J. Chosun Natural Sci.
Vol.16, No.4 (2023) pp.111-122
https://doi.org/10.13160/ricns.2023.16.4.111

Machine Learning Approaches for Anticancer Peptide Discovery: A Comprehensive Review

Priya Dharshini[1†]

[1]*Computational Biology Laboratory, Department of Genetic Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur-603203, Tamil Nadu, India*

Abstract

Invasive species are organisms that are introduced into places outside of their natural distribution range. The global pet trade is facilitating the introduction of invasive species into new countries and areas. Among the introduced alien species, turtles are one of the most common animal groups whether lives in wetland ecosystems, such as wetlands or reservoirs. Like other countries around the world, exotic turtles is becoming a growing concern for the wetland ecosystem in South Korea. In this study, we report new reports of subspecies of Painted turtle (Chrysemys spp.): Chrysemys picta marginata, C. p. bellii and C. dorsalis, from the reservoirs in downtown Cheongju and Gwangju, South Korea. We used morphological features, such as the characteristics of the legs, plastron, and carapace, to identify the turtles. It is assumed that all turtles were artificially released into nature. Considering the increasing number of reports on the introduction of alien invasive turtles in Korean wetlands, we recommend the formulation of an immediate and systematic management plan for pet trades and organized continuous monitoring programs.

Keywords: Cancer, Anticancer peptides, Machine learning.

(**Recceived** **August** 26, 2023; **Revised** September 14, 2023; **Accepted** **November** 13, 2023)

† Corresponding author: priyamedigen@gmail.com

1. Introduction

Cancer continues to be a significant and formidable health threat worldwide, characterized by uncontrolled cell growth and the ability to spread to other parts of the body. Its complex nature allows it to affect various tissues and organs, impacting people globally. As life expectancy rises in developed and developing countries, cancer prevalence has increased, leading to higher mortality rates[1]. Despite the availability of advanced clinical treatments such as chemotherapy, radiation therapy, and hormonal therapy, the recurrence of cancer remains alarmingly high. Moreover, these treatments often cause damage to normal, resulting in immunodeficiency among patients.

Hence, there is an urgent need to explore and create novel anti-cancer drugs that can reduce premature deaths and improve survival rates among affected populations. Peptide-based therapeutics have emerged as a promising drug class due to their perceived safety, high selectivity, good tolerability, lower production costs, ease of modification and synthesis, and favorable pharmacological properties. These attributes make them an attractive avenue for advancing cancer treatment and finding more effective and targeted therapies[2].

Peptides are short chains of amino acids, which are thefundamental building blocks of proteins. They play vital roles in various biological processes and are involved in numerous physiological functions within living organisms. In biological systems, peptides serve as signaling molecules, neurotransmitters, hormones, and enzymes, among other essential roles. They are responsible for transmitting information between cells, regulating various cellular activities, and coordinating complex physiological responses. Due to their diverse functions and potential therapeutic applications, peptides have garnered significant interest in the fields of medicine and biotechnology[3]. Peptide-based drugs have shown promising results in treating a wide range of medical conditions, including cancer, diabetes, cardiovascular diseases, and infectious diseases. One of the remarkable advantages of peptides is their specificity and selectivity[3]. They can be designed to interact with specific receptors or targets, making them highly effective and reducing the risk of off-target effects compared to traditional small molecule drugs. Additionally, peptides often exhibit lower toxicity and are well-tolerated by the human body, contributing to their safety profile in therapeutic applications. Moreover, the development of peptide synthesis techniques and delivery methods has expanded the possibilities of using peptides as therapeutic agents[4].

Recent research has identified certain peptides with diverse biological properties, making them potential candidates for novel therapeutics.These peptides include antiangiogenic peptides (AAPs), antibacterial peptides (ABPs), anticancer peptides (ACPs), antifungal peptides (AFPs), and more. The remarkable chemical and biological diversity exhibited by peptides adds to their attractiveness for therapeutic development. These discoveries raise the prospect of exploiting the unique properties of peptides to develop novel treatments for a variety of medical diseases[5].

Anticancer peptides represent a subset of antimicrobial peptides (AMPs) that have demonstrated the ability to exhibit anticancer activities. These ACPs can selectively recognize and interact with cancerous cells, leading to their destruction through various mechanisms, including induction of apoptosis (programmed cell death) and disruption of cancer cell membranes[6]. Indeed, the accurate prediction of Anticancer Peptides is of utmost importance for advancing cancer research and therapeutic development. The traditional experimental identification and development of ACPs can be a time-consuming and labor-intensive process. Therefore, bioinformatics tools and computational approaches are becoming increasingly crucial for effectively analyzing the vast amount of available data on existing peptides[7]. Machine learning (ML)-based computational approaches offer rapid and cost-effective pre-screening tools to navigate the vast combinatorial sequence space efficiently. By analyzing extensive peptide databases and learning from existing peptide data, ML models can predict and prioritize potential peptide candidates. This streamlined approach significantly accelerates and simplifies the burdensome process of peptide discovery, making it more efficient and accessible for researchers[8]. It allows for the identification of promising peptide sequences, which can then be further validated and optimized through experimental studies, leading to the potential development of novel and effective peptide-based therapeutics[9].

2. Peptides as Promising Anticancer Agents

ACPs are small peptides that usually contain 5 to 50 amino acid residues while possessing high hydrophobicity and a positive net charge. Thus, ACPs can interact with anionic cell membrane components of cancer cells and then selectively kill cancer cells. Additionally, ACPs can interfere with cancer cells by causing apoptosis mediated via mitochondrial disruption, triggering necrosis via cell lysis, stimulate the immune system of the host and prevent tumour angiogenesis. ACPs have biological properties, making them potential candidates for novel therapeutics[10]. These peptides include antiangiogenic peptides, antibacterial peptides, anticancer peptides, antifungal peptides and more. The remarkable chemical and biological diversity exhibited by peptides adds to their attractiveness for therapeutic development. These discoveries raise the prospect of exploiting the unique properties of peptides to develop novel treatments for a variety of medical diseases[11].

3. Data Sources and Datasets for Bioactive and Therapeutic Peptides

The landscape of bioactive and therapeutic peptides is illuminated by a range of comprehensive databases, each serving as a vital repository of essential peptide-related information. These data sources collectively contribute to advancing peptide research and therapeutic development. One prominent database is the APD3 (Antimicrobial Peptide

Database 3) which catalogues natural antimicrobial peptides characterized by defined sequences and activity profiles (http://aps.unmc.edu/AP/)[12]. BIOPEP is another valuable resource that focuses on biologically active peptide sequences, offering a diverse collection of peptides with various functionalities (http://www.uwm.edu.pl/bio-chemia/index.php/en/biopep)[13]. Delving into sequences, structures, and signatures of prokaryotic and eukaryotic AMPs, the CAMP (Collection of Anti-Microbial Peptides) database provides insights into antimicrobial peptides (http://www.camp.bic-nirrh.res.in/)[14]. Addressing the realm of anticancer peptides, the CancerPPD (CancerPPD: a database of anticancer peptides and proteins) database houses experimentally verified anticancer peptides and associated proteins, supporting cancer research and therapeutic innovations (http://crdd.osdd.net/raghava/cancerppd/)[15]. DRAMP (Database of Antimicrobial Peptides) takes an innovative approach by offering a resource for sequence- and structure-activity studies on AMPs (http://dramp.cpu-bioinfor.org)[16]. To aid in the discovery and design of antimicrobial agents, the LAMP (Landscape of Antimicrobial Peptides) database provides an invaluable tool (http://biotechlab.fudan.edu.cn/database/lamp)[17]. PeptideDB, in contrast, encompasses a wide spectrum of naturally occurring signaling peptides, contributing to various biological processes (http://www.peptides.be/)[18]. SATPdb (Structurally Annotated Therapeutic Peptide Database) focuses on structurally annotated therapeutic peptides with experimentally validated sequences, providing insights into therapeutic peptide design (http://crdd.osdd.net/raghava/satpdb/)[19]. Lastly, THPdb (Therapeutic Peptide Database) compiles FDA-approved therapeutic peptides and proteins, enhancing drug discovery efforts (http://crdd.osdd.net/raghava/thpdb/index.html)[20].

3. Existing Machine Learning approaches for the prediction of ACPs

In this section, we delve into the utilization of established techniques for classifying Anticancer Peptides through traditional machine learning approaches. The manual experimentation strategy for identifying new Anticancer Peptides is recognized for its time-consuming and costly nature. Given the pivotal role that ACPs play, both academic researchers and pharmaceutical companies have increasingly embraced automation as a viable alternative for ACP identification. Considering this, researchers have harnessed various automated intelligence algorithms to forecast Anticancer Peptides. In a notable anticancer investigation[21]. Chen et al. introduced the "iACP" framework tailored for peptide identification. Their approach synergistically employed an improved G-Gap DPC in conjunction with a refined peptide sequence formulation. Similarly, Manavalan et al. introduced an innovative model for ACP prediction[8]. The composite feature set, in this context, is strategically composed of optimal information, encompassing physicochemical properties, DPC, ionic attributes, and more. Employing K-fold cross-validation, their proposed system underwent training and testing phases, ensuring robustness and reliability. Additionally, Tyagi et al. devised in silico algorithms aimed at discerning ACPs from uncharacterized sequences[7]. The evaluation of the peptide's classification model was conducted using four distinct datasets.Conversely, two statistical methods, namely split AAC and binary profile, were employed for peptides encoding. Meanwhile, Li et al. introduced an innovative approach that leverages feature integration for enhanced ACP discrimination[22]. To extract robust features, a streamlined variant of AAC, individual amino acid properties, and traditional AAC were amalgamated. Through the application of Support Vector Machines (SVM), the predictive model showcased improved accuracy performance. Akbar et al. contributed by crafting a novel model dubbed "iACP-GAEnsC" for ACP identification[23]. In their pursuit, they adopted a hybrid encoding strategy to extract highly representative features from the target peptides. The integration of an evolving genetic algorithm served as a pivotal component in evaluating the performance implications of this newly devised technique. In a similar vein, Kabir et al. pioneered the "TargetACP" approach, harnessing revolutionary adaptive genetic algorithms and sequential insights[24]. Additionally, the synthetic minority oversampling technique emerged as a strategic solution to equitably distribute samples between minority and majority classes, culminating in balanced proportions. The proposed system underwent rigorous testing against diverse benchmark datasets, yielding superior performance outcomes.

Furthermore, Kumar et al. unveiled a noteworthy web server titled "ACPP," purpose-built to accurately discern positive peptides from their negative counterparts[25]. Their system encompasses a plethora of customizable settings, empowering operators to meticulously construct and identify ACPs. Notably, it extends its utility by offering insights into the cytotoxic function associated with each target peptide. Echoing a similar sentiment, Hajisharifi et al. ventured into the prediction of ACPs through the integration of pseudo amino acid composition (PseAAC) and a distinctive kernel featuring local alignment[26].

In a parallel vein, Xu et al. embarked on a subsequent investigation wherein the g-gap DPC approach was harnessed for the encoding of peptides[27]. To mitigate the presence of redundant and homogeneous features, they adopted the approach of maximum relevance-maximum distance. To enhance performance further, Boopathi et al. introduced two novel variant feature selection techniques, which judiciously select optimal yet informative descriptors from a feature space generated through seven distinct peptide encoding methods[28]. The challenge of inadequate performance stemming from high-dimensional descriptors confronted most of the machine learning models. To tackle this issue, Li et al. presented a model anchored in diverse feature extraction techniques, achieving remarkable performance even when utilizing a concise 19-dimensional vector[29]. Given the vast diversity inherent in genomic sequences, achieving precise classification of target peptides has emerged as a formidable challenge. Recognizing this complexity, Akbar et al. devised a novel strategy encompassing the fusion of three distinct peptide encoding methods. Subsequently, the incorporation of k-space amino acid pairs was employed to extract more intricately correlated features[30]. In a parallel line of inquiry, Agrawal et al. embarked on an analysis of the ETree classifier's performance in conjunction with AAC and DPC. Based on the outcomes of this investigation, they developed a web server tailor-made for edge devices, capitalizing on the model that exhibited the highest efficacy[31]. The overview of various studies focusing on Anticancer Peptide prediction was mentioned in Table 1.

4. Model refinement and Assessment of Model Performance

The primary objective of every machine learning algorithmis to effectively train the model to achieve accurate classifications for unseen data. During the process of model training, the feature descriptors extracted from the training dataset, along with their corresponding class labels (response variable: positive or negative), are fed into an ML classifier. In this stage, the classifier learns the underlying relationships between the feature descriptors (x) and the response variable (y). Consequently, the trained model becomes capable of making predictions for new, previously unseen datasets. The fundamental objective of a proficient machine learning (ML) model is to extrapolate its learning from the training dataset to external, independent datasets[32]. In the realm of computational biology and bioinformatics, a variety of classifiers are frequently employed. These encompass AdaBoost (AB), Artificial Neural Networks (ANN), Deep Learning (DL), Extreme Learning Machine (ELM), Extremely Randomized Tree (ERT), Gradient Boosting (GB), k-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB). The operational principles and applications of these machine learning classifiers have been elucidated in previous studies[33]. Indeed, each classifier comes with its own set of strengths and limitations, particularly in the context of data quantity, training speed, and feature encodings. The choice of classifier should be tailored to the specific characteristics of the dataset and the objectives of the analysis [32]. Cross-validation techniques are essential tools to prevent the overfitting of machine learning models. They help ensure that the model's performance generalizes well to unseen data. Among these techniques, one commonly used approach is K-Fold Cross-Validation, which divides the dataset into k subsets or folds. The model is trained on k-1 folds and validated on the remaining fold, repeating this process k times to cover all folds. This provides a robust assessment of the model's performance across different data subsets. Stratified K-Fold Cross-Validation is particularly useful when dealing with imbalanced datasets. It ensures that each fold maintains a proportional representation of different classes, which helps prevent biased performance estimates. Leave-One-OutCross-Validation (LOOCV) is an exhaustive technique where each data point is treated as validation once, with the rest used for training. This offers a thorough evaluation but can be computationally intensive for large datasets. Leave-P-Out Cross-Validation (LPOCV) strikes a balance by reserving p data points for validation while using the remaining for training. This approach manages computational complexity while maintaining robust partitioning[34]. Once the model is constructed, it is crucial to conduct an independent evaluation to determine its performance and generalizability. Four commonly used evaluation metrics are Sensitivity (SN), Specificity (SP), Accuracy (ACC), and the Matthews Correlation Coefficient (MCC)[35]. These metrics are calculated using the following formulas:

Sensitivity (SN) = $TP / (TP + FN)$
Specificity (SP) = $TN / (TN + FP)$
Accuracy (ACC) = $(TP + TN) / (TP + TN + FP + FN)$
Matthews Correlation Coefficient (MCC) = $(TP * TN - FP * FN) / \sqrt{}((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))$

Where TP represents true positives, TN is true negatives, FP is false positives, and FN is false negatives. These metrics provide a comprehensive understanding of the model's performance in terms of its ability to correctly classify positive and negative instances[3].

5. Web server development

Ideally, the refined predictive model along with the associated dataset should be made openly accessible, potentially through a dedicated web server infrastructure. Such an approach offers significant advantages for both experimentalists and computational researchers. By offering a web server interface, experimentalists can efficiently pinpoint potential peptide functions prior to embarking on resource-intensive experimental validation efforts. Concurrently, computational biologists can leverage this platform to cultivate sophisticated in silico prediction models[36]. The flowchart depicting the process of anticancerpeptide prediction using Machine Learning was illustrated in Fig1.

6. Conclusion

Anticancer peptides exhibit substantial promise in diverse realms, including apoptosis induction, cellular penetration, anti-inflammatory responses, and anti-angiogenic effects in both in vitro and in vivo contexts within cancer cells. While challenges persist in ACP-related investigations, the field boasts robust positive outcomes. Notably, computational methodologies employing Machine Learning andhybrid learning paradigms offer notable advantages in streamlining the identification of potent ACP candidates, reducing time and costs associated with pre-experimental phases. Moreover, prior to embarking on the resource-intensive journey of experimental validation encompassing biological functionality verification, optimization, preclinical assessments, and clinical trials for AI-predicted ACPs' therapeutic effects in cancer, the application of AI holds significant value in projecting diverse biological attributes of novel ACPs. Furthermore, acknowledging the limitations inherent in solitary cancer therapeutic modalities, an avenue to enhance efficacy emerges in the fusion of conventional therapies with the ACP strategy. This review serves to lay the groundwork for continued exploration into the development of ACPs, rooted in cancer cell characteristics. It also aims to foster comprehension of AI-driven predictions and the potential of combinational therapeutic approaches in cancer treatment.

References

[1] Ortega-García MB, Mesa A, Moya ELJ, Rueda B, Lopez-Ordoño G, García JÁ, et al. Uncovering Tumour Heterogeneity through PKR and nc886 Analysis in Metastatic Colon Cancer Patients Treated with 5-FU-Based Chemotherapy. Cancers (Basel) 2020;12(2).

[2] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. ACPred: A Computational Tool for the Prediction and Analysis of Anticancer Peptides. Molecules 2019;24(10).

[3] Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today 2015;20(1):122-8.

[4] Padhi A, Sengupta M, Sengupta S, Roehm KH, Sonawane A. Antimicrobial peptides and proteins in mycobacterial therapy: current status and future prospects. Tuberculosis (Edinb) 2014;94(4):363-73.

[5] Henninot A, Collins JC, Nuss JM. The Current State of Peptide Drug Discovery: Back to the Future? J Med Chem 2018;61(4):1382-414.

[6] Deslouches B, Di YP. Antimicrobial peptides with selective antitumor mechanisms: prospect for anticancer applications. Oncotarget 2017;8(28):46635-51.

[7] Tyagi A, Kapoor P, Kumar R, Chaudhary K, Gautam A, Raghava GPS. In silico models for designing and discovering novel anticancer peptides. Sci Rep 2013;3:2984.

[8] Manavalan B, Basith S, Shin TH, Choi S, Kim MO, Lee G. MLACP: machine-learning-based prediction of anticancer peptides. Oncotarget 2017;8(44):77121-36.

[9] Manavalan B, Shin TH, Lee G. DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. Oncotarget 2018;9(2):1944-56.

[10] Melo MN, Ferre R, Feliu L, Bardají E, Planas M, Castanho MARB. Prediction of antibacterial activity from physicochemical properties of antimicrobial peptides. PLoS One 2011;6(12):e28549.

[11] Al-Benna S, Shai Y, Jacobsen F, Steinstraesser L. Oncolytic activities of host defense peptides. Int J Mol Sci 2011;12(11):8027-51.

[12] Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 2016;44(D1):D1087-93.

[13] Minkiewicz P, Iwaniak A, Darewicz M. BIOPEP-UWM Database of Bioactive Peptides: Current Opportunities. Int J Mol Sci 2019;20(23).

[14] Waghu FH, Gopi L, Barai RS, Ramteke P, Nizami B, Idicula-Thomas S. CAMP: Collection of sequences and structures of antimicrobial peptides. Nucleic Acids Res 2014;42(Database issue):D1154-8.

[15] Tyagi A, Tuknait A, Anand P, Gupta S,

Sharma M, Mathur D, et al. CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res 2015;43(Database issue):D837-43.

[16] Kang X, Dong F, Shi C, Liu S, Sun J, Chen J, et al. DRAMP 2.0, an updated data repository of antimicrobial peptides. Sci Data [Internet] 2019;6(1):148. Available from: https://doi.org/10.1038/s41597-019-0154-y

[17] Zhao X, Wu H, Lu H, Li G, Huang Q. LAMP: A Database Linking Antimicrobial Peptides. PLoS One 2013;8(6):e66557.

[18] Liu F, Baggerman G, Schoofs L, Wets G. The construction of a bioactive peptide database in Metazoa. J Proteome Res 2008;7(9):4119–31.

[19] Singh S, Chaudhary K, Dhanda SK, Bhalla S, Usmani SS, Gautam A, et al. SATPdb: a database of structurally annotated therapeutic peptides. Nucleic Acids Res 2016;44(D1):D1119-26.

[20] Usmani SS, Bedi G, Samuel JS, Singh S, Kalra S, Kumar P, et al. THPdb: Database of FDA-approved peptide and protein therapeutics. PLoS One 2017;12(7):e0181748.

[21] Chen W, Ding H, Feng P, Lin H, Chou KC. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget 2016;7(13):16895–909.

[22] Li FM, Wang XQ. Identifying anticancer peptides by using improved hybrid compositions. Sci Rep [Internet] 2016;6(1):33910. Available from: https://doi.org/10.1038/srep33910

[23] Akbar S, Hayat M, Iqbal M, Jan MA. iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med 2017;79:62–70.

[24] Yu L, Jing R, Liu F, Luo J, Li Y.DeepACP: A Novel Computational Approach for Accurate Identification of Anticancer Peptides by Deep Learning Algorithm. Mol Ther Nucleic Acids 2020;22:862–70.

[25] Charoenkwan P, Chiangjong W, Lee VS, Nantasenamat C, Hasan MM, Shoombuatong W. Improved prediction and characterization of anticancer activities of peptides using a novel flexible scoring card method. Sci Rep 2021;11(1):3017.

[26] Hajisharifi Z, Piryaiee M, Mohammad Beigi 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 2014;341:34–40.

[27] Xu L, Liang G, Wang L, Liao C. A Novel Hybrid Sequence-Based Model for Identifying Anticancer Peptides. Genes (Basel) 2018;9(3).

[28] Boopathi V, Subramaniyam S, Malik A, Lee G, Manavalan B, Yang DC. mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. Int J Mol Sci 2019;20(8).

[29] Li Q, Zhou W, Wang D, Wang S, Li Q. Prediction of Anticancer Peptides Using a Low-Dimensional Feature Model. Front Bioeng Biotechnol 2020;8:892.

[30] Akbar S, Hayat M, Tahir M, Khan S, Alarfaj FK. CACP-DeepGram: Classification of Anticancer Peptides via Deep Neural Network and

Skip-Gram-Based Word Embedding Model. Artif Intell Med [Internet] 2022;131(C). Available from: https://doi.org/10.1016/j.artmed.2022.102349

[31] Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava GPS. AntiCP 2.0: an updated model for predicting anticancer peptides. Brief Bioinform 2021;22(3).

[32] Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 2019;18(6):463-77.

[33] Li F, Chen J, Leier A, Marquez-Lago T, Liu Q, Wang Y, et al. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. Bioinformatics 2020;36(4):1057-65.

[34] Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. Brief Bioinform 2020;21(4):1119-35.

[35] Cao R, Adhikari B, Bhattacharya D, Sun M, Hou J, Cheng J. QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics 2017;33(4):586-8.

[36] Schaduangrat N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. Meta-iAVP: A Sequence-Based Meta-Predictor for Improving the Prediction of Antiviral Peptides Using Effective Feature Representation. Int J Mol Sci 2019;20(22).

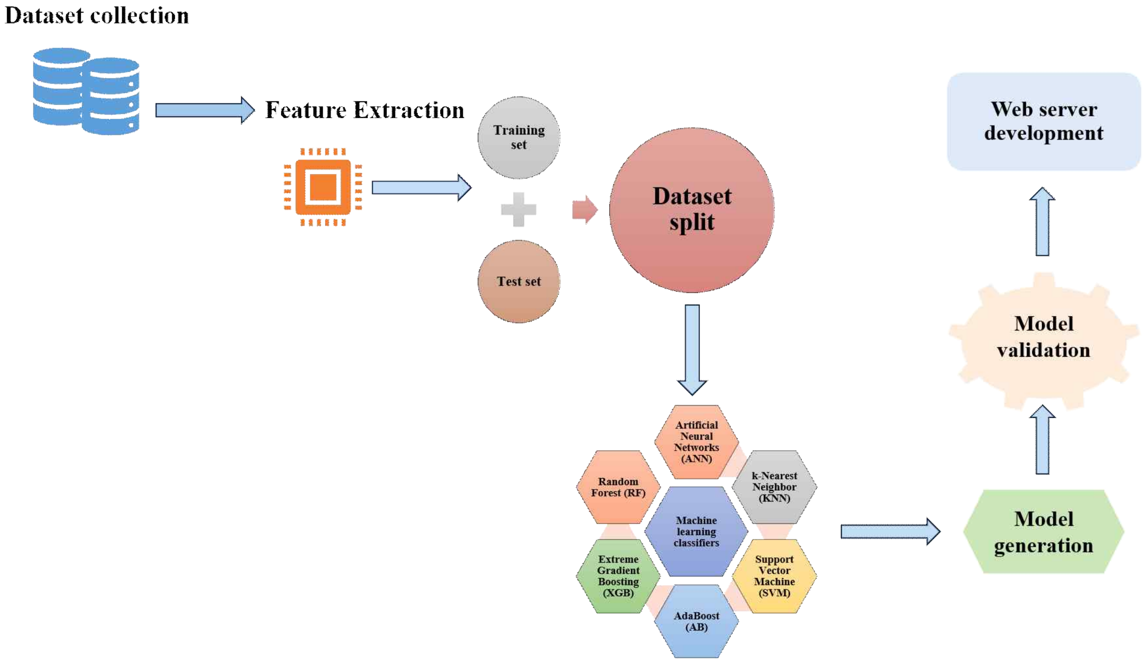
Table. 1. Comprehensive overview of various studies focusing on Anticancer Peptide (ACP) prediction

Predictor	Classifier	Feature encodings	Contributions	Reference
Chen et al	SVM	PseACC, g-gap dipeptide	Improved G-Gap DPC framework	(20)
Manavalan et al	SVM, RFT	AAC, DPC, ATC, and PCP	Composite feature set	(21)
Tyagi et al	SVM	Binary profile, DPC	Tree-based encoding	(22)
Li et al	SVM	ReduceAAC, AAC, average chemical shift	Enhancing feature selection	(23)
Akbar et al	SVM, KNN, PNN, RF, GRNN	PAAC, RAAC, g-gap dipeptide	Hybrid feature encoding	(24)
Kabir et al	SVM, KNN, PNN	Pseudo position specific scoring matrix, Composite protein sequence, Split-AAC	Diverse sequence encoding	(25)
Kumar et al	SVM, AdaBoost	Protein relatedness measure	Protein relatedness analysis	(26)
Hajisharifi et al	SVM	PAAC, Local alignment kernel	Local alignment-based encoding	(27)
Xu et al	SVM	g-gap dipeptide	g-gap DPC approach	(28)
Boopathi et al	SVM, LR, KNN, RF	Composition-based, physicochemical properties and profiles	Multi-dimensional encoding	(29)
Li et al	SVM, RFT, LibD3C	AAC, Conjoint triad, PAAC, GAAC	Comprehensive peptide features	(30)
Akbar et al	SVM, RFT, FKNN	K-space amino acid pair, Composite physiochemical properties, auto covariance,	Correlated feature extraction	(31)
Agrawal et al	Tree based	AAC, DPC, Terminus composition, binary profile	Combination of descriptors	(32)

Nomenclature

AAC	Amino acid composition
PseAAC	Pseudo amino acid composition
DPC	Dipeptide composition
ATC	Atomic composition
PCP	physicochemical properties
PAAC	Pseudo-amino acid composition
GAAC	Grouped amino acid composition
RF	Random Forest
SVM	Support vector machine
KNN	K-nearest Neighbor
PNN	Probabilistic neural network
GRNN	Generalize regression neural network

Fig. 1. Flowchart of Anticancer peptide prediction using Machine learning


Dataset collection
Feature Extraction
Training set
Test set
Dataset split
Artificial Neural Networks (ANN)
Random Forest (RF)
k-Nearest Neighbor (KNN)
Machine learning classifiers
Extreme Gradient Boosting (XGB)
Support Vector Machine (SVM)
AdaBoost (AB)
Model generation
Model validation
Web server development