IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received July 12, 2020, accepted August 8, 2020, date of publication August 11, 2020, date of current version August 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3015792

Prediction of Therapeutic Peptides Using Machine Learning: Computational Models, Datasets, and Feature Encodings

MUHAMMAD ATTIQUE[1,2], MUHAMMAD SHOAIB FAROOQ[1], (Member, IEEE), ADEL KHELIFI[3], AND ADNAN ABID[1], (Member, IEEE)

[1]Department of Computer Science, University of Management and Technology, Lahore 54000, Pakistan
[2]Department of Information Technology, University of Gujrat, Gujrat City 50700, Pakistan
[3]Department of Computer Science and Information Technology, Abu Dhabi University, Abu Dhabi, United Arab Emirates

Corresponding author: Muhammad Shoaib Farooq (shoaib.farooq@umt.edu.pk)

ABSTRACT Peptides, short-chained amino acids, have shown great potentials toward the investigation and evolution of novel medications for treatment or therapy. The wet-lab based discovery of potential therapeutic peptides and eventually drug development is a hard and time-consuming process. The computational prediction using machine learning (ML) methods can expedite and facilitate the discovery process of potential prospects with therapeutic effects. ML approaches have been practiced favorably and extensively within the area of proteins, DNA, and RNA to discover the hidden features and functional activities, moreover, recently been utilized for functional discovery of peptides for various therapeutics. In this paper, a systematic literature review (SLR) has been presented to recognize the data-sources, ML classifiers, and encoding schemes being utilized in the state-of-the-art computational models to predict therapeutic peptides. To conduct the SLR, fourty-one research articles have been selected carefully based on well-defined selection criteria. To the best of our knowledge, there is no such SLR available that provides a comprehensive review in this domain. In this article, we have proposed a taxonomy based on identified feature encodings, which may offer relational understandings to researchers. Similarly, the framework model for the computational prediction of the therapeutic peptides has been introduced to characterize the best practices and levels involved in the development of peptide prediction models. Lastly, common issues and challenges have been discussed to facilitate the researchers with encouraging future directions in the field of computational prediction of therapeutic peptides.

INDEX TERMS Anti-angiogenic, anti-cancer, anti-inflammatory, anti-microbial, feature extraction, encodings, machine learning, peptide therapeutics.

I. INTRODUCTION

Peptides are usually small chained amino acids (<50 in length), employing critical impact on humanoid physiological activities like neurotransmitters, anti-infective, ion channel ligands growth factors, and hormones [1]. Light toxicity and high specificity in regular environments make these short-chain peptides worthwhile and beneficial in the treatment [2]. Almost over seven thousand peptides with distinct characteristics have been discovered in the past, for instance, cell penetration peptides (CPP), anticancer (ACP), antiviral peptides (AVP), anti-inflammatory (AIP), and antimicrobial peptides (AMP) [3]. The peptides with these important characteristics make them vital candidates to propose new drugs for therapy. As a precedent, ACP's are being utilized for the treatment of cancer disease [4]–[7], and CCP's are observed supportive in intracellular communication and considered as a useful medium for drug delivery [8], [9].

Inflammation occurs in the reaction of any type of physical damage or injury, due to the non-specific response of the immune system [10]–[12]. AIP's are recently used in the remedy of several inflammatory disorders like Alzheimer disease [13]. Similarly, natural and synthetic peptides have the potential to constrain the manifestation of inflammatory promotors [14].

In several human pathologies, angiogenesis invokes pathological process that creates dense tumor in human body [15]. Antiangiogenic peptide's (AAP) discovery is an encouraging remedial channel for the treatment of tumor and cancerous cells by obstructing the angiogenic process. Therefore,

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

the accurate identification of AAP is mandatory to recognize their physicochemical characteristics. Another, globally preeminent mortality occurs due to cancer, and significant concentration has been made for the utilization of peptide-based therapies [16]. The most important uses where the peptide-based therapeutic methods have quickly grown is cancer treatment [6], [7], [17], [18]. The new peptide-based cancer treatment methods developed in recent years extending from naturally available and synthetic peptides to peptide conjugate drugs and small molecule peptides [7], [19]. AMP has a large-scale of actions and capabilities to operate in viral, bacterial, fungal infections, involved in mitogenic and signaling processes [20] and also shown anticancer activity. Sipuleucel-T with the trade name PROVENGE [21], [22] and Leuprolide with trade name LUPRON [1], [23], [24] are two examples from the list of FDA approved peptide-based cancer drugs for the treatment of prostate cancer. Furthermore, the HPRP-A1 peptide shown excellent antimicrobial and anticancer activity, according to the experimental study by Cuihua Hu *et al.* [6] and Wenjing Hao *et al.* [18]. Moreover, S. S. Usmani *et al.* has developed a repository having a list of FDA approved peptides and protein based drugs [5].

The production of peptide-based biologic medications are less complex and low cost comparatively of the traditional drug development process [1]. Thus, the identification of peptides with remedial characteristics is very important to invent new and effective therapeutic drugs, which accelerates their use in clinical treatment [25]. Consequently, there is immense potential for the discovery of generic, novel, and experimentally expandable solutions for precise prediction of therapeutic peptides.

The prime focus of this manuscript is to identify and discourse the diverse ML based peptide prediction models. Numerous studies have been reported in this field and gaining importance due to the effectiveness of peptide-based therapeutics. Thus, a comprehensive investigation is essential to recognize and summarize the current research developments in the field. However, there is no such review has been reported as per our best knowledge. This SLR reflects various aspects of the investigation in this area, including the use of various data-sources for dataset collection, encoding schemes to convert categorical data into ML understandable, current challenges and future prospects. Moreover, feature encoding taxonomy and framework model has also been proposed.

The next section of this paper is formulated as: Section II has been designated for the research methodology adopted in this review, describes research objectives, the research questions and motivations, search strategy, selection procedure to obtain relevant articles, abstract based keywording to classifying the articles and quality assessment criteria. Analysis and results representation has been presented in section III to deliberate the extracted outcomes and to response the research queries systematically. In section IV the research findings have been summarized by providing comprehensive discussions, proposed taxonomy and a generic framework for therapeutic peptide prediction model. Open issues and challenges have been discussed to offer imminent perspectives to the research community in section V. Lastly, the review has been concluded in section VI.

II. RESEARCH METHODOLOGY

The SLR provides a complete procedure to assist in the collection, investigation, and determination of main articles out of all available studies in the field under consideration. The guidelines for SLR proposed in 2004 by [26] has been followed in this study for unbiased data collection and demonstration of analyzed and extracted outcomes.

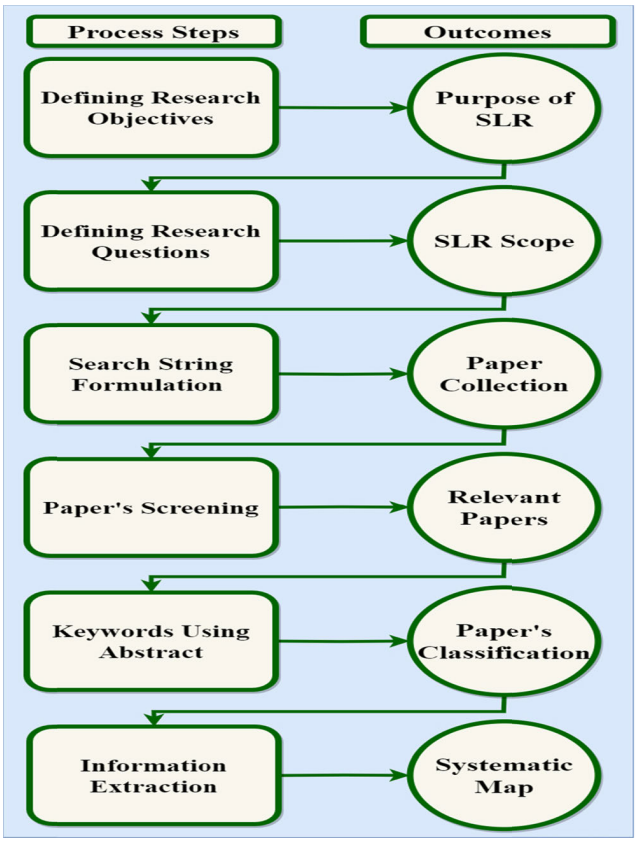

FIGURE 1. An SLR process model.

Figure 1 depicts the research process being adopted for this SLR. It is an appropriate and a reflective review procedure consists of six levels: 1) Definition of research objectives 2) definition of research questions 3) formulation of search strategy 4) screening and selection of papers 5) papers classification using keywords 6) information extraction and synthesis.

A. RESEARCH OBJECTIVES (RO)

The prime intents of this research are as follows:

RO1: The primary focus is to recognize the high-tech investigations in the area of computational prediction of therapeutic peptides.

RO2: Descriptive representation of the prevailing data-sources, feature encodings, and ML classifiers being used in the area of interest.

RO3: Hierarchal categorization of different feature encodings through proposed taxonomy, which underscores the successfully adopted schemes.

RO4: A generic framework proposal to provide a guideline to researchers for future developments in the area.

RO5: Identification of the main issues and unsolved challenges to recognize future research prospects.

B. RESEARCH QUESTIONS (RQ)

To carry out this SLR effectively, initially, the major research questions have been defined. Further, a comprehensive search planning required in the review for the identification and extraction of the most significant articles has been established. The research questions addressed in this review with their major motivation are mentioned in Table 1. The questions are addressed and responded in the light of well-defined procedure adopted in [26], [27].

TABLE 1. RQ and major motivations.

	Research Question	Major Motivation
RQ1	What is the change in publication rate relative to time in the computation prediction of therapeutic peptides?	To identify the field importance relative to time.
RQ2	What are the major data-sources / databases being used for computational prediction of therapeutic peptides?	To understand various databases, data collection and preprocessing mechanism used in the domain.
RQ3	Which encoding schemes are being used to train the computational prediction models?	To understand the features encoding schemes used specifically in the domain of peptide prediction.
RQ4	What are the major ML classifiers used in peptide prediction models?	To identify and understand the requirements of different models used for the prediction of therapeutic peptides.
RQ5	Which main publication sources are basically targeted for the publication of articles in the selected area of study?	To discover the prime journal sources from where the literature of the selected domain can be obtained or published.

C. SEARCH SCHEME

The crucial phase of SLR is the preparation of a search plan to adequately discover and collect possibly significant articles in the chosen field. This process entails the description of a search string, literature resources used to apply the search, and the segregation (inclusion/exclusion) plan to obtain the most relevant article's out of the collection. Numerous aspects of collected articles were assessed qualitatively and empirically to represent various perspectives associated with the research.

1) SEARCH STRING

An effective and fair investigation has been conducted by formulating a keyword-based string to search and gather available studies in the field using various well-known digital research repositories. To assure the authenticity of the search string regarding the relevancy of its results, the chief concepts have been analyzed in the light of research questions to obtain relevant keywords and terms used in the selected field of study. The finalized keywords and their alternative terms (synonyms) required to formalize a search string for the identification of most relevant articles are specified in Table 2. In Table 2 the '+' sign is used to express the inclusion and '-' sign for exclusion of studies having such terms.

TABLE 2. Terms and keywords used in search.

Terms (Keywords)	Synonyms / Alternate Keywords
+ Computer Based (CB)	Computational (CL), Machine Learning (ML)
+ Therapeutic Peptides Prediction (TPP)	Peptide Prediction (PP), Therapeutic Peptides (TP), Prediction of Peptides (PoP)
- Mass Spectrometry (MS)	Spectrometry (SY), Spectrometric (SC), Spectrum (SM)
- Binding (B)	-

The logical "AND" and "OR" operators were used to combine the finalized keywords and alternative terms to form a search string. The "$*$" wild character was also utilized to indicate zero or more characters where required. The "OR" operator allows extra options to search in, while the "AND" is to concatenate the terms to specify the search options and to confine the query to obtain relevant search results.

The finalized search string has three fragments. The first fragment of the string is used to confine the results related to terms computer based or computational and the next fragment relates to the prediction of peptides, while the last is used to confine the results from the inclusion of studies that are based on non-computation methods (i.e. Lab-based experiments). Equation (1) representing the mathematical formulation of the search string.

$$R = \forall[(CB \vee CL \vee ML) \wedge (PP \vee TP \vee PoP) \not\equiv (MS \vee SY \vee SC \vee SM \vee B)] \quad (1)$$

Here in (1), R stands for search results obtain against search string, '$\forall$' representing 'for all', '$\vee$' used for 'OR' operator and '$\wedge$' for 'AND' operator combining with the search terms expressed in Table 2 to formalize the complete search string according to each selected repository. The generic search term using (1) can be expressed as:

((computer based OR computational OR machine learning) AND ("peptide prediction" OR "prediction of peptide" OR "therapeutic peptides") NOT (spectrometry OR binding))

2) LITERATURE RESOURCES

Field specific and most prominent journals have been selected to conduct the literature search from online repositories, dedicated for research publication and collection. The details of selected repositories, applied search strings and the results are mentioned in Table 3.

TABLE 3. Publisher wise search strings.

Repository	Search Strings
PLOS	((("COMPUTER BASED" OR COMPUTATIONAL OR MACHINE LEARNING) AND (THERAPEUTIC PEPTIDES PREDICTION OR PEPTIDE PREDICTION OR PREDICTION OF PEPTIDES)) NOT MASS SPECTROMETRY
PMC	("COMPUTER BASED"[ALL FIELDS] OR COMPUTATIONAL [ALL FIELDS] OR MACHINE LEARNING[ALL FIELDS]) AND ("PEPTIDE PREDICTION"[ALL FIELDS] OR "THERAPEUTIC PEPTIDES"[ALL FIELDS] OR "PEPTIDES THERAPEUTICS"[ALL FIELDS] OR "PREDICTION OF PEPTIDES"[ALL FIELDS]) NOT (SPECTROMETRY OR SPECTROMETRIC [ALL FIELDS] OR SPECTRUM [ALL FIELDS])
Oxford Academics	("COMPUTER BASED" OR COMPUTATIONAL OR MACHINE LEARNING AND THERAPEUTIC PEPTIDES PREDICTION OR PEPTIDE PREDICTION OR PREDICTION OF PEPTIDES AND NOT MASS SPECTROMETRY NOT BINDING)
Springer Link	(("COMPUTER BASED" OR COMPUTATIONAL OR MACHINE LEARNING) AND (THERAPEUTIC PEPTIDES PREDICTION OR PEPTIDE PREDICTION OR PREDICTION OF PEPTIDES) AND NOT (MASS SPECTROMETRY) AND NOT (BINDING))
Science Direct	((COMPUTER BASED OR COMPUTATIONAL OR MACHINE LEARNING) AND ("PEPTIDE PREDICTION" OR "PREDICTION OF PEPTIDE" OR "THERAPEUTIC PEPTIDES") NOT (SPECTROMETRY OR BINDING))
IEEE Xplore	((((("ALL METADATA":"COMPUTER BASED") OR "ALL METADATA":COMPUTATIONAL) OR "ALL METADATA":MACHINE LEARNING) AND "ALL METADATA":PEPTIDE PREDICTION) OR "ALL METADATA":THERAPEUTIC PEPTIDES PREDICTION) NOT "ALL METADATA":SPECTROMETRY)
PUBMED	("COMPUTER BASED"[ALL FIELDS] OR COMPUTATIONAL [ALL FIELDS] OR MACHINE LEARNING[ALL FIELDS]) AND ("PEPTIDE PREDICTION"[ALL FIELDS] OR "THERAPEUTIC PEPTIDES"[ALL FIELDS] OR "PEPTIDES THERAPEUTICS"[ALL FIELDS] OR "PREDICTION OF PEPTIDES"[ALL FIELDS]) NOT (SPECTROMETRY OR SPECTROMETRIC [ALL FIELDS] OR SPECTRUM [ALL FIELDS])
ACM Digital Library	(((("COMPUTERS" OR "COMPUTER") AND BASED) OR COMPUTATIONAL OR (("MACHINE" AND "LEARNING") OR "MACHINE LEARNING")) AND ((("THERAPEUTICS" OR "THERAPEUTIC") AND ("PEPTIDES" OR "PEPTIDE") AND PREDICTION) OR (("PEPTIDES" OR "PEPTIDE") AND PREDICTION) OR ("PREDICTION OF" AND ("PEPTIDES" OR "PEPTIDE")))) NOT (("MASS SPECTROMETRY" OR ("MASS" AND "SPECTROMETRY")))

3) INCLUSION AND EXCLUSION CREITERIA

Parameters defined for inclusion criteria (IC) are:

IC 1) Include studies that were primarily conducted for computational prediction of peptides

IC 2) If the study was targeting the therapeutic effects of peptides.

The exclusion criteria applied on all the articles to exclude such studies that involved in lab-based experimental identification/prediction of peptides for their therapeutic effects or involved in other irrelevant procedures, like

EC 1) If the study did not involve any computational prediction and only involved in-vivo/wet-lab experiments based identification of peptides.

EC 2) Haven't any therapeutic activity.

EC 3) If the study was conducted for structural representation of peptides or peptide-protein mapping.

EC 4) If the study involved binding mechanism.

D. SELECTION OF RELEVANT PAPERS

To intact with relevancy, January 2010 to December 2019 was selected as publication period to search in. The primary search process produces numerous research articles, not all of them were specifically appropriate in accordance to the defined research questions, and also have duplication. Accordingly, the searched papers require re-assessment and screening to acquire truly important papers. For the determination of articles relevancy and screening, we utilized the process outlined in [28].

The first step is to filter the studies based on the titles and duplication removal. In this study a number of papers were available that are irrelevant to the selected domain, so filtered on the basis of title and irrelevant papers were barred as a beginning measure. In selection continuation, abstracts were carefully reviewed and subjected articles were included that describe the computational efforts in the selected area. Furthermore, next to examining the articles following inclusion criteria, the exclusion criterion was also used to refine the articles that deliberate the field of therapeutic peptide prediction however did not submit any substantial study contribution or involved clinical experiments. Finally, the selected articles after the described process were included in the succeeding assessment level.

E. ABSTRACT BASED KEYWORDING

Further, the screening and classification of articles have been carried out using the two-staged abstract based keywording procedure, presented in [22], to obtain relevant articles. The abstract was analyzed initially to perceive the primary idea of the article, its contribution to the domain and to discover the most pertinent keywords. The keywords identified from various articles are then combined through which a tremendous comprehension has been developed about the contribution of the research in the domain. Ultimately, these keywords have been used to classify the articles for mapping.

Based on the significance of therapeutics, the selected articles have been separated into four major categories; AIP, AAP, ACP, and AMP. These areas were chosen due to their direct importance in peptide-based therapeutics. Other domains like ''Cell-penetrating peptides (CPP)'', ''quorum-sensing peptides (QSP)'', etc., are out of scope of this study. Furthermore, just for clarity that AMP has several recognized therapeutic activities like anti-viral, anti-bacterial, anti-fungal, etc [29], [30] considered and grouped here as AMP.

F. QUALITY ASSESSMENT CRITERIA

A systematic quality assessment of the selected articles is an important task of a systematic review. Since these studies differ according to the design, the assessment of quality has been carried out by the mean of the sequential evaluation process suggested in [31]. The process contains standards for the quality assessment considering all significant factors involved in research [32], comprising notion, plan of the study, way to collect data, analysis of data, discussion, and outcomes. Based on the above points to extend the quality of research, a questionnaire has created jointly with other authors (examine Table 4). Internal and external quality criteria have also been used to improve the assessment quality as adopted by [33].

TABLE 4. Questionnaire to assess quality.

Sr.	Assessment Questions	Expected Answers	Score
Internal Scoring			
1	Was abstract well described?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
2	Was background / literature-review described in detail?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
3	Was data collection mechanism clearly defined?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
4	Was the feature description / selection defined clearly?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
5	Was methodology section clearly defined?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
6	Was result assessment well described valid and reliable?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
7	Is open accessible tool, service / API or a web service available for test results on providing dataset?	a. Yes b. No	a. 1 b. 0
8	Was conclusion relevant and effectively based on results?	a. Yes b. Intermediate c. No	a. 1 b. 0.5 c. 0
External scoring (based on publication-source)			
9	A study published in CORE ranked conference, proceedings and symposium.	a. CORE rank A b. CORE rank B c. CORE rank C d. No CORE ranking	a. 1.5 b. 1 c.0.5 d. 0
10	A study published in JCR listed ranked journal	a. JCR rank Q1 b. JCR rank Q2 c. JCR rank Q3/Q4 d. No JCR ranking	a. 2 b. 1.5 c.1 d. 0

The internal criteria belong to the internal quality assessment of an article while the external quality has been determined by considering the stability and reliability of the publication source of an article. The "Journal Citation Reports (JCR)" and the "Computer Science Conference rankings (CORE)" have been used to rate and measure the external quality [34]. The total score is the sum of the individual score of each criterion. The final score may be varied between minimum score 0 and maximum 10 scores and categorized as high ranked if the score is greater than 8, ranked average if the score is between 6 to 8 and low ranked if the score is less than 6.

III. DATA ANALYSIS

This segment compiles the results and presents a clear evaluation of all selected articles. The selected articles were investigated to effectively answer the research questions. The first part of this section discusses the search results obtained through the defined search string. Next to it is the description of the assessment score and the final part is dedicated to the comprehensive discussions to answer the research questions.

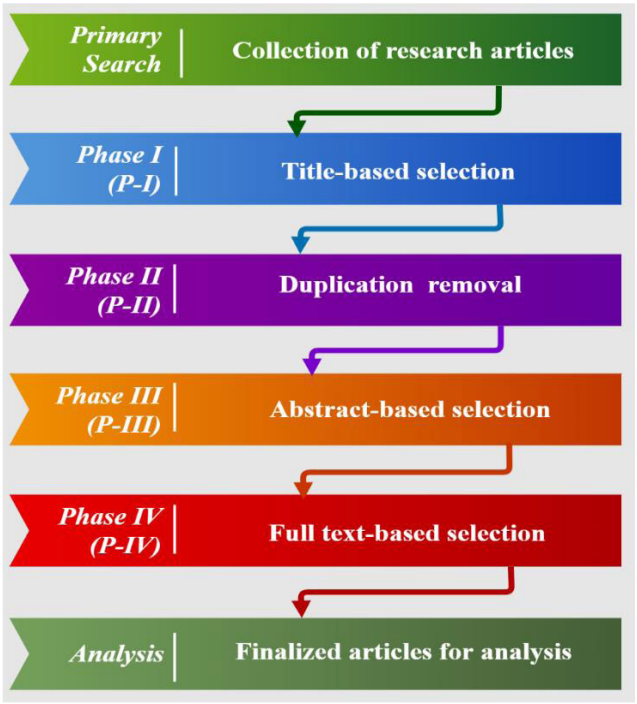

FIGURE 2. Selection procedure.

A. SEARCH RESULTS

Modern computational prediction of therapeutic peptide involves various aspects like the use of data-sources for the collection of benchmark dataset, feature extraction, computational model. The primary search process yields a total 2269 articles from various online data sources. On this collection, selection process described in the previous section was applied. The phases involved in the selection process have also been described below in Figure 2 and phase wise selection outcomes are expressed in Table 5.

The title based selection was performed by two authors in phase I (P-I), results in the selection of 216 articles. Next, the duplicate articles were removed in phase (P-II), and domain irrelevant articles were also screened on the basis of inclusion and exclusion criteria defined in previous section.

TABLE 5. Publisher based stage wise selection process.

Database/Repository	Primary Search	P-I	P-II	P-III	P-IV
PLOS	810	49	25	9	5
PMC	577	35	18	11	6
OXFORD ACADEMICS	448	28	21	7	4
SPRINGER LINK	171	26	16	5	1
SCIENCE DIRECT	75	17	12	5	2
IEEE XPLORE	68	15	9	5	2
PUBMED	62	33	29	24	17
ACM DIGITAL LIBRARY	58	13	7	6	4
TOTAL	**2269**	**216**	**137**	**72**	**41**

For instance, the search procedure has also produced papers that were involved in wet-lab experimental identification of peptides or related to their binding capabilities, but not involved in computational prediction of therapeutic peptides, so excluded due to their complete irrelevance. The consensus was measured 89% using the Kappa factor [35] indicating the high agreement among authors in selection. Abstract based screening was applied in phase III (P-III) on the 137 resultant articles obtained from the previous phase, and finally in phase IV (P-IV) full text-based analysis was applied on 72 articles and total 41 articles were found most pertinent and finalized to include in this SLR for data extraction and analysis.

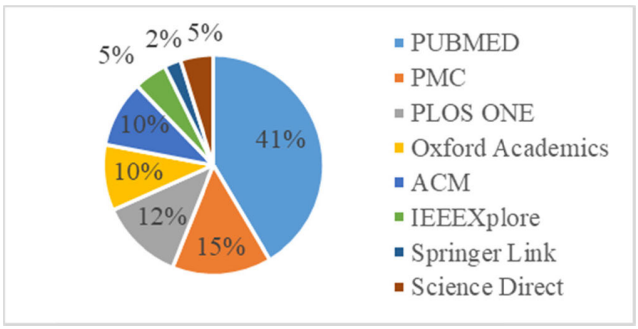

FIGURE 3. DL-wise ratio of selected studies.

Extremely realized digital libraries (DL) to publish research studies for various journals, conferences and workshops were employed to select studies for this systematic literature review as per search strategy shown in Table 3. Figure 3 shows the DL-wise distribution ratio of selected articles, includes PUBMED as the top-ranked with 41% share, 15% of PMC, 12% of PLOS ONE, 10% of Oxford Academics and ACM each, 5% of IEEEXplore and Science Direct each, and 2% of Springer Link. Publisher based staged wise selection status and distribution ratio of selected studies has already been shown in Table 5 above.

B. QUALITY ASSESSMENT SCORE

As per the scoring mechanism defined above the scores were given to each selected study, evaluated according to the internal and external criteria are mentioned in Table 6. The scores

TABLE 6. Quality assessment score and classification.

Ref. No.	P. Type	Domain	I-Score	E-Score	Total Score
[16]	Journal	ACP	7.5	2	9.5
[36]	Journal	AMP	4.5	2	6.5
[37]	Journal	AMP	8.0	2	10
[38]	Journal	AMP	7.0	2	9
[15]	Journal	AAP	6.5	2	8.5
[39]	Journal	AMP	5.5	2	7.5
[40]	Journal	AMP	5.0	2	7
[41]	Journal	AMP	6.5	2	8.5
[42]	Journal	AMP	4.5	2	6.5
[43]	Journal	ACP, AMP	5.0	2	7
[44]	Journal	ACP	5.5	2	7.5
[12]	Journal	AIP	7.5	2	9.5
[45]	Journal	AAP	7.5	2	9.5
[3]	Journal	ACP, AMP, AAP, AIP	6.5	2	8.5
[46]	Journal	ACP, AMP	6.5	2	8.5
[47]	Journal	ACP	6.0	2	8
[48]	Journal	AMP	6.0	1.5	7.5
[49]	Journal	ACP	8.0	1.5	9.5
[50]	Journal	AMP	4.5	1.5	6
[51]	Journal	AMP	4.5	2	6.5
[52]	Journal	AAP	6.0	2	8
[53]	Journal	ACP	5.5	2	7.5
[54]	Journal	AIP	7.0	2	9
[55]	Journal	AMP	4.0	2	6
[56]	Journal	AMP	5.5	2	7.5
[57]	Journal	ACP	6.5	2	8.5
[58]	Journal	AAP	4.5	2	6.5
[13]	Journal	AIP	6.0	2	8
[59]	Journal	AAP	4.5	2	6.5
[60]	Journal	ACP	4.0	2	6
[61]	Journal	ACP	5.5	2	7.5
[62]	Journal	AMP	4.5	1.5	6
[63]	Journal	ACP	6.0	1	7
[64]	Journal	ACP	6.5	2	8.5
[65]	Conference	AMP	3.5	1	4.5
[66]	Journal	AMP	2.5	1.5	4
[20]	Journal	AMP	2.5	1.5	4
[67]	Conference	ACP	5.0	1	6
[68]	Conference	AMP	3.5	1	4.5
[69]	Journal	ACP	6	2	8
[70]	Journal	ACP	5.5	2	7.5

obtained through internal and external criteria are represented as I-Score and E-Score respectively and publication type as P. Type. Out of all selected articles, 32% received a highest score greater than 8 and categorized in the high ranked articles, 58% articles obtain average rank, while merely 4 out of 41 articles (10%) viewed in low rank according to the criteria. These facts are represented in Figure 4, demonstrating a high trust in the quality of selected articles. Though, no elimination was done on a quality basis.

C. ASSESSMENT AND DISCUSSION OF RESEARCH QUESTIONS

In this part, the fourty-one main articles were analyzed on the basis of research questions described in Table 1. The facts extracted after the analysis of the selected studies were discussed on the basis of questions for the evaluation of piecemeal information.

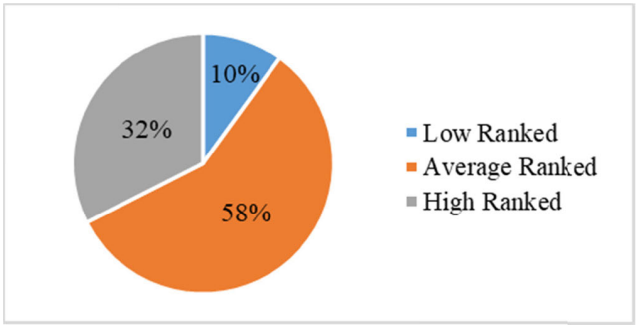

FIGURE 4. Percentage of score-based articles ranking.

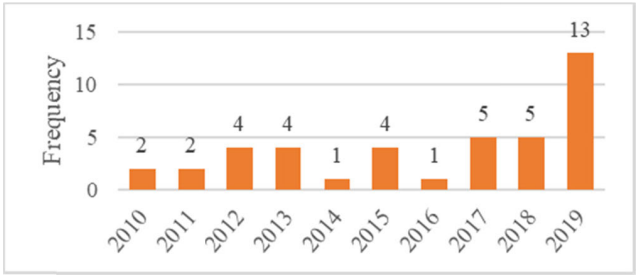

FIGURE 5. Year-wise publication frequency of selected studies.

1) ASSESSMENT OF QUESTION 1. WHAT IS THE CHANGE IN PUBLICATION RATE RELATIVE TO TIME IN THE COMPUTATIONAL PREDICTION OF THERAPEUTIC PEPTIDES?

Domain importance can be measured by gradually increased interest in the selected field of study. The published articles were selected in the years ranged from January 2010 to December 2019. Figure 5 represents year-wise frequency distribution of selected articles showing a linear increase of domain per year. Maximum 13 papers Out of 41 selected articles, 32% of the total were identified in 2019, 12% in 2018 and 12% in the year of 2017. The overall trend is showing the continuous growth and the importance of the domain per year.

2) ASSESSMENT OF QUESTION 2. WHAT ARE THE MAJOR DATA-SOURCES/DATABASES BEING USED FOR COMPUTATIONAL PREDICTION OF THERAPEUTIC PEPTIDES?

Data collection is the primary task to build a prediction model based on ML methods. The major data-sources are listed in Table 7, identified during the detailed review of selected articles. Most of the articles use multiple data-sources to collect their relevant dataset. Domain wise listing of data sources used in selected articles are mentioned in Table 8. Due to the limitations of available data most of the research has been conducted by collecting data from previous studies. Sixty percent of the included studies use the datasets of previously published articles combining various online resources according to the facts shown in Table 8.

TABLE 7. List of major data sources obtained from selected articles.

Sr. No.	Name of Data Source	Data Source Link	Accessible
1	APD	*aps.unmc.edu/AP/main.php*	*Yes*
2	CAMP	*bicnirrh.res.in/antimicrobial*	*Yes*
3	DADP	*split4.pmfst.hr/dadp/*	*Yes*
4	Swiss-Prot	*ebi.ac.uk/swissprot/*	*Yes*
5	UniProt	*uniprot.org/*	*Yes*
6	PDB	*rcsb.org/*	*Yes*
7	NCBI	*ncbi.nlm.nih.gov/*	*Yes*
8	IEDB	*iedb.org/*	*Yes*
9	CancerPPD	*crdd.osdd.net/raghava/cancerppd*	*No*
10	Previous Studies (PS)	*N.A*	*N.A*

TABLE 8. Domain wise listing of data sources used in selected articles.

Domain Type	Name of Data Source	References (Used in)
AMP	APD	[17], [36], [37], [40], [48], [50], [55], [70]
	CAMP	[17], [36], [38], [39], [41], [42]
	DADP	[17]
	Swiss-Prot	[17], [3], [55]
	UniProt	[37], [38], [39], [42], [48], [70]
	PDB	[40]
	NCBI	[38], [39], [41]
	PS	[17], [36], [3], [46], [51], [56], [62], [65]
ACP	APD	[16], [17], [44], [53], [60], [63], [69]
	CAMP	[16], [17], [53], [63], [66]
	DADP	[16], [17], [46], [53], [61], [63]
	Swiss-Prot	[17], [44], [3], [46], [53]
	UniProt	[20], [60], [63], [66]
	CPPD	[17], [44]
	PS	[16], [17], [44], [3], [46], [47], [49], [57], [60], [61], [63], [64], [67], [68], [69]
AAP	Swiss-Prot	[15], [3]
	PS	[15], [45], [3], [52], [58], [59]
AIP	Swiss-Prot	[3]
	IEDB	[13], [12], [54]
	PS	[3], [54]

An APD named as "Antimicrobial Peptide Database" had come to 2003 initially and expanded into APD3 on regular updates, comprises on 185 anti-cancer peptides, 80 anti-parasitic, 172 anti-viral, 959 anti-fungal, 105 anti-HIV and 2,169 antibacterial [71]–[73]. CAMP stands for "Collection of Antimicrobial Peptides Database", freely available online and established for the improvement of current understanding of antimicrobial peptides. It is composed manually and recently has 3782 sequences of antimicrobial peptides, out of these 2766 were experimentally tested and 1016 predicted datasets according to their reference material [38], [74]. "DADP (database of anuran defense peptides)" was manually built, presently comprises of 2571 items out of which 921 sequences were mainly collected from previously published articles as well as from UniProt [75].

The "Swiss Bioinformatics Institute (SIB)", "Protein Information Resource (PIR)" and "European Bioinformatics Institute (EBI)" collaboratively hosting the UniProt consortium aimed to provide valuable protein knowledge-base. It provides a free and convenient way to access primary protein sequences and necessary explanations. Core features include a user-friendly website, sequence archiving, manually curated protein sequences and procurement of valuable information across a reference to other knowledge bases [76].

The Swiss-Prot is also the part of the UniProt knowledgebase having analyzed and manually interpreted sequences [77], [78]. The "Protein Data Bank (PDB)" is a primary source of data used in various online data sources like UniProt, providing a firm base for research in Bioinformatics with approximately 159670 structures [79]. The "National Center for Biotechnology Information (NCBI)" is another large data source of nucleic acid sequences, molecular modeling and protein clusters, i.e. GenBank [80].

IEDB lists experimentally validated data almost 120000 structures investigated in human and non-human species. Scientific researchers can easily access online data from the provided web interface, see Table 7. The "Immune Epitope Database (IEDB)" also contains tools to help predict and analyze sequences [81]. As claimed in [82], cancerPPD contains anti-cancer based 3491 peptides and 121 protein entries with comprehensive information. It is manually curated and data gathered from other databases, patents and previously published research papers.

3) ASSESSMENT OF QUESTION 3. WHICH FEATURE ENCODING SCHEMES ARE BEING USED TO TRAIN THE COMPUTATIONAL PREDICTION MODELS?

The peptides are of divers length sequences of amino acids containing biological information [16], while the machine learning methods required feature vectors of fixed length [83]. Accordingly, the first step should be the fixed length numerical formulation of these non-numeric sequences. Over time, various feature encoding schemes have been developed for this purpose. These schemes are developed to precisely indicate the sequential knowledge and exhibit associations within the peptide sequences. [29]. The sequence of a peptide can be mathematically expressed as:

$$P = \{p_1, p_2, p_3 \ldots p_N\} \tag{2}$$

where p_i represents the first amino acid residues, p_2 is the second residue, ..., p_N is the last amino acid residue in the peptide sequence P and N denotes sequence length. Each of the amino acid residues represented in (2) belongs to one of the twenty natural amino acids. The list of 20 natural amino acids include A, D, C, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y in a single letter symbolic form.

The feature encodings schemes being used in selected articles are summarized in Table 9. Common supporting feature encoding tools may use to encode several types of features are detailed at the end of this section. However, distinctive tools are mentioned within the description of some feature encoding.

a: AMINO ACID COMPOSITION (AAC)

The peptides are the sequences of amino acids, and AAC is its compositional representation, where the relational arrangement of individual amino acid in a sequence is determined. It produces a twenty dimensional (20-D) numeric vector that describes the number of each type of amino acids normalized with the total number of amino acids in the length of a provided peptide sequence [56], [84], and can be calculated as:

$$AAC_i = \sum N_i * 100/L \tag{3}$$

where L is the sequence length, i represents 20 natural amino acids, and N_i is the amino acid of type i.

As AAC gives the occurrence frequency of a given peptide, so it supports to analyze the compositional features of a class of peptides and facilitate to distinguish one class from another on the basis of this information [13]. For example, the amino acids Gln, Tyr, Ser, Arg and Leu were found in abundance in anti-inflammatory peptides, while in non-anti-inflammatory peptides Thr, Asp, Pro, Ala and Gly were found in greater extent [13].

TABLE 9. Summary of feature encoding methods.

FEATURE ENCODINGS	REFERENCES (USED IN)
AAC	[12], [13], [15], [16], [36], [37], [38], [39], [40], [44], [45], [3], [47], [48], [49], [53], [55], [56], [59], [66]
DC	[12], [13], [15], [38], [44], [45], [3], [47], [49], [53], [58], [59], [66], [67]
TC	[13], [38], [58], [59]
PSE-AAC	[39], [45], [47], [49], [58], [67], [69]
AM-PSEAAC	[45], [47], [56], [70]
CTD	[16], [38], [3], [52]
PCP	[12], [15], [16], [43], [36], [38], [40], [41], [42], [44], [45], [3], [47], [48], [50], [51], [52], [53], [54], [58], [68]
GDC	[16], [3], [60], [70]
ASDC	[16], [3]
MBF	[13], [15], [36], [41]
KSAAP	[54]
CGR	[57]
DFT	[52]
KMC	[46]
SA	[36], [41], [69]
SSE	[38], [40], [41], [51], [54], [65]
PRM	[63]
BPF	[16], [64]

b: DIPEPTIDE COMPOSITION (DC)

DC represents the percentage of all the potential amino acid pairs available in a particular peptide. As 20 diverse amino acids exist in nature, so the 400 (20^2) possible peptide pair may exist, accordingly, produce a feature vector of size 400 vector size. DC carries information regarding the adjacent pair (neighboring residues) of amino acids in a peptide sequence [13], [59], [84]. The ratio of dipeptide may vary from one class to another class, thereby, this kind of extracted information may support in peptide type specification. For instance, the anti-inflammatory peptides were found enriched in the different amino acid pair compared to non-anti-inflammatory peptides [13].

Equation (3) represents mathematical formulation to calculate DC feature vector [59]:

$$DC_{ij} = \sum D_{ij} * 100/\sum_{i=1}^{20}\sum_{j=1}^{20} D_{ij} \tag{4}$$

where D_{ij} represents the number of dipeptides with specific type of amino acid pair i and j. $\sum_{i=1}^{20}\sum_{j=1}^{20} D_{ij}$ represents all possible dipeptides in the given peptide sequence respectively.

c: TRIPEPTIDE COMPOSITION (TC)

Similar to DC, the TC measures the rate of all potential tripeptides that a sequence may contain. The TC is being used to transform the amino acid order of varying sizes into a fixed-size vector of 8000 (20^3) dimension in size [13].

Likewise, other composition based features, it was being analyzed that different peptide classes may have different frequency of tripeptides combination so TC may also helpful in peptide classification or recognition [13]. The calculation of TC is similar to dipeptide composition, however the only difference is it takes into account the information of three consecutive peptides [59] and can be express as:

$$TC_{ijk} = \sum T_{ijk} * 100 / \sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20} T_{ijk} \quad (5)$$

where T_{ijk} represents the number of tripeptides with specific type of adjacent amino acid residues i, j, *and* k. $\sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{k=1}^{20} T_{ijk}$ represents all possible tripeptides in the given peptide sequence respectively.

d: G-GAP DIPEPTIDE COMPOSITION (GDC)

The GDC is an extension of DC (Dipeptide Composition) defined above, numerically represents the relationship of adjoining amino acids (AA_i, AA_{i+1}) in a given sequence, whereas the DC descriptor was unable to encapsulate similar information about non-adjacent amino acids $(AA_i, AA_j; j - i > 1)$. However, the more essential properties may be placed in the higher level of correlation of amino acid residues. The non-adjacent amino acid residues are found closer spatially in tertiary structure reveals the extended significance than the adjoining amino acid residues in biology. Specially in some common secondary structures, like beta-sheet and alpha-helix are two non-adjacent residues, but spatially closer and connected with hydrogen bonds [85]. Thus, non-adjacent correlation may be useful for type categorization of a peptide. Therefore, GDC introduced, a 400-dimension descriptor capable to encompass the information on relational arrangement between non-adjacent amino acids from the sequence of a given peptide [16]. GDC descriptor can be express as:

$$GDC_g = (f_1^g, f_2^g, \cdots f_{400}^g) \quad (6)$$

where f_i^g represents the frequency of i^{th} g-gap dipeptide existence in a given peptide sequence with $i = 1, 2, 3, \cdots, 400$, and f_i^g is evaluated as:

$$f_i^g = N_i^g / (\sum_{i=1}^{400} N_i^g) \quad (7)$$

where N_i^g is the number of existence of i^{th} g-gap dipeptide in a sequence.

e: ADAPTIVE SKIP DIPEPTIDE COMPOSITION (ASDC)

Wei *et al.* proposed ASDC, a modified version of dipeptide composition to retain the correlation information between adjacent and intervening amino acids in a peptide sequence [86]. Therefore, the ASDC carries the properties of both the DC and GDC, and is a of 400 dimension feature vector [61]. The feature descriptor of ASDC for a given peptide sequence is expressed as:

$$ASDC = (v_1, v_2, \cdots, v_{400}) \quad (8)$$

and v_i is evaluated as:

$$v_i = (\sum_{g=1}^{L-1} n_i^g) / (\sum_{i=1}^{400}\sum_{g=1}^{L-1} n_i^g) \quad (9)$$

where v_i denotes the frequency vector contains the occurrence of all potential pairs of intervening amino acid residues less than or equal to *L-1*.

f: K-SPACED AMINO ACID PAIRS (KSAAP)

For bio-informatics studies, KSAAP is widely used feature descriptor. KSAAP descriptor is used to obtain the compositional information of amino acids from a peptide sequence by varying the space (K) between the pair of amino acids [54], [87]. A sequence of peptide has 400 (20×20) types of different amino acid pairs, for instance, $(AA, AB, AC, \ldots, YY)_{400}$, when taking $K = 0$. KSAAP can be expressed as [88]:

$$KSAAP = p_i\{K\}p_j \quad (i, j = 1, 2, 3 \cdots 20) \quad (10)$$

where K represents the space between pairs of amino acids p_i and p_j.

KSAAP behaves like dipeptide composition of 400 dimensions of 20×20 pairs with value of K equal to 0, while K can be adjusted to any optimal value to obtain useful ordered information of amino acid pairs. As discussed above, this kind of compositional information may be useful to categorize different sequences. Refer to Hasan *et al.* [88] for further detailed process. KSAAP encoding can be computed by using iFeature, available as a standalone tool and web server as well [89].

g: PSEUDO-AMINO ACID COMPOSITION (PSE-AAC)

The AAC, DC and TC were practiced widely in the field of bioinformatics for the classification or prediction of several proteins/peptides, while these compositions may lose the sequence order information. To address the issue, Pse-AAC was proposed [49], [90]. Unlike the previously described compositions, the Pse-AAC takes into the account of sequence order information of amino acids. Pse-AAC rapidly gain increased use in computational proteomics and Bioinformatics [83] later-after it was proposed by Chou [91]. The Pse-AAC measures the correlation among different ranges of amino acid pairs which may vary in the different classes of peptides. According to the Chou's concepts Pse-AAC can be expressed as [92]:

$$P = [p_1, p_2, \cdots p_{20}, p_{20+1}, \cdots, p_{20+\lambda}]^T \quad (11)$$

where the first 20 components represent the composition of the 20 naturally occurring amino acids similar to AAC. Whereas the 20 + 1 to 20 + λ are additional components describe the correlations and sequence order information of amino acids via several modes along the length of the peptide sequence, and T is the transpose operator. Equation (10) can underline all the modes of Pse-AAC as [92], [69]:

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{\lambda} w_j p_j}, & (1 \le k \le 20) \\ \dfrac{^{w}(k-20)^{p}(k-20)}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{\lambda} w_j p_j}, & (20+1 \le k \le 20+\lambda) \end{cases} \tag{12}$$

where $f_i\,(i = 1, 2, \cdots, 20)$ is same to AAC representation and $p_j\,(j = 1, 2, \cdots, \lambda)$ represents the pseudo amino acid components, and w_j is for weight factors. Refer to [49], [90]–[92] for further detailed descriptions.

Currently, three compelling tools; ''propy'' [93], and ''Pse-AAC-Builder'' [94], to calculate ''Pse-AAC'' based features in several forms [90], while ''Pse-AAC-General'' [95] has been developed to derive general features based on ''Pse-AAC'' [92]. Furthermore, two highly robust, and effective, publicly available servers had been implemented entitled as ''Pse-in-One'' [96] and its next variant ''Pse-in-One 2.0'' [97] to extract features from given RNA, DNA, and peptide/protein sequences according to the requirements. Refer to [91] and [90] for additional details and comprehensions.

h: AMPHIPHILIC PSEUDO-AMINO ACID COMPOSITION (AM-PSE-AAC)

Various kinds of peptides have several characteristics like lipophilic and hydrophobic properties, which are commonly called as amphiphilic (AM) property. This AM property plays a central part in protein-folding, which determines its interaction and functional activity. To utilize these important features, Chou suggests a new scheme named as Am-Pse-AAC into the array of encodings [56], [98]. Similar to Pse-AAC, the Am-Pse-AAC also take into account the order information of peptide sequences. Am-Pse-AAC is a 20 + 2 λ dimensional feature vector, where the former 20 represents the composition of 20 natural amino acids, whereas the rest 2 λ numbers are to describe and determines the correlation factor of amphiphilic property distribution in a peptide/protein and can be formulated as [56]:

$$P = [p_1, p_2, \cdots p_{20}, p_{20+\lambda}, p_{20+\lambda+1}, \cdots, p_{20+2\lambda}]^T \tag{13}$$

where $p_1, p_2, \cdots p_{20}, p_{20+\lambda}$ are the amino acid compositions, $p_{20+\lambda+1}, \cdots, p_{20+2\lambda}$ the sequence correlation factors and T is the transpose operator. Refer to [98] for further details and mathematical calculation.

i: COMPOSITION-TRANSITION-DISTRIBUTION (CTD)

As the name suggests, the CTD feature descriptor comprised of three different attributes; composition, transition, and distribution to represent the global description of a given peptide sequence [99], [100]. The design goal of CTD is to derive the significant biological information from the amino acid sequences. For feature description, peptide sequences (amino acids) are divided into groups according to their property types and so numerically encoded by *1, 2, …, n* according to the group to which it belongs. These properties may be charge, hydrophobicity, secondary structure, polarity, polarizability, solvent accessibility, and normalized van der Waals volume, etc. [16], [101]. On the basis of these groups Dubchak *et al.* [102] proposed CTD to encode features on the basis of global composition, sequence order, and physicochemical properties of amino acid sequences to predict different protein folding classes. Afterward, CTD has been applied in several sequences based classification models [12], [87].

In CTD, the C (Composition) computes the frequency of every amino acid group with some particular property in a peptide sequence, can be defined as [52]:

$$CTD_C = \left(\frac{a_1}{L}, \frac{a_2}{L}, \cdots, \frac{a_n}{L}\right) \tag{14}$$

where a_i, $i \in \{1, 2, \cdots, n\}$ is the number of groups, n is the total number of groups, and L is the length of the sequence.

The T (transition) describes the occurrence of amino acids of a particular property accompanied with the amino acids of another property, which can be expressed as [52]:

$$CTD_T = {}^{(a_{i,j} + a_{j,i})}\!/_{L} - 1 \tag{15}$$

where $i, j \in \{1, 2, \cdots, n\}$ represents the corresponding property group, n is the total number of groups, L is the length of sequence, and $a_{i,j}$ represents the number of dipeptide having amino acid residues belongs to two different groups.

Lastly, the D (distribution) is the distribution measure of all amino acids belongs to a particular property group along the sequence length, measured by the location of first, 25%, 50%, 75% and 100% of the amino acid sequence [99], [102], which can be formulated as [52]:

$$CTD_D = \left(\frac{a_{1,1}}{L}, \cdots, \frac{a_{1,n}}{L}, \frac{a_{2,1}}{L}, \cdots, \frac{a_{2,n}}{L}, \cdots, \frac{a_{n-1,1}}{L}, \cdots, \frac{a_{n-1,n}}{L}\right) \tag{16}$$

where $a_{i,1}$ is the sequence length in which the first i^{th} group residue is placed. Similarly, $a_{i,2}, a_{i,3}, \ldots, a_{i,n}$ are the measures of i^{th} group residues in the sequence length at the positon of 25%, 50%, 75% and 100% respectively. Referred to [52], and [101] for additional details.

j: PHYSICO-CHEMICAL PROPERTIES (PCP)

PCP is a most instinctive descriptor related to physicochemical properties, where the residues of a peptide translated into a specific property class. Numerous bio-chemical and bio-physical properties have been determined experimentally determined properties. Amino acid sequence based properties alone is sometimes insufficient to categorize the peptide sequences. In such a case the blend of several physicochemical and structural features like amino acid composition, amphiphilic and hydrophilic property, size, overall charge and

secondary structure are useful to represent inclusive characteristics of a peptide [36]. For instance, the prediction of AMP has been carried out using physicochemical properties and obtained good results by Torrent *et al.* [42].

Various tools have been established to obtain physicochemical properties. Khawashima *et al.* [103] established amino acid index (AAI) database containing bio-chemical and bio-physical properties as numerical indices [29]. The AAI is classified into 3 divisions: AAindex1 (AAI-1), AAindex2 (AAI-2), and AAindex3 (AAI-3). The AAI-1 comprised of biochemical characteristics of 20 natural amino acids that can be represented as numerical values and referred to as an amino acid index. Currently, AAI-1 has 544 amino acid indices. The AAI-2 provides aggregates of various substitution matrices like BLOUSM62 or PAM26 [29] and currently have 94 substitution matrices; 67 symmetric and 27 non-symmetric. The last AAI-3 section is to provide protein contact potentials [103], so empirical values for spatially adjacent amino acids, like Gibbs energy change, to specify the interactions among amino acid pairs [29]. AAI-3 currently has 47 entries with 44 symmetric and 3 non-symmetric amino acid contact potential matrices [103].

k: MOTIF BASED FEATURES (MBF)

A motif is a repeated pattern of sequences in a collection of peptide/protein. Previously, various bio-informatics articles have examined motifs pertaining immunological characteristics in peptides for several purposes [104]–[106]. The repeated pattern (motifs) in amino acids having specific properties are responsible for the interaction of peptides. Therefore, the occurrence of these particular repeated patterns can be used in the classification and identification of the peptide types subject-wise [13]. Identifying positive dataset motifs by comparing with motifs of negative dataset provides high-class motifs for positive dataset characterization. Numerous publicly accessible tools were developed to determine motifs, i.e. MERCI [107] and MEME [108]. MERCI is one of them, uses physicochemical properties to identify motifs for a given class of amino acids. It provides functionally relevant, recurrent and high-class motifs of the positive dataset which are not available in negative sample [13], [107]. MEME-Suite is an integrated platform with web interface to provide multiple functions like database searching (motif-sequence and motif-motif), and the primary one motif detection [36], [108].

l: SEQUENCE ALIGNMENT (SA)

Sequence alignment in bio-informatics is a method of organizing nucleotides or amino acid sequences to recognize similar areas in sequences [109]. This similarity may be due to evolutionary relationships or based on structural or functional associations. The sequences are presented as rows of the matrix, where spaces are inserted within the residues so that columns may be aligned based on similar characteristics. The segments of a sequence that have high similarity are expected to have relatively same structural and functional activities [110]. For example, the peptide sequence P is considered to share the same characteristics as of Pk in a given set of peptides ($P1$, $P2$, ..., Pn) if P obtains high-scoring segment pairs (HSPs) score among the query peptide P and the peptides set. This similarity measure can be obtained by using several well defined tools [39].

Various sophisticated alignment based tools have been created like BLAST [111], BLASTP [112], HMMER [113], FASTA [114], PSI-BLAST [115], and Smith-Waterman algorithm [116]. These sequence alignment programs have been extensively practiced for protein/peptide prediction previously [36], [39], [110], [117]. Alignment based approaches are fairly simple and straightforward, however, failed to perform with a peptide sequence that did not have any substantial similarities to any known peptides [60].

m: CHAOS GAME REPRESENTATION (CGR)

CGR was suggested by Jeffrey in 1990 specifically to represent DNA sequences [118] and various generalize variant of CGR has been reported to represent amino acid (protein/peptide) sequences both in numerical and graphical form for their analysis. Different CGR's are outlined for different sequences in the injective behavior of the repeated function, which is quite an important characteristic of CGR [57]. It has been analyzed that, different kind of proteins/peptides demonstrate specific CGR patterns. This CGR feature established its effective use in the identification of new members of the specific peptide type. CGR has further benefits over the sequence based alignment tools usually use to search sequence homology that the CGR has a great potential to unveil the functional and evolutionary association among the proteins/peptides that have no substantial sequence similarity (homology) [119], [120]. C-GRex [121] is a standalone tool available to derive the CGR features.

The sequence mapping with CGR is invertible, therefore the original sequences can be recovered. CGS can be mathematically represented by a repeated function as [57]:

$$x_0 = (0, 0) \quad and\ x_n = (x_{n-1} + v_n)1/2 \tag{17}$$

where v_n represents the nth vertex related to the n^{th} base.

n: DISCRETE FOURIER TRANSFORMATION (DFT)

DFT is used to discover the periodic patterns and patterns relative strength by transforming the numerical data into the frequency domain. In bioinformatics, the DFT is applied nicely in bioinformatics like image processing, DNA hierarchical analysis, gene prediction, protein/peptide sequence analysis and prediction and to identify the coding regions. It is successfully applied to DNA, proteins and peptide sequences because the DFT spectrum of sequences represents the distribution and periodic patterns in the sequences [122].

The amino acid sequences of proteins or peptides are first transformed into numerical representations usually using physicochemical properties like hydrophobicity, hydrophilicity, etc. Then this numerical information is transformed to its corresponding frequency domain using discrete Fourier transformation (DFT) to obtain discrete components, which

may reveal the hidden significant features from the sequences without loss of information [52], [123], [124]. The frequency domain shows the discrete component's periodicity and power distribution contained in the peptide sequence over frequencies [52]. The DFT of a peptide sequence at frequency k can be calculated as:

$$f_k = \sum_{n=0}^{L-1} H(P_n) e^{-j\frac{2\pi}{L}nk}, \quad k = n = 0, 1, 2, \cdots, L-1 \tag{18}$$

where f_k denotes the compositional patterns and periodic properties of the sequence by sinusoidal components with various frequencies and $H(P_n)$ represent the values of physicochemical properties of each amino acid in the given peptide sequence of length L. The power spectrum of DFT with frequency k can be calculated as:

$$PS_k = |f_k|^2, \quad k = 0, 1, 2, \cdots, L-1 \tag{19}$$

o: K-MER COMPOSITION (KMC)

K-mer composition is used to encode the frequency of subsequences (K-mers) of the length k exists in the biological sequence of DNA, protein or peptide [46], [125]. For example, for a peptide sequence (ADAAYAAC), the k-mers of length $k = 2$ will be AD, DA, AA, AY, YA, AA, AC. In these subsequences AA has a frequency of 2 while all other k-mers appear once. K-mer counting used in bio-informatics for several purposes like sequence analysis, multiple sequence alignment, substring matching, and recurrent pattern detection [126]. Several publically accessible tools are available to obtain k-mer occurrences of the given peptide sequences, such as word2vec [46], [127], KMC [128], KAnalyze [129], Pse-in-One [96], Pse-in-One 2.0 [97], k-mer in scikit-bio (http://scikit-bio.org/) and kmer of cran project (https://cran.r-project.org/web/packages/kmer/index.html).

p: SECONDARY STRUCTURE BASED ENCODINGS (SSE)

Several studies have reported the use of structural features for the classification of proteins or peptides [40], [51]. The simplest structure of a protein or peptide is termed as the primary structure, consists of a sequential order of amino acids. While the secondary structure is the higher level representation of the protein or peptide. The secondary structures are highly associated with the functional activity of proteins or peptides. For instance, Alpha-Helix, and Beta-sheet are two core categories of secondary structure. The combine use of both (primary and secondary) structure-based descriptor enhances the overall accuracy by allowing the generalization capabilities of the classifier [29].

Numerous tools are available to predict the secondary structure from a given amino acid sequence such as Spider2 [130], PEP2D (http://crdd.osdd.net/raghava/pep2d/), and Protein secondary structure prediction (PSSpred) [51]. Khatun *et al.* use Spider2 and PEP2D to obtain structure-based descriptors such as alpha-Helix, beta-Sheet, and random Coil along with accessible surface area (ASA) and beta-torsion angles [54].

q: PROTEIN RELATEDNESS MEASURE (PRM)

Currently, the sequence alignment methods being used to measure protein's similarities are usually slow and required assumptions and evolutionary mechanisms about relationships. Carr *et al.* proposed PRM, which exhibits quickness as well as usefulness in the revelation of the functional and phylogenetic connections among protein/peptide sequences [131].

The PRM property is used to measure the degree of deviation of amino acids in a specific sequence compared to theoretical protein/peptide [63]. PRM descriptor is comprised of three measures; Type I (compositional), Type II (centrodial), and Type III (distributional). The first is to estimate the degree of deviation from the expected proportion of amino acid in the sequence. Second is to estimate the presence of some particular amino acids in the specific area of a sequence. And the last is to estimate the proportion to which amino acid groups occur alongside the length of the amino acid sequence [63], [132]. PRM is a 60-dimensional feature vector contains 20 dimensions for each measure. Referred to [131] for detailed calculation process and source code in C++ and Mathematica to encode PRM feature.

r: BINARY PROFILE FEATURE (BPF)

Proteins/peptides are based on the combination of several amino acids, that signifies their functional activities. In this descriptor each residue is encoded in binary form (0/1) according to amino acid type out of twenty natural amino acids already mentioned above. The presence of an amino acid is coded as a function of each amino acid in order form as f(A) = (1, 0, . . ., 0), for the first type of amino acid "A" and f(c) = (0, 1, . . ., 0), for the second type "C" and so on [64]. For a certain peptide sequence with amino acid length K, BPF can be encoded as [16]:

$$BPF_k = \{f(p_1), f(p_2), \cdots, f(p_k)\} \tag{20}$$

where k is the length of peptide sequence as a whole or can be specified in the form of N-terminus (amine-terminus/start of the amino acid chain) or C-terminus (carboxyl-terminus/end of the amino acid chain) of protein or peptide sequence [16], [64]. Therefore, the BPF is a $20 \times k$ dimension descriptor.

Numerous common tools or web servers are available to encode several features from the given protein/peptide sequences. For instance, anyone of propy [93], iLearn [133], iFeature [89], Pfeature [134], Protr [135], PyBioMed [136], PseAAC-General, Pse-in-one, and Pse-in-one 2.0 can be used to obtain feature descriptors of type AAC, DC, TC, Pse-AAC, Am-Pse-AAC, and CTD, however BPF can be derived by using anyone of the iFeature, Pfeature, Pse-in-One, and Pse-in-One 2.0.

4) ASSESSMENT OF QUESTION 4. WHAT ARE THE MAJOR ML CLASSIFIERS USED IN PEPTIDE PREDICTION MODELS?

In the answer of the previous question, we have extensively elaborate the feature encoding schemes. Likewise, in this

section we will elaborate the various ML classifiers, conventional as well as advanced, used in the field of study. Numerous statistical or artificial intelligence (AI) based algorithms such as logistic regression and ML-based methods have been used in literature to build classifiers for the prediction or classification of peptides and protein to explore the functional activities [137].

Machine learning algorithms such as Support Vector Machine (SVM) [55], [57], Random Forest (RF) [12], [38], and Advanced Machine Learning like Neural Network (NN) [42], [50] and Deep Neural Network (DNN) [46], [64], etc. All the classifiers that have been identified in this study are shown in Table 10 along with their references. SVM has been identified as the widely used algorithm, whereas RF is the second-most used classifier. However, the combinatorial use of different classifiers has also been reported, used as an ensemble classifier.

TABLE 10. Summary of ML classifiers used in selected articles.

Sr. No.	*Classifiers*	*Used in*
1	SVM	[13], [16], [15], [43], [36], [40], [44], [47], [48], [52], [53], [55], [57], [59], [60], [61], [62], [63], [65], [66], [67], [69], [70]
2	RF	[12], [38], [44], [45], [3], [47], [49], [51], [52], [54], [56], [59], [66]
3	DNN	[37], [46], [64], [50]
4	ANN	[42], [70], [52]
5	CRF	[41]
6	LR	[52], [68]
7	NNA	[39], [59], [70]
8	IDQD	[20]
9	NB	[52]
10	DT	[58]

a: SUPPORT VECTOR MACHINE (SVM)

SVM is an ML-based algorithm works under the supervisory mode and trained with the labeled dataset to produce a model which can then be utilized to determine the correct label for unmarked data [16]. SVM practiced a statistic-based learning concept, originally used for twofold classification and was then adept successfully for multiclass problems [66]. SVM utilizes the basic concept of hyperplane to maximize the separation-margin among the classes of the dataset to classify. The optimal separation is achieved by converting the inputs into a higher feature order [67].

But it usually happens that the proposed feature set is not distinct and discrete. This issue is being solved by using the kernels to change the primary non-separable characteristics into a linear separable feature set, where an optimal hyper-plane classification is possible. Four classes of core functions that often employed for this purpose is a linear function, radial function (RBF), polynomial function, and sigmoid function [16], [67]. However, the most commonly used function is RBF in the classification of various therapeutic peptides.

b: RANDOM FOREST (RF)

The RF model uses two powerful (regulatory) procedures: bootstrap aggregation (bagging) and a random selection of feature, so also considered as an ensemble classifier works in the supervisory mode. It is widely practiced algorithm to solve the problems related to regression as well as classification. RF model builds on the basis of growing decision trees and split sampling technique with a replacement policy (bootstrap-sampling). Training dataset splits into two parts, one is used for model training, while the remaining (usually one third of the whole dataset) portion is used to test the model. This approach is known as out of the bag (OOB) approach offers better prediction performance and improve the reliability of the model for a dataset [3], [45].

c: ARTIFICIAL NEURAL NETWORKS (ANN)

NN is another machine learning model to provide artificial intelligence based solutions and extensively practiced in the area of Bioinformatics to recognize patterns, classification, and etc. [138]. In general, NN is a collection or web of artificial neurons or nodes, an artificial representation of the human brain, so usually termed as artificial neural network (ANN). These artificial neurons are interlinked in various ways to communicate with each other arranged in connected layers [139]. Input layer is the first layer, where the number of neurons are related to the number of input variables. After the input layer there could be zero or more hidden layers, while the last layer is termed as output layer where the number of neurons depends on the number of outputs, in example one for binary classification and could be more than one for multi-class labeling [139], [140]. The flow of information in NN is as follows: the information patterns (inputs) are accepted at the input layer of NN, which activates the neurons of hidden layer for outcome estimation and pass these outcomes to the output layer.

The neurons are computational/mathematical components, takes input and passed to the other neurons after performing the computation. The links among neurons have connected weights, where positive values reveal the exciting contact while non-positive considered as hindering. There are two phases of an ANN; learning (training) phase and operating phase (capable of classifying the patterns after being learned). In learning phase, NN is being trained with given inputs and labeled data are provided in advance, transform data by multiplying each node value with the relevant link's weight. Further, summing the outcomes of same node, regulate the resultant with node specific bias value and lastly, the output is achieved by normalizing the resultant value with a function termed as ''activation function'' [138], [140]. Mostly used activation functions are ''sigmoid'', ''hyperbolic tangent (TanH)'', radial basis function (RBF), and ''rectified linear unit (ReLU)'' [37], [46], [64]. To avoid overfitting, dropout layers are generally implemented in NN [46].

Feedforward neural network (FNN) [42], radial basis function (RBF) network (RBFN), generalize regression neural network (GRNN) and probabilistic neural network (PNN) [70] are variants of ANN being used in the selected studies. FNN has been used for several types of classification models. FNN is a simple neural network with one input, sigmoid hidden and output layers. The flow of information is unidirectional, from input through a hidden layer and then to output layer [42], [141]. Similar to FNN, the RBFN also comprised of three layers; the input, hidden with RBF (a non-linear activation function), and a linear output layer [52]. GRNN/PNN can be used for classification and decision making problems and both have a similar architecture comprised of four layers; the input, hidden, summation/pattern, and decision layers [70], [142], [143]. The input and hidden layers are same in both neural networks relative to functionality, while there is a difference in numbers and functions of neurons in summation/pattern, and decision layers. For instance, summation layer in case of GRNN has only two neurons; the denominator neuron and numerator neuron while PNN has one neuron for each target class. Furthermore, GRNN is used to implement regression problems that have continuous target variables [142], while PNN for classification with categorical target variables [143]. For further detailed descriptions refer to [142]–[145].

d: DEEP NEURAL NETWORKS (DNN)

DNN is an enhanced form of ANN comprised of multiple or higher number of network layers also known as the multilayer perceptron (MLP) uses well defined and improved algorithmic design. The enhanced architecture of DNN able it to extract best representative features of the given data set without any manual involvement or guide plan. Adaptive neuro-fuzzy inference system (ANIFS) [50], convolutional neural networks (CNN) [46], long short term memory (LSTM) a type of recurrent neural network (RNN) [37], [46], [64], are some variants of deep neural networks mostly used in the selected studies for their proposed prediction models, explained next.

CNN is a neural network, modeled on the basis of human visual system, applied in various fields to predict, recognize or identify targeted object(s) from the given input. CNN architecture typically comprises on multiple layers including the combination of several convolutional, pooling (subsampling) and dropout (regularization) layers accompanied by fully connected layer(s) [116].

The grouping of several convolution layers provides highly refined features at each layer level through the input layer to the output layer. Pooling or dropout layers are often interleaved among convolutional layers. The purpose of pooling layer (max-pooling or average-pooling) is to subsample or reduce the input representation. On the other end, the dropout layer is used to reduce the unnecessary dependencies of network on the features by dynamically lowering the number of neuron connections to avoid the overfitting and also reduce the training time. Thus, the neurons in each convolutional (feature extraction) layer are not fully connected. The final layers of CNN are fully connected layers and responsible for classification on the basis of collective features retrieved from previous layers [46], [139], [146].

LSTM is an advanced type of RNN that have feedback connections and allowing to store previous states in network memory. This state awareness feature of LSTM distinguishes it from other NN models and supports in the training process by incorporating previous state history when required. LSTM use gradient descent and time-based backpropagation training algorithm and also address the gradient vanishing problem by allowing the gradients to pass unaltered, which was a limitation of RNN [64], [139]. LSTM architecture consists of memory blocks (LSTM units) and control gates (input, forget, and output gates) to control the information flow between LSTM units. LSTM based models can monitor and retain the long-standing states or dependencies. That's why they are well learned from sequence inputs and building patterns that depends on context and previous states. The LSTM memory block keeps the relevant data from previous states to utilize in future calculation when required. The input gate determines the extent to which a new value flows into a cell, the forget gate determines the extent to which the value remains in the cell, and the output gate determines the usage of block value to calculate the output [37], [64], [139].

The ANIFS is a type of deep neural networks; an ensemble of adaptive control, fuzzy system and artificial neural networks with fuzzy inference system (FIS) as hidden layer [50], [147]. Fuzzy inference systems are widely used to understand the behavior of systems that are very easy to interpret [148]. These systems models human-like reasoning using plausible semantic terms. The specialist can develop the basic knowledge base using the predefined rules in the inference system, whereas the system is incomprehensible if no specialist is available [149]. Neural networks, on the other hand, have capabilities of learning, but cannot be interpreted [148]. However, due to this neuro-fuzzy combination and adaptive control technique, the ANFIS have the capability of both, the fuzzy logic (reasoning and interpretability) and neural networks of learning [147], [150].

e: CONDITIONAL RANDOM FIELD (CRF)

CRF models are statistical models deployed for various machine learning and pattern recognition based prediction problems. For example, classifying biological sequences [151], functional activity identification of peptides [41] and etc. It is an undirected graphical discriminative model overcome the biases in labeling. The CRF has the capability to predict a data label without considering the adjacent, however, could work in a contextual manner by realizing the perdition dependencies in a graphical model. For details, see [41].

f: LOGISTIC REGRESSION (LR)

LR is another statistical based classifier used to model the likelihood of some class, in example the input value belongs

to a class under a determination or not [68]. LR is based on the logistic function (sigmoid), which converts any real value in between 0 and 1. LR provides dualistic associations among dependent and independent variables, however, extensions for multi-label classifications are also exists [52], [152].

g: NEAREST NEIGHBOR ALGORITHM (NNA)

NNA is a non-parametric and modest learning method used effectively for the recognition of patterns. It works on the basis of weights of its specified ''k'' nearest neighbors and classify an unidentified sample into a class, where k could be some optimum value for the neighbor to consider for weight-based membership calculation. The weight measure is usually a degree of distance among the unknown sample and its neighbors to classify [39].

h: INCREMENT OF DIVERSITY WITH QUADRATIC DISCRIMINANT (IDQD) ANALYSIS

Quadratic discriminant (QD) analysis is a modified form of linear discriminant analysis (LDA) that evaluates a covariance model for each observation class. QDA is especially beneficial in a situation when there is a piece of prior knowledge about each class that shows quite distinctive behavior [153]. Whereas IDQD has two fragments, one is an increment of diversity (ID) used to extract diverse information from the sample sequences as a class of sequence can be predicted with such information and the second fragment QD provides a scheme by assimilating the various extracted features into a system to predict [20].

i: NAIVE BAYES (NB)

NB is normally recognized as a simplest classifier effectively used in the field of bio-informatics [52], based on Bayes theorem [154]. Bayes theorem supports to find an occurrence probability of one variable on the basis of other when given. NB adopts robust assumptions about the independence of each feature to contribute its part equally and independently to get optimum results, so reduced the complications to develop a classifier [52].

j: DECISION TREE (DT)

DT is a classifier, effectively used in numerous fields to solve the prediction, classification or recognition problems. Conceivably the supreme property of DT is to provide easily deducible results by breaking-down the complex problems into simplest ones. It works great with both classification and regression related problems. DT has various variants and implementations, like Gordon *et al.* introduced ''CART'' (classification and regression tree) for analysis, having both procedural capabilities of classification and regression [155]. Greedy algorithm is used for learning purpose to decide splits in a tree at and optimum level and Gini-index to measure the degree of inequality.

''PART'', a partial decision tree, an algorithm that provides appropriate decision rules without considering global optimization [156]. ''RPART'' (recursive partitioning) is another R package specific implementation of ''CART'' [157]. Another ensemble classifier named as ''Deep-Boost'', use decision tree as a member, have achieved better accuracy by handling the data over-fitting problem [158]. Zahiri *et al.* used a combination of classifier by using ''PART'', ''RPART'' and ''Deep-Boost'' for their prediction model.

5) ASSESSMENT OF QUESTION 5. WHICH MAIN PUBLICATION SOURCES ARE BASICALLY TARGETED FOR THE PUBLICATION OF ARTICLES IN THE SELECTED AREA OF STUDY?

The motivation behind this RQ is to discover the key publication sources from where the domain related articles can be acquired, and eventually the identification of excellent channels to publish articles to support the research community to have a predetermined source set. Table 11 represents the summarized description of target publication sources and the articles published in each source along with the (JCR/CORE) rank to present the quality of each selected article.

Of the total fourty-one selected articles, thirty-seven (93%) were published in journals and only three (7%) articles were presented in conferences. Moreover, out of total, fourty-one articles 76% (31 articles) were published in highly recognized journals with Q1 JCR rank, 15% (6 articles) publication sources have Q2 JCR rank, only 2% (1 article) have Q3, and all the conference articles, 7% (3 articles) as whole have B4 CORE rank, as shown in Figure 6. It can be analyzed from the facts presented in Table 11 and Figure 6 that the articles in the selected field were normally published in well-reputed publication sources.

IV. DISCUSSIONS

This section provides a detailed discussion of different ML based predictors applied in different areas of the selected domain. In the first instance, a feature encoding taxonomy has been proposed to sum up the results of this study. The taxonomy is depicted in Figure 7. In addition, a standard framework is proposed to support the development of a prediction model for the identification of peptides, shown in Figure 8.

A. TAXONOMY OF FEATURE ENCODINGS

Feature encodings identified in this study are being classified in hierarchal manner and presented in the taxonomical form here. Initially the feature encodings are classified in three major modes; structure based, property based, and sequence based, centered around the way of their formulation as recognized from the selected articles. The hierarchy is being tried to be developed on the basis of parent-child relationship. All these encoding approaches have already been discussed extensively in section III and its hierarchal classification is presented here in Figure 7.

B. FRAMEWORK FOR PEPTIDE PREDICTION MODEL

Proficient manual examination could serve better for the likelihood of candidate peptides in therapy of several diseases. Though, computational prediction is better than manual

TABLE 11. Summary of publication sources and their JCR/CORE ranks of selected articles.

Publication Source	JCR / CORE Rank	Publication Type	References	No of Articles	Ratio (%)
Bioinformatics	Q1	Journal	[16], [37], [3]	3	7.32%
Nucleic Acids Research	Q1	Journal	[36], [38]	2	4.88%
PLoS ONE	Q1	Journal	[15], [39], [40], [41], [42], [51]	6	14.63%
Journal of Biomedical Informatics	Q1	Journal	[43]	1	2.44%
Oncotarget	Q1	Journal	[44], [60]	2	4.88%
Frontiers in Pharmacology	Q1	Journal	[12]	1	2.44%
International Journal of Molecular Sciences	Q1	Journal	[45], [56], [61]	3	7.32%
BMC Bioinformatics	Q1	Journal	[46], [55]	2	4.88%
Molecules	Q1	Journal	[47]	1	2.44%
Computers in Biology and Medicine	Q2	Journal	[48], [49]	2	4.88%
Biopolymers	Q2	Journal	[50]	1	2.44%
Scientific Reports	Q1	Journal	[52], [53], [59]	3	7.32%
Frontiers in Genetics	Q1	Journal	[54]	1	2.44%
Journal of Mathematical Biology	Q1	Journal	[57]	1	2.44%
Journal of Translational Medicine	Q1	Journal	[13], [58]	2	4.88%
BioMed Research International	Q2	Journal	[62]	1	2.44%
International Journal of Peptide Research and Therapeutics	Q3	Journal	[63]	1	2.44%
Molecular Therapy - Nucleic Acids	Q1	Journal	[64]	1	2.44%
Journal of Theoretical Biology	Q1	Journal	[69]	1	2.44%
Artificial Intelligence in Medicine	Q1	Journal	[70]	1	2.44%
IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2012	B4	Conference	[65]	1	2.44%
IEEE/ACM Transactions on Computational Biology and Bioinformatics	Q2	Journal	[20], [66]	2	4.88%
International Conference on Biomedical and Bioinformatics Engineering, ICBBE 2017	B4	Conference	[67]	1	2.44%
ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics, ACM-BCB 2013	B4	Conference	[68]	1	2.44%

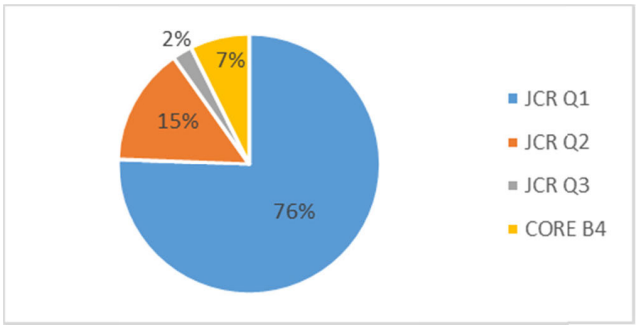

FIGURE 6. Ranks of publication sources of selected articles.

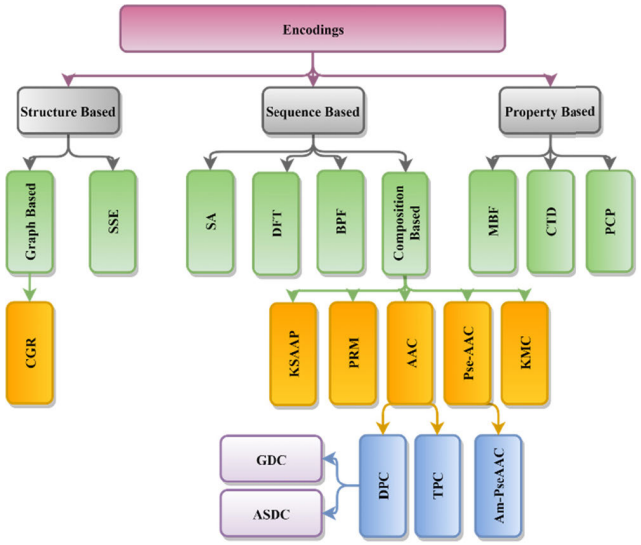

FIGURE 7. Feature encoding hierarchy.

inspection due to several reasons. It has been identified computationally that peptide sequences with several therapeutic activates share very low similarity with other peptide sequences have same functional properties. Consequently, it's quite difficult for an expert human to identify manually. Furthermore, high performance of computational prediction models permits the exclusive discovery and selection of unforeseen proteins/peptides predecessors. This phenomenon provides provision in an effective manner to experimentally validate the discovered predecessors in labs before the design of drugs [159].

A generic framework to build a prediction model is being proposed and depicted in Figure 8. The proposed framework comprised on five staged processes; dataset collection, dataset preprocessing, feature (encodings) selection and optimization, selection of ML classifier (MLC), and implementation of the finalized model in the form of a publicly accessible web-server or tool for experimental purposes.

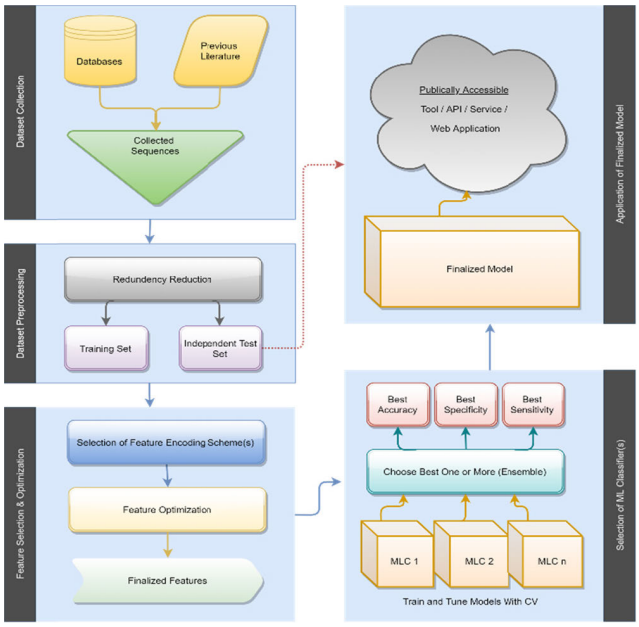

FIGURE 8. Framework for ML-based therapeutic peptide prediction.

Dataset collection is the primary but cumbersome task to build a successful predictive model, as there is no automated procedure is available for the collection of related datasets. Researchers gather it by examining extensive published articles, inspect, and curate the dataset to use in model, also some groups convert these collections into publicly accessible databases for the researchers to reveal new ways. During this review the identified databases are listed in Table 7 in section III and the number of peptide sequences searched in the selected studies are mentioned in Table 12 along with their references.

To improve the credibility of prediction model it is necessary that the collected dataset should have a lower sequence identity (homology), which may lead to models biasness and eventually affects its performance, normed as accuracy. Consequently, the next step is the preprocessing of collected dataset to clean it up from the redundant and identical sequences. The accuracy of the prediction model is based on various parameters including the availability of experimentally validated datasets, preprocessing and elimination of redundant data. ''CD-HIT (Cluster Database at High Identity with Tolerance)'' is a freely available online suite mostly used in the domain to reduce data redundancy. It takes amino acid sequences (proteins/peptides) or nucleotide sequences (DNA/RNA) in Fasta format as inputs and produces non-redundant sequence sets (datasets) and clusters of sequencing as output at some provided identity-threshold. Only the sequences having high similarity than the provided threshold are removed to reduce the dataset size without losing useful sequence information [160], [161].

Table 12 further represents paper-wise datasets related useful information regarding the breakup of a benchmark dataset into training, independent/test dataset (including number of positive and negative samples), programming languages or Tools used, framework/package / library used, and sequence identity tool with the threshold value used to remove the duplication in the dataset. It is being analyzed from Table 12 that mostly used sequence identity threshold values ranging from minimum 40% (0.4) and 90% (0.90) as maximum in different articles according to the size of available datasets. It is being noticed that the lower threshold for sequence identity may increase the credibility of constructing models, but the smaller size of the dataset restricts to apply such a threshold value specifically in the case of peptide prediction [61]. After the cleanup, the dataset should be distributed in the training set and independent/test set. For such a distribution no criterion yet has been defined, however, according to the identified best practices, the ratio of the distribution may vary from 70:30 to 80:20 for training and independent sets respectively, depending on the size of the whole dataset. The training set will be used to train the ML classifier model while the successor will be used to evaluate the model's transparency and actual performance towards unseen data.

The next stage in the development of classifier is feature selection & optimization, which is a threefold process. As a first step, the feature encoding(s) (single/multiple) are being selected to convert the diverse nature of dataset (amino acid sequences) into fixed-length feature vectors, the ML Classifiers understandable. Primarily identified feature encodings during this study has been extensively discussed in section III and the usage percentage is presented in Figure 9.

Most frequently used feature in this domain is PCP with 21% usage share, AAC is second with 20%, DC at third with 14%, PseAAC and SSE have usage share of 7% and 6% respectively, among the selected studies. Mostly published articles utilized various combinations of feature encodings to optimize the prediction credibility. However, [46], [57], and [63] referenced articles utilized single feature encodings which are k-mer, CGR, and PRM respectively. The trade-off among the selection of local or global feature encodings is still an open challenge that needs attention.

Combinatorial use of diverse encodings has a great chance to contain noise and redundancy in feature attributes, which may affect the training of model and cause overfitting, later such trained models performed poor with the independent or unseen dataset. Subsequently, feature optimization methods are applied to finalize the effective features for model training. Various optimized feature selection techniques have been practiced by means of different feature ranked methodologies. At first input features are ranked using ''F-score algorithm'' [61], ''minimum redundancy maximum relevance (mRMR)'' [3], [16] or simply RF based ranking and finally the optimized feature vectors based on best ''area under curve AUC'' are being determined by using ''sequential forward search (SFS)'' method. RF model with Gini-importance or mean decrease of Gini-index (MDGI)'' based ''rigorous feature elimination (RFE)'' was adopted for highly ranked features selection in several studies [12], [45], [49], [51], [56]. Both, RFE and SFS have been practiced

TABLE 12. Article and domain-wise list of sequence identity threshold (use in CD-HIT), validation methods, number and distribution of dataset, and web server links.

Ref No.	Domain Type	Sequence Identity Threshold (<%)	Training Dataset		Independent / Test Dataset		Evaluation Technique	Programming Language / Tool	Framework / Package / Library	Tool / API / Service / Web Server	
			Positive	Negative	Positive	Negative	Models	(if provided)		Link (if provided)	Accessible
[16]	ACP	90	250	250	82	82	10-Fold CV	Weka	LIBSVM	http://server.malab.cn/ACPred-FL/	Yes
[36]	AMP	Not Described	544	407	120	105	5-Fold CV	SVM^{light}	Not Provided	http://crdd.osdd.net/servers/avppred	Yes
[37]	AMP	90	712	712	712	712	10-Fold CV	Python	keras, Tensorflow	http://aps.unmc.edu/AP, http://www.ampscanner.com	Yes
[38]	AMP	90	1805	2807	733	1204	10-Fold CV	R	Kernlab	http://www.camp.bicnirrh.res.in/predict/	Yes
[15]	AAP	70	107	107	28	28	10-Fold CV	SVM^{light}	Not Provided	http://clri.res.in/subramanian/tools/antiangiopred/	Yes
[39]	AMP	70	870	8661	1136		Jackknife CV	Not Provided	Not Provided	http://amp.biosino.org/	No
[40]	AMP	40	310	310	75	75	5-Fold CV	SVM^{light}	Not Provided	http:// sourceforge.net/projects/csamppred/	Yes
[41]	AMP	70	-	-	-	-	10-Fold CV	CRF++	Not Provided	Not Provided	No
[42]	AMP	Not Described	1157	991	-	-	Jackknife CV	Matlab	Not Provided	Not Provided	No
[43]	ACP	Not Described	225	225	-	-	10-Fold CV	Matlab	LIBSVM	https://sourceforge.net/projects/application-of-q-wiener-index	Yes
	AMP	Not Described	929	1263	-	-					
[44]	ACP	90	187	398	422	422	10-Fold CV	Python	scikit-learn	www.thegleelab.org/mlacp.html	No
[12]	AIP	80	1258	1887	420	629	5-Fold CV	Python	Not Provided	http://www.thegleelab.org/AIPpred/	Yes
[45]	AAP	90	181	181	55	57	5-Fold CV	R	Not Provided	http://codes.bio/targetantiangio/	Yes
[3]	ACP	Not Described	250	250	164	2082	10-Fold CV	Weka	Not Provided	http://server.malab.cn/PEPred-Suite/	Yes
	ABP		800	800	398	2199					
	AVP		544	407	120	2045					
	AIP		1258	1887	840	2629					
	AAP		107	107	56	2028					
[46]	ACP	Not Described	225	2250	50	50	Independent Test	Not Provided	Not Provided	Not Provided	No
	AVP / AMP	Not Described	469	703	138	206					
[47]	ACP	90	138	205	-	-	Jackknife CV	R	protr, seqinr, caret,	https://github.com/chaninlab/acpred-webserver	Yes
[48]	AMP	40	2704	5156	879	2405	5-Fold CV	Python	Not Provided	http://amap.pythonanywhere.com/	Yes
[49]	ACP	90	972	1172	324	391	5-Fold CV	R	protr, seqinr, randomForest,	http://codes.bio/thpep/	Yes
[50]	AMP	50	115	58	115	58		Matlab	Not Provided	Not Provided	No

Ref No.	Domain Type	Sequence Identity Threshold (<%)	Training Dataset		Independent / Test Dataset		Evaluation Technique	Programming Language / Tool	Framework / Package / Library	Tool / API / Service / Web Server	
			Positive	Negative	Positive	Negative	Models	(if provided)		Link (if provided)	Accessible
[51]	AMP	Not Described	544	951	120	105	Independent Test	R	randomForest	Not Provided	No
[52]	AAP	Not Described	107	107	-	-	10-Fold CV	Weka	Not Provided	Not Provided	No
[53]	ACP	Not Described	225	2250	50	50	10-Fold CV	SVMlight	Not Provided	http://crdd.osdd.net/raghava/anticp/	Yes
[54]	AIP	80	1258	1887	420	629	10-Fold CV	Weka, R	randomForest	http://kurata14.bio.kyutech.ac.jp/PreAIP/	Yes
[55]	AMP	Not Described	999	999	466	-	5-Fold CV	SVMlight	Not Provided	http://www.imtech.res.in/raghava/antibp2/	No
[56]	AMP	Not Described	544	951	120	105	5-Fold CV	R	Caret	http://codes.bio/meta-iavp/	Yes
[57]	ACP	90	225	2250	138	206	Jackknife	R	LIBSVM	Not Provided	No
[58]	AAP	Not described	108	108	27	27	Independent Test	R	Caret	https://cran.r-project.org/web/packages/AntAngioCOOL/index.html	Yes
[13]	AIP	Not Described	690	1009	173	253		R	Caret, SVMLight	http://metagenomics.iiserb.ac.in/antiinflam/	Yes
[59]	AAP	Not Described	107	105	-	-	10-Fold CV	R	Mlr	Not Provided	No
[60]	ACP	Not Described	138	206	150	150	Jackknife	Not Provided	LIBSVM	http://lin.uestc.edu.cn/server/iACP	No
[61]	ACP	80	256	256	157	157	10-Fold CV	Python	scikit-learn	http://www.thegleelab.org/mACPpred/	Yes
[62]	AMP	Not Described	870	8661	3556	-	Jackknife	Perl	LIBSVM	Not Provided	No
[63]	ACP	10	217	1000	40	40	10-Fold CV	Weka	WLSVM (Weka LIBSVM)	http://acpp.bicpu.edu.in/predict.php	No
[64]	ACP	90	376	364	129	111	5-Fold CV	Python	scikit-learn, keras, Tensorflow	https://github.com/haichengyi/ACP-DL	Yes
[65]	AMP	Not Described	-	-	-	-	-	Weka	Not Provided	http://ida.felk.cvut.cz/peptides/BIBM2012.zip	Yes
[66]	AMP	Not Described	569	1138	1524	-	10-Fold CV	R	e1071	http://www.bicnirrh.res.in/classamp/	Yes
[20]	AMP	40	787	-	-	-	Jackknife	Not Provided	Not Provided	Not Provided	No
[67]	ACP	Not Described	138	206	-	-	Jackknife	Not Provided	Not Provided	Not Provided	No
[68]	AMP	Not Described	115	116	216	145	Independent Test	Not Provided	Not Provided	http://binf.gmu.edu/dveltri/cgi-bin/AMP-PASS.cgi	No
[69]	ACP	90	138	206	-	-	5-Fold CV	Not Provided	LIBSVM	Not Provided	No
[70]	ACP	90	138	206	-	-	Jackknife	Not Provided	Not Provided	Not Provided	No

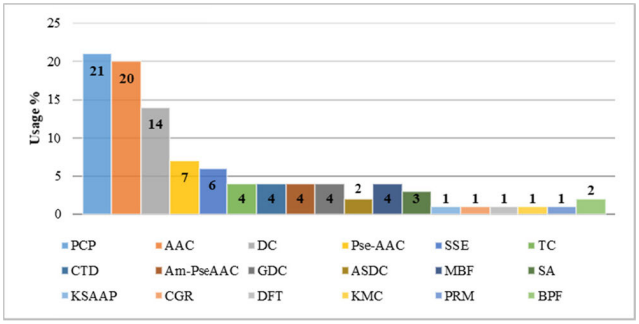

FIGURE 9. Usage percentage of feature encodings.

effectively, however, RFE has performance benefits over the SFS due to its non-complexed computational mechanism.

Next and most important phase is the selection of most appropriate ML based classification or regression model(s) according to the problem under consideration. The main focus behind the careful training of ML-based algorithm is that it should adopt the same predicted behavior on providing unknown dataset and classify it accurately. ML-based classifiers are being trained by providing the optimized features and known expected answers (labeled data) to make associations among the features and expected answers to learn. The most frequently used classifiers identified in the selected studies have been discussed in details in section III, and the usage share in the percentage of each classifier shown in Figure 10. Among the selected articles, 44% of the studies utilize SVM, second most frequent classifier is RF with 25% usage, DNN has 7% share, ANN and NNA has and equal share of 6% share, LR has a usage share of 4%.

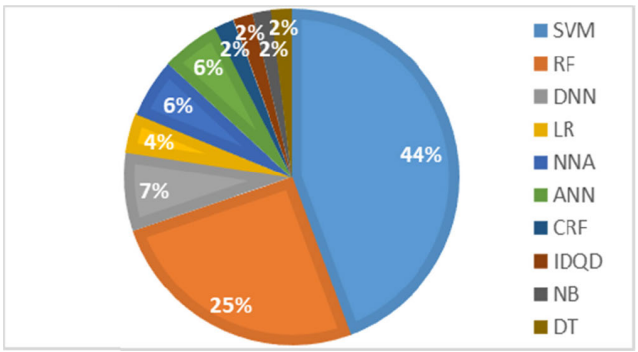

FIGURE 10. Frequency of ML classifier's usage.

Generally, a ML classifier has customary hyper-parameters, required regulation and tuning to obtain superlative grouping and therefore to acquire best prediction performance [59]. The ML classifiers learn using the datasets, and it's not feasible to use the same dataset for various purposes like, tuning of parameter, learning, and finally for validation. Normally, the cross validation (CV) methods are being used for training, validation and evaluation of success of a classifier. The main CV techniques are sub-sampling also known as "k-fold test" [12], [13], [44], "leave-one-out"/"jackknife" test [20], and independent test [45], [58] has often practiced in training, tuning and validation of classifiers. The use of these validation techniques is also represented in Table 12.

Numerous researchers practiced a single best fitted classifier for their prediction model, though, make a comparison of their results with other classifiers too. It has been advised here to train and test different ML-based classifiers, and decide the finalized model single or combination of multiple classifiers (classifier fusion) on the basis of performance. Some hybrid or customize use of several classifiers was also identified. For instance, D. Veltri *et al.* proposed a deep neural network model by employing embedding, convolutional and max-pooling layers for sequence pattern generalization and LSTM layers to characterize the given sequences [37]. Derived discrete features are fed to the embedding layer to convert into a fixed size representation which may semantically nearest to each other in the feature map. These features were convolved through the one dimensional convolutional layer having sixty-four filters and max pooling layer to down-sample the filter outputs of convolution layer. These downsized features were passed to a dense layer with sigmoid function through LSTM network to predict the outputs in 0-1 range and reported accuracy of the model is 91.01%. With the name of PTPD a deep learning based new computational model was proposed to predict therapeutic peptides. In PTPD, the first input layer was utilized to map the peptide sequences into a k-mer based numeric vectors by incorporating bag-of-words and continuous skip-gram (two learning algorithms) through word2vec tool. These numeric representations (vectors) were fed to a convolutional layer implemented with three types of 1-dimensional convolution filters of different sizes and ReLU activation function to perform convolution on the whole vectors to create feature maps. A dense layer with ReLU activation function was adopted to combine the results of three convolutional filters after dimensionality reduction of feature maps through max-pooling layer and dropout function to avoid overfitting. Finally, a fully connected layer with sigmoid function was adopted for output. According to Chuanyan *et al.*, PTPD effectively perform with 90.2% accuracy with several datasets. Torrent *et al.* developed an AMP prediction model by using a two-layer feedforward NN, sigmoid activation function in the hidden layer with 50 neurons and scaled conjugate gradient (SCG) based backpropagation algorithm to adjust the weights during learning process to minimize the error loss [42], and achieved 90% accuracy.

The classifier fusion approach was adopted by Zhang *et al.* [52] for AAP and Akbar *et al.* [70] for ACP classification models and achieved the accuracy of 83.2% and 96.45% respectively. Classifier fusion is the assembly of several individual basic classification models with different learning practices and then combines the results from all autonomous classification models to address the same classification problem. The outcomes in the sense of sensitivity and specificity from different individual models are usually different for a particular task. However, the fusion of best classifiers on sensitivity and specificity usually makes them a more suitable

classification performer than an individual classifier [52], [162], [163], which has also been proved theoretically by Hansen and Salamon [164]. Zhang *et al*. [52] assessed the performance of various classification models, including RBFN, NB, LR, NNA and RF. The final result was then determined on the basis of average probability of outcomes from an individual classifier, which is best in specificity (best to predict negative samples) and another, which is best in sensitivity (best to predict positive samples). Similarly, Akbar *et al.* [70] utilized SVM, RF, PNN, GRNN, and NNA as operative classifiers and the optimal results were obtained by fusing the outputs of the individual classifier with evolutionary genetic algorithm and majority voting.

In ML, the quality or prediction performance is measured using various parameters. One of them is confusion matrix, provides the measure of accuracy of identified values. The accuracy measure is not the only parameter that can be considered sufficient for the evaluation of performance of a predictor. Therefore, various parameters are being considered to measure performance: sensitivity, usually the percentage of truly identified positive samples, specificity measures the truly identified negative samples while is the measure of correctly identified samples (positive + negative), and generalized measure termed as ''Matthews correlation coefficient (MCC)'' generates a quantitative measure between $+1$ to -1 by considering proportion of whole set of true positive, false positive, true negative and false negative. The quantitative measure near or equal to $+1$ considered as accurate estimation, 0 for poor prediction, while -1 considered as complete divergence among the predicted and known annotations. Refer to [15], [41], [52], [67] for detailed overview and mathematical calculation. MCC might also be directly calculated using confusion matrix [165].

Receiver operating characteristic (ROC) curve has widely been adopted to measure the classifiers performance and for comparison purposes as well. ROC plots a graph to demonstrate the performance of a classifier under the relationship of truly identified positive (sensitivity) and negative (specificity) at various threshold levels. And the quantitative comparison among ROC's of various classifiers to rank is being accessed by using ''area under curve (AUC)'' [3], [12], [30].

The last step is to develop actual classification model based on the selected classifier(s) and its possible deployment on web. The web application of the trained model is not obligatory, though, the research community must have some way to access the prediction model and the dataset, either in the form of downloadable tools, APIs, or other ways. Such provision will be significant to extend the experiments in the domain.

V. ISSUES AND CHALLENGES

There are notable challenges for the utilization of ML-based models for the therapeutic prediction of peptide sequences. ML classifier demand accurately validated and quantitative datasets. Such quantitative selection and dissemination of datasets is a challenging task due to the costly lab-based experiments and other associated complexities, leading to the inadequate exposure of the entire sequence set of the domain.

To develop an ML classifier with the best-fitted model, several repetitive experiments are obligatory. Amongst numerous available encoding schemes and classification models, the scholars of the field normally practice their personal expertise to select the encodings and classifiers. Due to all these, different trained models use to predict the same peptide sequences, may produce different outcomes. The difficult procedures to extract features may also affect the selection of unbiased and comprehensive features. Furthermore, the exceeding number of features from the size of the actual dataset may impact on the performance of a model [166].

The significant problem in implementing an ML model for some prospective therapeutic is the inadequate explanation of the data and repeatability of the experimental results obtained with ML. Various already implemented prediction models do not provide the datasets, ML models or model codes to the new researcher community, which is the main hindrance in designing new models for experimentation in the same field. So, it is very important to provide them the datasets and model codes to find out the new ways [167].

Data repetition is another challenge, which may head to the lower and biased performance of predictor due to overfitting of the model. Ordinarily, when a predictor is trained over the noisy/repetitive data, it gains the irrelevant information while learning, subsequently, a predictor becomes over-fitted and perform inadequately with the unseen dataset and providing unexpected results. Redundancy removal is its remedy by setting the sequence identity threshold using CD-HIT. Normally, the insufficient volume of the dataset does not support the use of threshold value at some recommended level, i.e. 30% [104]. The counterpart of overfitting is underfitting, occurs while a model is unable to learn well and produces biased results on an independent dataset. This may usually cause due to the limited size of the dataset. The underfitting and overfitting is a prominent challenge in the absence of testing or an independent dataset.

Typically, to access the performance of the created prediction model, the researcher utilizes own created independent dataset, which causes biases in the assessment. However, for neutral evaluation of the created model as well as for the comparison of performance with other models, the construction of a distinctive independent dataset is quite sensible, so suggested here. Another matter that intrudes biases in the results of a classifier is dataset disparity. In such a case the accuracy is not the only criterion; though various metrics should be used to evaluate a model to avoid the effects of dataset disparity, like, confusion metric, precision score, recall score, F-Score or ROC [168].

Many ML classifier models are available only for the prediction of the specific type of peptides. There is a dearth of the availability of common models to predict or classify multiple types of peptides as a whole. However, Chunrui *et al.* developed a method to predict ACP, allergenic, and virulence prospects from peptides [43], and Wei *et al.* established a

model for the prediction of eight types of peptides connected to therapy [3]. Conclusively, underlining the necessity for the development of general and progressive prediction models, proficient to predict several types of peptides and their functional activities.

VI. CONCLUSION

In this article, a systematic literature review has been conducted that provides a comprehensive discussion based on qualitatively selected research papers available in the field of computational prediction of therapeutic peptides. The study was carried out using a well-defined, systematic method for the selection of the fourty-one published articles. An analysis of the various ML classifiers, feature extraction schemes, and data-sources have been introduced in this study. Based on identified best practices from the selected articles, a framework has been presented as a guideline for the upcoming developments in the domain. Similarly, the feature encodings have been classified and presented in the taxonomic order. Moreover, issues, challenges and future prospects have been discussed, which may provide guidelines for the researchers of the domain. Furthermore, it is being advised to the researchers of the domain to only provide the most reliable prediction models for equitable future comparisons, instead to share all experimental models as practiced previously. However, various peptide predictors are available in the field, but the rapidly growing amount of datasets yet emphasizing the demand of new approaches and qualitative ML methods for the exploration of information beneficial in therapeutics and drug design.

REFERENCES

[1] K. Fosgerau and T. Hoffmann, "Peptide therapeutics: Current status and future directions," *Drug Discovery Today*, vol. 20, no. 1, pp. 122–128, Jan. 2015, doi: 10.1016/j.drudis.2014.10.003.

[2] C. de la Fuente-Núñez, O. N. Silva, T. K. Lu, and O. L. Franco, "Antimicrobial peptides: Role in human disease and potential as immunotherapies," *Pharmacol. Therapeutics*, vol. 178, pp. 132–140, Oct. 2017, doi: 10.1016/j.pharmthera.2017.04.002.

[3] L. Wei, C. Zhou, R. Su, and Q. Zou, "PEPred-suite: Improved and robust prediction of therapeutic peptides using adaptive feature representation learning," *Bioinformatics*, vol. 35, no. 21, pp. 4272–4280, Nov. 2019, doi: 10.1093/bioinformatics/btz246.

[4] C. Borghouts, C. Kunz, and B. Groner, "Current strategies for the development of peptide-based anti-cancer therapeutics," *J. Peptide Sci.*, vol. 11, no. 11, pp. 713–726, Nov. 2005, doi: 10.1002/psc.717.

[5] S. S. Usmani, "THPdb: Database of FDA-approved peptide and protein therapeutics," *PLoS ONE*, vol. 12, no. 7, 2017, Art. no. e0181748.

[6] C. Hu, X. Chen, Y. Huang, and Y. Chen, "Co-administration of iRGD with peptide HPRP-A1 to improve anticancer activity and membrane penetrability," *Sci. Rep.*, vol. 8, no. 1, pp. 1–14, Dec. 2018, doi: 10.1038/s41598-018-20715-4.

[7] A. C. Conibear, A. Schmid, M. Kamalov, C. F. W. Becker, and C. Bello, "Recent advances in peptide-based approaches for cancer treatment," *Current Medicinal Chem.*, vol. 27, no. 8, pp. 1174–1205, Mar. 2020, doi: 10.2174/0929867325666171123204851.

[8] L. Wei, P. Xing, R. Su, G. Shi, Z. S. Ma, and Q. Zou, "CPPred-RF: A sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency," *J. Proteome Res.*, vol. 16, no. 5, pp. 2044–2053, May 2017, doi: 10.1021/acs.jproteome.7b00019.

[9] J. Habault and J. L. Poyet, "Recent advances in cell penetrating peptide-based anticancer therapies," *Molecules*, vol. 24, no. 5, p. 927, 2019, doi: 10.3390/molecules24050927.

[10] L. Ferrero-Miliani, O. H. Nielsen, P. S. Andersen, and S. E. Girardin, "Chronic inflammation: Importance of NOD2 and NALP3 in interleukin-1β generation," *Clin. Exp. Immunol.*, vol. 147, no. 2, pp. 227–235, 2007, doi: 10.1111/j.1365-2249.2006.03261.x.

[11] R. Medzhitov, "Origin and physiological roles of inflammation," *Nature*, vol. 454, no. 7203, pp. 428–435, Jul. 2008, doi: 10.1038/nature07201.

[12] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, "AIPpred: Sequence-based prediction of anti-inflammatory peptides using random forest," *Frontiers Pharmacol.*, vol. 9, p. 276, Mar. 2018, doi: 10.3389/fphar.2018.00276.

[13] S. Gupta, A. K. Sharma, V. Shastri, M. K. Madhu, and V. K. Sharma, "Prediction of anti-inflammatory proteins/peptides: An insilico approach," *J. Transl. Med.*, vol. 15, no. 1, pp. 1–11, Jan. 2017, doi: 10.1186/s12967-016-1103-6.

[14] P. Zhou, C. Wang, Y. Ren, C. Yang, and F. Tian, "Computational peptidology: A new and promising approach to therapeutic peptide design," *Current Medicinal Chem.*, vol. 20, no. 15, pp. 1985–1996, Apr. 2013, doi: 10.2174/0929867311320150005.

[15] A. S. E. Ramaprasad, S. Singh, R. Gajendra, and P. S. S. Venkatesan, "AntiAngioPred: A server for prediction of anti-angiogenic peptides," *PLoS ONE*, vol. 10, no. 9, Sep. 2015, Art. no. e0136990, doi: 10.1371/journal.pone.0136990.

[16] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, Jun. 2018, doi: 10.1093/bioinformatics/bty451.

[17] C. Xu, L. Ge, Y. Zhang, M. Dehmer, and I. Gutman, "Prediction of therapeutic peptides by incorporating q-Wiener index into Chou's general PseAAC," *J. Biomed. Inform.*, vol. 75, pp. 63–69, Nov. 2017, doi: 10.1016/j.jbi.2017.09.011.

[18] W. Hao, C. Hu, Y. Huang, and Y. C. Id, "Coadministration of kla peptide with HPRP-A1 to enhance anticancer activity," *PLoS ONE* vol. 15, no. 11, 2019, Art. no. e0223738.

[19] Y. Xiao, "Peptide-based treatment: A promising cancer therapy," *J. Immunol. Res.*, vol. 2015, Oct. 2015, Art. no. 761820.

[20] P. Feng, Z. Wang, and X. Yu, "Predicting antimicrobial peptides by using increment of diversity with quadratic discriminant analysis method," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1309–1312, Jul. 2019, doi: 10.1109/TCBB.2017.2669302.

[21] Y.-F. Xiao, "Peptide-based treatment: A promising cancer therapy," *J. Immunol. Res.*, vol. 2015, Oct. 2015, Art. no. 761820, doi: 10.1155/2015/761820.

[22] S. K. Ghosh, T. S. McCormick, and A. Weinberg, "Human beta defensins and cancer: Contradictions and common ground," *Frontiers Oncol.*, vol. 9, pp. 1155–1163, May 2019, doi: 10.3389/fonc.2019.00341.

[23] M. D'Hondt, N. Bracke, L. Taevernier, B. Gevaert, F. Verbeke, E. Wynendaele, and B. De Spiegeleer, "Related impurities in peptide medicines," *J. Pharmaceutical Biomed. Anal.*, vol. 101, pp. 2–30, Dec. 2014, doi: 10.1016/j.jpba.2014.06.012.

[24] A. C.-L. Lee, J. L. Harris, K. K. Khanna, and J.-H. Hong, "A comprehensive review on current advances in peptide drug development and design," *Int. J. Mol. Sci.*, vol. 20, no. 10, p. 2383, May 2019, doi: 10.3390/ijms20102383.

[25] S. Vázquez-Prieto, E. Paniagua, H. Solana, F. M. Ubeira, and H. González-Díaz, "A study of the immune epitope database for some fungi species using network topological indices," *Mol. Diversity*, vol. 21, no. 3, pp. 713–718, Aug. 2017, doi: 10.1007/s11030-017-9749-4.

[26] B. Kitchenham, "Procedures for performing systematic reviews," Keele Univ. Nat., ICT Sydney, NSW, Australia, Tech. Rep., 2004, pp. 1–26.

[27] Z. Mushtaq, G. Rasool, and B. Shehzad, "Multilingual source code analysis: A systematic literature review," *IEEE Access*, vol. 5, pp. 11307–11336, 2017, doi: 10.1109/ACCESS.2017.2710421.

[28] T. Dybå and T. Dingsøyr, "Empirical studies of agile software development: A systematic review," *Inf. Softw. Technol.*, vol. 50, nos. 9–10, pp. 833–859, Aug. 2008, doi: 10.1016/j.infsof.2008.01.006.

[29] S. Spänig and D. Heider, "Encodings and models for antimicrobial peptide classification for multi-resistant pathogens," *BioData Mining*, vol. 12, no. 1, p. 7, Dec. 2019, doi: 10.1186/s13040-019-0196-x.

[30] M. N. Gabere and W. S. Noble, "Empirical comparison of Web-based antimicrobial peptide prediction tools," *Bioinformatics*, vol. 33, no. 13, pp. 1921–1929, Jul. 2017, doi: 10.1093/bioinformatics/btx081.

[31] M. Rowe, J. Frantz, and V. Bozalek, "The role of blended learning in the clinical education of healthcare students: A systematic review," *Med. Teacher*, vol. 34, no. 4, pp. e216–e221, Apr. 2012, doi: 10.3109/0142159X.2012.642831.

[32] M. Heyvaert, K. Hannes, B. Maes, and P. Onghena, "Critical appraisal of mixed methods studies," *J. Mixed Methods Res.*, vol. 7, no. 4, pp. 302–327, Oct. 2013, doi: 10.1177/1558689813479449.

[33] A. Fernandez, E. Insfran, and S. Abrahão, "Usability evaluation methods for the Web: A systematic mapping study," *Inf. Softw. Technol.*, vol. 53, no. 8, pp. 789–817, Aug. 2011, doi: 10.1016/j.infsof.2011.02.007.

[34] S. Ouhbi, A. Idri, J. L. Fernández-Alemán, and A. Toval, "Requirements engineering education: A systematic mapping study," *Requirements Eng.*, vol. 20, no. 2, pp. 119–138, Jun. 2015, doi: 10.1007/s00766-013-0192-5.

[35] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012, doi: 10.11613/bm.2012.031.

[36] N. Thakur, A. Qureshi, and M. Kumar, "AVPpred: Collection and prediction of highly effective antiviral peptides," *Nucleic Acids Res.*, vol. 40, no. W1, pp. W199–W204, Jul. 2012, doi: 10.1093/nar/gks450.

[37] D. Veltri, U. Kamath, and A. Shehu, "Deep learning improves antimicrobial peptide recognition," *Bioinformatics*, vol. 34, no. 16, pp. 2740–2747, Aug. 2018, doi: 10.1093/bioinformatics/bty179.

[38] S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, and S. Idicula-Thomas, "CAMP: A useful resource for research on antimicrobial peptides," *Nucleic Acids Res.*, vol. 38, no. 1, pp. D774–D780, Jan. 2010, doi: 10.1093/nar/gkp1021.

[39] P. Wang, L. Hu, G. Liu, N. Jiang, X. Chen, J. Xu, W. Zheng, L. Li, M. Tan, Z. Chen, H. Song, Y.-D. Cai, and K.-C. Chou, "Prediction of antimicrobial peptides based on sequence alignment and feature selection methods," *PLoS ONE*, vol. 6, no. 4, Apr. 2011, Art. no. e18476, doi: 10.1371/journal.pone.0018476.

[40] W. F. Porto, Á. S. Pires, and O. L. Franco, "CS-AMPPred: An updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides," *PLoS ONE*, vol. 7, no. 12, Dec. 2012, Art. no. e51444, doi: 10.1371/journal.pone.0051444.

[41] K. Y. Chang, T.-P. Lin, L.-Y. Shih, and C.-K. Wang, "Analysis and prediction of the critical regions of antimicrobial peptides based on conditional random fields," *PLoS ONE*, vol. 10, no. 3, Mar. 2015, Art. no. e0119490, doi: 10.1371/journal.pone.0119490.

[42] M. Torrent, D. Andreu, V. M. Nogués, and E. Boix, "Connecting peptide physicochemical and antimicrobial properties by a rational prediction model," *PLoS ONE*, vol. 6, no. 2, Feb. 2011, Art. no. e16968, doi: 10.1371/journal.pone.0016968.

[43] C. Xu, L. Ge, Y. Zhang, M. Dehmer, and I. Gutman, "Computational prediction of therapeutic peptides based on graph index," *J. Biomed. Informat.*, vol. 75, pp. 63–69, Nov. 2017, doi: 10.1016/j.jbi.2017.09.011.

[44] B. Manavalan, S. Basith, T. H. Shin, S. Choi, M. O. Kim, and G. Lee, "MLACP: Machine-learning-based prediction of anticancer peptides," *Oncotarget*, vol. 8, no. 44, pp. 77121–77136, Sep. 2017, doi: 10.18632/oncotarget.20365.

[45] V. Laengsri, C. Nantasenamat, N. Schaduangrat, P. Nuchnoi, V. Prachayasittikul, and W. Shoombuatong, "TargetAntiAngio: A sequence-based tool for the prediction and analysis of anti-angiogenic peptides," *Int. J. Mol. Sci.*, vol. 20, no. 12, p. 2950, Jun. 2019, doi: 10.3390/ijms20122950.

[46] C. Wu, R. Gao, Y. Zhang, and Y. De Marinis, "PTPD: Predicting therapeutic peptides by deep learning and word2vec," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–8, Dec. 2019, doi: 10.1186/s12859-019-3006-z.

[47] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, "ACPred: A computational tool for the prediction and analysis of anticancer peptides," *Molecules*, vol. 24, no. 10, p. 1973, May 2019, doi: 10.3390/molecules24101973.

[48] S. Gull, N. Shamim, and F. Minhas, "AMAP: Hierarchical multi-label prediction of biologically active and antimicrobial peptides," *Comput. Biol. Med.*, vol. 107, pp. 172–181, Apr. 2019, doi: 10.1016/j.compbiomed.2019.02.018.

[49] W. Shoombuatong, N. Schaduangrat, R. Pratiwi, and C. Nantasenamat, "THPep: A machine learning-based approach for predicting tumor homing peptides," *Comput. Biol. Chem.*, vol. 80, pp. 441–451, Jun. 2019, doi: 10.1016/j.compbiolchem.2019.05.008.

[50] F. C. Fernandes, D. J. Rigden, and O. L. Franco, "Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application," *Biopolymers*, vol. 98, no. 4, pp. 280–287, 2012, doi: 10.1002/bip.22066.

[51] K. Y. Chang and J.-R. Yang, "Analysis and prediction of highly effective antiviral peptides based on random forests," *PLoS ONE*, vol. 8, no. 8, Aug. 2013, Art. no. e70166, doi: 10.1371/journal.pone.0070166.

[52] L. Zhang, R. Yang, and C. Zhang, "Using a classifier fusion strategy to identify anti-angiogenic peptides," *Sci. Rep.*, vol. 8, no. 1, p. 14062, Dec. 2018, doi: 10.1038/s41598-018-32443-w.

[53] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, and G. P. S. Raghava, "In silico models for designing and discovering novel anticancer peptides," *Sci. Rep.*, vol. 3, no. 1, p. 2984, Dec. 2013, doi: 10.1038/srep02984.

[54] M. S. Khatun, M. M. Hasan, and H. Kurata, "PreAIP: Computational prediction of anti-inflammatory peptides by integrating multiple complementary features," *Frontiers Genet.*, vol. 10, p. 129, Mar. 2019, doi: 10.3389/fgene.2019.00129.

[55] S. Lata, N. K. Mishra, and G. P. Raghava, "AntiBP2: Improved version of antibacterial peptide prediction," *BMC Bioinf.*, vol. 11, no. 1, p. S19, 2010, doi: 10.1186/1471-2105-11-S1-S19.

[56] N. Schaduangrat, C. Nantasenamat, V. Prachayasittikul, and W. Shoombuatong, "Meta-iAVP: A sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation," *Int. J. Mol. Sci.*, vol. 20, no. 22, p. 5743, Nov. 2019, doi: 10.3390/ijms20225743.

[57] L. Ge, J. Liu, Y. Zhang, and M. Dehmer, "Identifying anticancer peptides by using a generalized chaos game representation," *J. Math. Biol.*, vol. 78, nos. 1–2, pp. 441–463, Jan. 2019, doi: 10.1007/s00285-018-1279-x.

[58] J. Zahiri, B. Khorsand, A. A. Yousefi, M. Kargar, R. Shirali Hossein Zade, and G. Mahdevar, "AntAngioCOOL: Computational detection of anti-angiogenic peptides," *J. Transl. Med.*, vol. 17, no. 1, pp. 1–6, Mar. 2019, doi: 10.1186/s12967-019-1813-7.

[59] J. L. Blanco, A. B. Porto-Pazos, A. Pazos, and C. Fernandez-Lozano, "Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection," *Sci. Rep.*, vol. 8, no. 1, p. 15688, Dec. 2018, doi: 10.1038/s41598-018-33911-z.

[60] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "IACP: A sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, pp. 16895–16909, Mar. 2016, doi: 10.18632/oncotarget.7815.

[61] V. Boopathi, S. Subramaniyam, A. Malik, G. Lee, B. Manavalan, and D.-C. Yang, "MACPpred: A support vector machine-based meta-predictor for identification of anticancer peptides," *Int. J. Mol. Sci.*, vol. 20, no. 8, p. 1964, Apr. 2019, doi: 10.3390/ijms20081964.

[62] X. Y. Ng, B. A. Rosdi, and S. Shahrudin, "Prediction of antimicrobial peptides based on sequence alignment and support vector machine-pairwise algorithm utilizing LZ-complexity," *BioMed Res. Int.*, vol. 2015, pp. 1–13, 2015, doi: 10.1155/2015/212715.

[63] S. Vijayakumar and L. Ptv, "ACPP: A Web server for prediction and design of anti-cancer peptides," *Int. J. Peptide Res. Therapeutics*, vol. 21, no. 1, pp. 99–106, Mar. 2015, doi: 10.1007/s10989-014-9435-7.

[64] H.-C. Yi, Z.-H. You, X. Zhou, L. Cheng, X. Li, T.-H. Jiang, and Z.-H. Chen, "ACP-DL: A deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation," *Mol. Therapy-Nucleic Acids*, vol. 17, pp. 1–9, Sep. 2019, doi: 10.1016/j.omtn.2019.04.025.

[65] A. Szaboova, O. Kuzelka, and F. Zelezny, "Prediction of antimicrobial activity of peptides using relational machine learning," in *Proc. IEEE Int. Conf. Bioinf. Biomed. Workshops*, Oct. 2012, pp. 575–580, doi: 10.1109/BIBMW.2012.6470203.

[66] S. Joseph, S. Karnik, P. Nilawe, V. K. Jayaraman, and S. Idicula-Thomas, "ClassAMP: A prediction tool for classification of antimicrobial peptides," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 5, pp. 1535–1538, Sep. 2012, doi: 10.1109/TCBB.2012.89.

[67] F. Khan, S. Akbar, A. Basit, I. Khan, and H. Akhlaq, "Identification of anticancer peptides using optimal feature space of Chou's split amino acid composition and support vector machine," in *Proc. 4th Int. Conf. Biomed. Bioinf. Eng. (ICBBE)*, 2017, pp. 91–96, doi: 10.1145/3168776.3168787.

[68] E. G. Randou, D. Veltri, and A. Shehu, "Binary response models for recognition of antimicrobial peptides," in *Proc. Int. Conf. Bioinf., Comput. Biol. Biomed. Informat. (BCB)*, 2007, pp. 76–85, doi: 10.1145/2506583.2506597.

[69] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test," *J. Theor. Biol.*, vol. 341, pp. 34–40, Jan. 2014, doi: 10.1016/j.jtbi.2013.08.037.

[70] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "IACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artif. Intell. Med.*, vol. 79, pp. 62–70, Jun. 2017, doi: 10.1016/j.artmed.2017.06.008.

[71] Z. Wang, "APD: The antimicrobial peptide database," *Nucleic Acids Res.*, vol. 32, pp. D590–D592, Jan. 2004, doi: 10.1093/nar/gkh025.

[72] G. Wang, X. Li, and Z. Wang, "APD2: The updated antimicrobial peptide database and its application in peptide design," *Nucleic Acids Res.*, vol. 37, no. 1, pp. D933–D937, Jan. 2009, doi: 10.1093/nar/gkn823.

[73] G. Wang, X. Li, and Z. Wang, "APD3: The antimicrobial peptide database as a tool for research and education," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1087–D1093, Jan. 2016, doi: 10.1093/nar/gkv1278.

[74] F. H. Waghu, L. Gopi, R. S. Barai, P. Ramteke, B. Nizami, and S. Idicula-Thomas, "CAMP: Collection of sequences and structures of antimicrobial peptides," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D1154–D1158, Jan. 2014, doi: 10.1093/nar/gkt1157.

[75] M. Novković, J. Simunić, V. Bojović, A. Tossi, and D. Juretić, "DADP: The database of anuran defense peptides," *Bioinformatics*, vol. 28, no. 10, pp. 1406–1407, May 2012, doi: 10.1093/bioinformatics/bts141.

[76] B. Boeckmann, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 365–370, Jan. 2003, doi: 10.1093/nar/gkg095.

[77] A. Bairoch, "The universal protein resource (UniProt)," *Nucleic Acids Res.*, vol. 33, pp. D154–D159, Dec. 2004, doi: 10.1093/nar/gki070.

[78] A. Bateman, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017, doi: 10.1093/nar/gkw1099.

[79] H. M. Berman, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000, doi: 10.1093/nar/28.1.235.

[80] NCBI Resource Coordinators, "Database resources of the national center for biotechnology information," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D12–D17, Jan. 2017, doi: 10.1093/nar/gkw1071.

[81] R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters, "The immune epitope database (IEDB) 3.0," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D405–D412, Jan. 2015, doi: 10.1093/nar/gku938.

[82] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, and G. P. S. Raghava, "CancerPPD: A database of anticancer peptides and proteins," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D837–D843, Jan. 2015, doi: 10.1093/nar/gku892.

[83] Y. D. Khan, N. Rasool, W. Hussain, S. A. Khan, and K.-C. Chou, "IPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC," *Anal. Biochem.*, vol. 550, pp. 109–116, Jun. 2018, doi: 10.1016/j.ab.2018.04.021.

[84] M. Bhasin and G. P. S. Raghava, "Classification of nuclear receptors based on amino acid composition and dipeptide composition," *J. Biol. Chem.*, vol. 279, no. 22, pp. 23262–23266, May 2004, doi: 10.1074/jbc.M401932200.

[85] H. Lin, W.-X. Liu, J. He, X.-H. Liu, H. Ding, and W. Chen, "Predicting cancerlectins by the optimal g-gap dipeptides," *Sci. Rep.*, vol. 5, no. 1, p. 16964, Dec. 2015, doi: 10.1038/srep16964.

[86] L. Wei, J. Tang, and Q. Zou, "SkipCPP-pred: An improved and promising sequence-based predictor for predicting cell-penetrating peptides," *BMC Genomics*, vol. 18, no. S7, pp. 1–11, Oct. 2017, doi: 10.1186/s12864-017-4128-1.

[87] M. M. Hasan, D. Guo, and H. Kurata, "Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information," *Mol. Biosyst.*, vol. 13, no. 12, pp. 2545–2550, 2017, doi: 10.1039/c7mb00491e.

[88] M. M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song, and Z. Zhang, "Computational identification of protein pupylation sites by using profile-based composition of k-Spaced amino acid pairs," *PLoS ONE*, vol. 10, no. 6, Jun. 2015, Art. no. e0129635, doi: 10.1371/journal.pone.0129635.

[89] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "IFeature: A Python package and Web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018, doi: 10.1093/bioinformatics/bty140.

[90] K.-C. Chou, "Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology," *Current Proteomics*, vol. 6, no. 4, pp. 262–274, Dec. 2009, doi: 10.2174/157016409789973707.

[91] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins: Struct., Function, Genet.*, vol. 43, no. 3, pp. 246–255, May 2001, doi: 10.1002/prot.1035.

[92] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011, doi: 10.1016/j.jtbi.2010.12.024.

[93] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "Propy: A tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, Apr. 2013, doi: 10.1093/bioinformatics/btt072.

[94] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Anal. Biochem.*, vol. 425, no. 2, pp. 117–119, Jun. 2012, doi: 10.1016/j.ab.2012.03.015.

[95] P. Du, S. Gu, and Y. Jiao, "PseAAC-general: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *Int. J. Mol. Sci.*, vol. 15, no. 3, pp. 3495–3506, Feb. 2014, doi: 10.3390/ijms15033495.

[96] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-one: A Web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W65–W71, Jul. 2015, doi: 10.1093/nar/gkv458.

[97] B. Liu, H. Wu, and K.-C. Chou, "Pse-in-One 2.0: An improved package of Web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Natural Sci.*, vol. 9, no. 4, pp. 67–91, 2017, doi: 10.4236/ns.2017.94007.

[98] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005, doi: 10.1093/bioinformatics/bth466.

[99] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen, "PROFEAT: A Web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence," *Nucleic Acids Res.*, vol. 34, no. Web Server, pp. W32–W37, Jul. 2006, doi: 10.1093/nar/gkl305.

[100] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S. H. Kim, "Recognition of a protein fold in the context of the SCOP classification," *Proteins: Struct., Function, Bioinf.*, vol. 35, no. 4, pp. 4001–407, 1999, doi: 10.1002/(SICI)1097-0134(19990601)35:4<401::AID-PROT3>3.0.CO;2-K.

[101] G. Govindan and A. S. Nair, "Composition, Transition and Distribution (CTD)—A dynamic feature for predictions based on hierarchical structure of cellular sorting," in *Proc. Annu. IEEE India Conf.*, Dec. 2011, pp. 1–6, doi: 10.1109/INDCON.2011.6139332.

[102] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 19, pp. 8700–8704, 1995, doi: 10.1073/pnas.92.19.8700.

[103] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino acid index database, progress report 2008," *Nucleic Acids Res.*, vol. 36, pp. D202–D205, Dec. 2007, doi: 10.1093/nar/gkm998.

[104] S. Gupta, M. K. Madhu, A. K. Sharma, and V. K. Sharma, "ProInflam: A webserver for the prediction of proinflammatory antigenicity of peptides and proteins," *J. Transl. Med.*, vol. 14, no. 1, p. 178, Dec. 2016, doi: 10.1186/s12967-016-0928-3.

[105] N. Petrovsky and V. Brusic, "Computational immunology: The coming of age," *Immunol. Cell Biol.*, vol. 80, no. 3, pp. 248–254, Jun. 2002, doi: 10.1046/j.1440-1711.2002.01093.x.

[106] D. Sirskyj, F. D. Mitoma, A. Golshani, A. Kumar, and A. Azizi, "Innovative bioinformatic approaches for developing peptide-based vaccines against hypervariable viruses," *Immunol. Cell Biol.*, vol. 89, no. 1, pp. 81–89, Jan. 2011, doi: 10.1038/icb.2010.65.

[107] C. Vens, M.-N. Rosso, and E. G. J. Danchin, "Identifying discriminative classification-based motifs in biological sequences," *Bioinformatics*, vol. 27, no. 9, pp. 1231–1238, May 2011, doi: 10.1093/bioinformatics/btr110.

[108] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: Tools for motif discovery and searching," *Nucleic Acids Res.*, vol. 37, pp. W202–W208, Jul. 2009, doi: 10.1093/nar/gkp335.

[109] M. Gollery, *Bioinformatics: Sequence and Genome Analysis*, D. W. Mount, Ed., 2nd ed. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 2004, p. 692, doi: 10.1373/clinchem.2005.053850.

[110] A. Zielezinski, S. Vinga, J. Almeida, and W. M. Karlowski, "Alignment-free sequence comparison: Benefits, applications, and tools," *Genome Biol.*, vol. 18, no. 1, p. 186, Dec. 2017, doi: 10.1186/s13059-017-1319-7.
[111] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, no. 3, pp. 403-410, 1990, doi: 10.1016/S0022-2836(05)80360-2.
[112] S. F. Altschul, "Evaluating the statistical significance of multiple distinct local alignments," in *Theoretical and Computational Methods in Genome Research*. Boston, MA, USA: Springer, 1997.
[113] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics.*, vol. 14, no. 9, pp. 755–763, 1998, doi: 10.1093/bioinformatics/14.9.755.
[114] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison.," *Proc. Nat. Acad. Sci. USA*, vol. 85, no. 8, pp. 2444–2448, Apr. 1988, doi: 10.1073/pnas.85.8.2444.
[115] S. Altschul, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997, doi: 10.1093/nar/25.17.3389.
[116] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, no. 1, pp. 195–197, 1981, doi: 10.1016/0022-2836(81)90087-5.
[117] J. Svenson, R. Karstad, G. E. Flaten, B.-O. Brandsdal, M. Brandl, and J. S. Svendsen, "Altered activity and physicochemical properties of short cationic antimicrobial peptides by incorporation of arginine analogues," *Mol. Pharmaceutics*, vol. 6, no. 3, pp. 996–1005, Jun. 2009, doi: 10.1021/mp900057k.
[118] H. J. Jeffrey, "Chaos game representation of gene structure," *Nucleic Acids Res.*, vol. 18, no. 8, pp. 2163–2170, 1990, doi: 10.1093/nar/18.8.2163.
[119] S. Basu, A. Pan, C. Dutta, and J. Das, "Chaos game representation of proteins," *J. Mol. Graph. Model.*, vol. 15, no. 5, pp. 279–289, 1997, doi: 10.1016/S1093-3263(97)00106-X.
[120] H. F. Löchel, D. Eger, T. Sperlea, and D. Heider, "Deep learning on chaos game representation for proteins," *Bioinformatics*, vol. 36, no. 1, pp. 272–279, Jan. 2020, doi: 10.1093/bioinformatics/btz493.
[121] A. S. Nair, V. V. Nair, K. S. Arun, K. Kant, and A. Dey, "Bio-sequence signatures using chaos game representation," in *Bioinformatics: Applications in Life and Environmental Sciences*. Dordrecht, The Netherlands: Springer, 2007.
[122] J. Pal, S. Ghosh, B. Maji, and D. K. Bhattacharya, "Use of FFT in protein sequence comparison under their binary representations," *Comput. Mol. Biosci.*, vol. 6, no. 2, pp. 33–40, 2016, doi: 10.4236/cmb.2016.62003.
[123] T. Hoang, C. Yin, H. Zheng, C. Yu, R. Lucy He, and S. S.-T. Yau, "A new method to cluster DNA sequences using Fourier power spectrum," *J. Theor. Biol.*, vol. 372, pp. 135–145, May 2015, doi: 10.1016/j.jtbi.2015.02.026.
[124] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Comput. Biol. Chem.*, vol. 34, nos. 5–6, pp. 320–327, Dec. 2010, doi: 10.1016/j.compbiolchem.2010.09.002.
[125] R. A. Edwards, R. Olson, T. Disz, G. D. Pusch, V. Vonstein, R. Stevens, and R. Overbeek, "Real time metagenomics: Using k-mers to annotate metagenomes," *Bioinformatics*, vol. 28, no. 24, pp. 3316–3317, Dec. 2012, doi: 10.1093/bioinformatics/bts599.
[126] S. C. Manekar and S. R. Sathe, "A benchmark study of k-mer counting methods for high-throughput sequencing," *GigaScience*, vol. 7, no. 12, p. giy125, Oct. 2018, doi: 10.1093/gigascience/giy125.
[127] M. K. Mejía-Guerra and E. S. Buckler, "A k-mer grammar analysis to uncover maize regulatory architecture," *BMC Plant Biol.*, vol. 19, no. 1, pp. 1–17, Dec. 2019, doi: 10.1186/s12870-019-1693-2.
[128] M. Kokot, M. Długosz, and S. Deorowicz, "KMC 3: Counting and manipulating k-mer statistics," *Bioinformatics*, vol. 33, no. 17, pp. 2759–2761, Sep. 2017, doi: 10.1093/bioinformatics/btx304.
[129] P. Audano and F. Vannberg, "KAnalyze: A fast versatile pipelined K-mer toolkit," *Bioinformatics*, vol. 30, no. 14, pp. 2070–2072, Jul. 2014, doi: 10.1093/bioinformatics/btu152.
[130] Y. Yang, "Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks," in *Methods in Molecular Biology*. New York, NY, USA: Humana Press, 2017.
[131] K. Carr, E. Murray, E. Armah, R. L. He, and S. S.-T. Yau, "A rapid method for characterization of protein relatedness using feature vectors," *PLoS ONE*, vol. 5, no. 3, p. e9550, Mar. 2010, doi: 10.1371/journal.pone.0009550.
[132] W. Shoombuatong, N. Schaduangrat, and C. Nantasenamat, "Unraveling the bioactivity of anticancer peptides as deduced from machine learning," *EXCLI J.*, vol. 17, pp. 734–752, 2018, doi: 10.17179/excli2018-1447.
[133] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "ILearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, vol. 21, no. 3, pp. 1047–1057, May 2020, doi: 10.1093/bib/bbz041.
[134] A. Pande, "Computing wide range of protein/peptide features from their sequence and structure," 2019. [Online]. Available: https://www.biorxiv.org/content/10.1101/599126v1, doi: 10.1101/599126.
[135] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, "Protr/ProtrWeb: R package and Web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, Jan. 2015, doi: 10.1093/bioinformatics/btv042.
[136] J. Dong, Z.-J. Yao, L. Zhang, F. Luo, Q. Lin, A.-P. Lu, A. F. Chen, and D.-S. Cao, "PyBioMed: A Python library for various molecular representations of chemicals, proteins and DNAs and their interactions," *J. Cheminform.*, vol. 10, no. 1, Dec. 2018, doi: 10.1186/s13321-018-0270-2.
[137] C. Wu, M. Berry, S. Shivakumar, and J. McLarty, "Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition," *Mach. Learn.*, vol. 21, nos. 1–2, pp. 177–193, 1995, doi: 10.1007/bf00993384.
[138] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[139] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE Access*, vol. 7, pp. 53040–53065, 2019, doi: 10.1109/ACCESS.2019.2912200.
[140] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer*, vol. 29, no. 3, pp. 31–44, Mar. 1996, doi: 10.1109/2.485891.
[141] N. Zhang, W. Wu, and G. Zheng, "Convergence of gradient method with momentum for two-layer feedforward neural networks," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 522–525, Mar. 2006, doi: 10.1109/TNN.2005.863460.
[142] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Nov. 1991, doi: 10.1109/72.97934.
[143] B. Mohebali, A. Tahmassebi, A. Meyer-Baese, and A. H. Gandomi, "Probabilistic neural networks: A brief overview of theory, implementation, and application," in *Handbook of Probabilistic Models*. Oxford, U.K.: Butterworth-Heinemann, 2019.
[144] N. Sharma and H. Om, "Usage of probabilistic and general regression neural network for early detection and prevention of oral cancer," *Sci. World J.*, vol. 2015, pp. 1–11, 2015, doi: 10.1155/2015/234191.
[145] D. F. Specht, "Probabilistic neural networks," *Neural Netw.*, vol. 3, no. 1, pp. 109–118, 1990, doi: 10.1016/0893-6080(90)90049-Q.
[146] J. Yan, P. Bhadra, A. Li, P. Sethiya, L. Qin, H. K. Tai, K. H. Wong, and S. W. I. Siu, "Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning," *Mol. Therapy-Nucleic Acids*, vol. 20, pp. 882–894, Jun. 2020, doi: 10.1016/j.omtn.2020.05.006.
[147] S.-J. Heo, Z. Chunwei, and E. Yu, "Response simulation, data cleansing and restoration of dynamic and static measurements based on deep learning algorithms," *Int. J. Concrete Struct. Mater.*, vol. 12, no. 1, p. 82, Dec. 2018, doi: 10.1186/s40069-018-0316-x.
[148] K. Łapa, K. Cpałka, and L. Rutkowski, "New aspects of interpretability of fuzzy systems for nonlinear modeling," in *Studies in Computational Intelligence*. Cham, Switzerland: Springer, 2018.
[149] H. Li, J. Wang, H. Du, and H. R. Karimi, "Adaptive sliding mode control for Takagi–Sugeno fuzzy systems and its applications," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 2, pp. 531–542, Apr. 2018, doi: 10.1109/TFUZZ.2017.2686357.
[150] M. Hesami, R. Naderi, M. Tohidfar, and M. Yoosefzadeh-Najafabadi, "Application of adaptive neuro-fuzzy inference System-Non-dominated sorting genetic algorithm-II (ANFIS-NSGAII) for modeling and optimizing somatic embryogenesis of chrysanthemum," *Frontiers Plant Sci.*, vol. 10, p. 869, Jul. 2019, doi: 10.3389/fpls.2019.00869.
[151] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 1–10, doi: 10.1038/nprot.2006.61.
[152] J. Tolles and W. J. Meurer, "Logistic regression: Relating patient characteristics to outcomes," *JAMA*, vol. 316, no. 5, p. 533, Aug. 2016, doi: 10.1001/jama.2016.7653.
[153] Lu, "Increment of diversity with quadratic discriminant analysis-an efficient tool for sequence pattern recognition in bioinformatics," *Open Access Bioinf.*, vol. 2, p. 89, Aug. 2010, doi: 10.2147/oab.s10782.

[154] I. Rish, "An empirical study of the naive Bayes classifier," in *Proc. Workshop Empirical Methods Artif. Intell. (IJCAI)*, 2001, pp. 1–6, doi: 10.1039/b104835j.
[155] A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression Trees.," *Biometrics*, vol. 40, no. 3, p. 874, Sep. 1984, doi: 10.2307/2530946.
[156] S. Ali and K. A. Smith, "On learning algorithm selection for classification," *Appl. Soft Comput.*, vol. 6, no. 2, pp. 119–138, Jan. 2006, doi: 10.1016/j.asoc.2004.12.002.
[157] W. Loh, "Classification and regression trees," *WIREs Data Mining Knowl. Discovery*, vol. 1, no. 1, pp. 14–23, Jan. 2011, doi: 10.1002/widm.8.
[158] C. Cortes, M. Mohri, and U. Syed, "Deep boosting," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 1–9.
[159] Y. Kliger, E. Gofer, A. Wool, A. Toporik, A. Apatoff, and M. Olshansky, "Predicting proteolytic sites in extracellular proteins: Only halfway there," *Bioinformatics*, vol. 24, no. 8, pp. 1049–1055, Apr. 2008, doi: 10.1093/bioinformatics/btn084.
[160] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.
[161] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT suite: A Web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010, doi: 10.1093/bioinformatics/btq003.
[162] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, pp. 1–39, Nov. 2010, doi: 10.1007/s10462-009-9124-7.
[163] R. Xu, J. Zhou, B. Liu, L. Yao, Y. He, Q. Zou, and X. Wang, "EnDNA-prot: Identification of DNA-binding proteins by applying ensemble learning," *BioMed Res. Int.*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/294279.
[164] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 10, pp. 993–1001, Oct. 1990, doi: 10.1109/34.58871.
[165] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, Jan. 2011.
[166] F. Li, Y. Wang, C. Li, T. T. Marquez-Lago, A. Leier, N. D. Rawlings, G. Haffari, J. Revote, T. Akutsu, K.-C. Chou, A. W. Purcell, R. N. Pike, G. I. Webb, A. Ian Smith, T. Lithgow, R. J. Daly, J. C. Whisstock, and J. Song, "Twenty years of bioinformatics research for protease-specific substrate and cleavage site prediction: A comprehensive revisit and benchmarking of existing methods," *Briefings Bioinf.*, vol. 20, no. 6, pp. 2150–2166, Nov. 2019, doi: 10.1093/bib/bby077.
[167] B. Peters, S. E. Brenner, E. Wang, D. Slonim, and M. G. Kann, "Putting benchmarks in their rightful place: The heart of computational biology," *PLOS Comput. Biol.*, vol. 14, no. 11, Nov. 2018, Art. no. e1006494, doi: 10.1371/journal.pcbi.1006494.
[168] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*. Boston, MA, USA: Springer, 2009.

MUHAMMAD ATTIQUE was born in Bahawalpur, Punjab, Pakistan. He received the M.Sc. and M.S. degrees in computer science from the Islamia University of Bahawalpur. He has extensive experience to work in multinational organizations and serve at various designative levels. He is currently an Assistant Professor with the Department of Information Technology, University of Gujrat, Pakistan. He is the author of three journal articles. His research interests include digital image processing, pattern recognition, machine learning, and bioinformatics.

MUHAMMAD SHOAIB FAROOQ (Member, IEEE) has more than 25 years of teaching experience in the field of computer science. He is currently an Associate Professor of computer science with the University of Management and Technology, Lahore. He is also an Affiliate Member with George Mason University, USA. He has published many peer-reviewed international journal and conference papers. His research interests include the theory of programming languages, big data, the IoT, the Internet of Vehicles, machine learning, distributed systems, and education.

ADEL KHELIFI received the Ph.D. degree from the Engineering School of High Technology, Canada, in 2005. He was the Dean of the Computer Information Technology, American University in the Emirates, Dubai, United Arab Emirates. He has held impressive past careers. He was a Lecturer with the Engineering School of Technology, Canada. He was also a United Nations MSF, Canada, the Ministry of Relations with the Citizen and Immigration, Canada, and the Ministry of Finances, Tunisia. He is currently a Faculty Member with Abu Dhabi University. He holds a high-level of knowledge and expertise. He is involved in prompting the open source software paradigm in the region. As a Canadian ISO Member in software engineering, he is contributing in developing software measurement standards. His research interest includes archeology.

ADNAN ABID (Member, IEEE) was born in Gujranwala, Pakistan, in 1979. He received the B.S. degree from the National University of Computer and Emerging Science, Pakistan, in 2001, the M.S. degree in information technology from the National University of Science and Technology, Pakistan, in 2007, and the Ph.D. degree in computer science from the Politecnico Di Milano, Italy, in 2012. He spent a period of one year with EPFL, Switzerland, to complete the M.S. thesis. He is currently an Associate Professor with the Department of Computer Science, University of Management and Technology, Pakistan. He has almost 40 publications in different international journals and conferences. His research interests include computer science education, information retrieval, and data management. He served as a Reviewer for many international conferences.