-term memory and convolutional neural networks. And as a machine learning model, the gradient boosting machine was used. Their proposed algorithm has reached 79.9% of accuracy, 81.5% of sensitivity, and 76.6% of specificity.

Agrawal et al. [1] proposed a machine-learning algorithm that can classify peptide sequences into ACP and non-ACP. The set of extracted features includes amino acid composition [11] (AAC), dipeptide composition [7] (DPC), terminus composition, binary profile, and hybrid features. Their proposed algorithm scored 92% accuracy on the alternate dataset and 77% accuracy on the main dataset.

Schaduangrat et al. [10] introduced a computer-aided drug design system to predict ACP. They selected 138 of ACP and 205 of non-ACP. They extract AAC, DPC, and physiochemical properties (PCP). As classifiers, they used a random forest [8] and an SVM. To evaluate their results, they used the k-fold and the leave-one-out cross-validation. Their proposed algorithm reached 91% of Mathew correlation coefficient (MCC).

# 3 Dataset

We have used the alternate dataset from the AntiCP 2.0 [1] repository that contains 776 validated ACP and 776 non-ACP. Also, this study considers using a small subset from the CancerPPD database [12] i.e. 116 ACP and 116 non-ACP.

# 4 Methodology

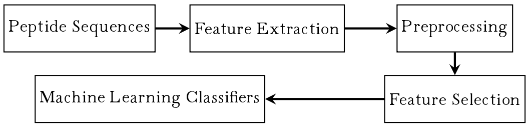

Figure 1: The proposed ACP classification pipeline

The proposed pipeline starts with extracting features from peptide sequences. As a feature extractor, we used the Peptides Python package. The set of generated features includes physiochemical and biological properties, QSAR descriptors, and sequence profiles. We use five-fold cross-validation and we balance the training set for each fold using undersampling techniques. We use the MinMaxScaler to normalize the data and an RFE algorithm to select the best 50 features. As machine learning classifiers, we used the KNN, the ANN, and the SVM with radial basis function. The proposed ANN has two blocks. Each block has a fully connected layer along with a batch normalization and an exponential linear unit (ELU) activation function. To deal with overfitting, we added a dropout layer before the last dense layer. We train our network using an Adam optimizer and a learning rate of 1e-4. As a loss function, we used the binary cross-entropy.

As evaluation metrics, we used precision, recall, accuracy, and the area under the curve.

# 5 Results and Findings

As a deep learning library, we used Pytorch and, as a machine learning library, we used Scikit-learn.

Table 1 shows the recorded metrics on the CancerPPD and AntiCP 2.0 Alternate datasets. For instance, the proposed ANN model outperforms all other classifiers while scoring an accuracy of 92%. Also, this model can predict ACP with a precision of 88%.

| Model | Dataset | Accuracy(%) | Precision(%) | Recall(%) |
|---|---|---|---|---|
| SVM | CancerPPD | 86 | 86 | 86 |
| KNN | CancerPPD | 85 | 85 | 85 |
| ANN | CancerPPD | 92 | 92 | 92 |
| SVM | AntiCP 2.0 | 89 | 89 | 89 |
| KNN | AntiCP 2.0 | 89 | 89 | 89 |
| ANN | AntiCP 2.0 | 91 | 91 | 91 |

Table 1: The recorded performance of the machine learning models on the AntiCP 2.0 Alternate and CancerPPD datasets

Table 3 shows the confusion matrix of the best-performing model. The true positive value is 718, the false positive value is 58, the false negative value is 76, and the true positive value is 700.

| Pred/Actual | ACP | non-ACP |
|---|---|---|
| ACP | 718 | 58 |
| non-ACP | 76 | 700 |

Table 2: The confusion matrix of the ANN model recorded on the CancerPPD dataset

Figure 1 shows the ROC curves of the ANN models recorded on the cancerPPD and the AntiCP 2.0 datasets. The proposed deep learning model has a high specificity and sensitivity while classifying the extracted features.

Table 3 shows a comparaison between our proposed method and the method covered in previous sections. Our proposed machine learning model has

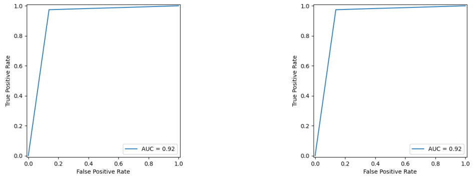

Figure 2: The receiver operating curves (ROC) of the ANN model recorded on both datasets

outperformed all other models while scoring an accuracy of 92% on the AntiCP 2.0 dataset and 91% on the CancerPPD dataset.

| Study | Dataset(s) | Metic(s) | Results |
|---|---|---|---|
| Yan et al.[13] | CancerPPD | accuracy | 79.9% |
| Agrawal et al. [1] | AntiCP main | accuracy | 77% |
| Schaduangrat et al. [10] | APD2 | MCC | 91% |
| proposed method | AntiCP 2.0 & CancerPPD | accuracy | 92% 91% |

Table 3: A comparaison between our proposed method and other studies

# 6 Conclusion

This study introduced a machine-learning pipeline to detect ACP. The best performance is recorded using the proposed artificial neural network.

In further steps of the study, we suggest using long short-term memory (LSTM) and one-dimensional convolutional neural networks (1D CNN) to directly classify peptide embeddings.

# References

[1] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, and G. P. Raghava. Anticp 2.0: an updated model for predicting anticancer peptides. *Briefings in bioinformatics*, 22(3):bbaa153, 2021.

[2] I. A. Basheer and M. Hajmeer. Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods*, 43(1):3–31, 2000.

[3] X.-w. Chen and J. C. Jeong. Enhanced recursive feature elimination. In *Sixth international conference on machine learning and applications (ICMLA 2007)*, pages 429–435. IEEE, 2007.

[4] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[5] A. U. Khan et al. Descriptors and their selection methods in qsar analysis: paradigm for drug design. *Drug discovery today*, 21(8):1291–1302, 2016.

[6] L. E. Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[7] P. Petrilli. Classification of protein sequences by their dipeptide composition. *Bioinformatics*, 9(2):205–209, 1993.

[8] Y. Qi. Random forest for bioinformatics. *Ensemble machine learning: Methods and applications*, pages 307–323, 2012.

[9] C. Research. `https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk`.

[10] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong. Acpred: a computational tool for the prediction and analysis of anticancer peptides. *Molecules*, 24(10):1973, 2019.

[11] M. H. Smith. The amino acid composition of proteins. *Journal of Theoretical Biology*, 13:261–282, 1966.

[12] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, and G. P. Raghava. Cancerppd: a database of anticancer peptides and proteins. *Nucleic acids research*, 43(D1):D837–D843, 2015.

[13] Q. Yuan, K. Chen, Y. Yu, N. Q. K. Le, and M. C. H. Chua. Prediction of anticancer peptides based on an ensemble model of deep learning and machine learning using ordinal positional encoding. *Briefings in Bioinformatics*, 24(1):bbac630, 2023.