UNIVERSITY OF PLOVDIV "PAISII HILENDARSKI"

FACULTY OF MATHEMATICS AND INFORMATICS

DEPARTMENT OF COMPUTER INFORMATICS

Zhelyazko Petrov Terziyski

Using techniques from artificial intelligence to analyze and predict the properties of peptides

ABSTRACT

of a dissertation

for awarding the educational and scientific degree "Doctor"

Field of higher education: 4. Natural sciences, mathematics and informatics

Professional direction: 4.6. Informatics and Computer Science

Doctoral program: "Informatics"

Scientific supervisor**: Assoc. Prof. Stanka Hadzhikoleva, PhD**

Plovdiv , 2024

The dissertation work was discussed and directed for defense before a scientific jury, at a meeting of the "Computer Informatics" department at the Faculty of Mathematics and Informatics of the University of Plovdiv "Paisii Hilendarski", on 16.01.2024.

The dissertation "Using techniques from artificial intelligence to analyze and predict the properties of peptides" contains 189 pages. The list of used literature includes 150 sources. The list of author publications on the topic consists of 5 titles.

The defense of the dissertation work will take place on 11.03.2024 at ..... in the meeting room in the New Building of PU "Paisiy Hilendarski".

Defense materials are available in the Dean's Office of the Faculty of Mathematics and Informatics, New Building of PU "Paisiy Hilendarski", every working day from 8:30 am to 5:00 pm

Author: Zhelyazko Petrov Terziyski

Title: Using techniques from artificial intelligence to analyze and predict the properties of peptides

Plovdiv, 2024

Table of contents

Abbreviations

AAS	-	Amino Acid Composition
BAV	-	Biologically active substances
DB	-	databases
AI	-	Artificial Intelligence
NN	-	Neural Network
AUC	-	Area under ROC curve
FFNN	-	Feed Forward Neural Network
pI	-	Isoelectric point
RF	-	Random forest
SVM	-	Support vector machines
QSAR	-	Quantitative structure-property relationships

Introduction

Peptides are chemical substances of natural or synthetic origin that consist of amino acids. Their chain includes from 2 to about 50 amino acids. In the presence of more than 50 amino acids, they are transformed into proteins.

Peptides have a wide range of biological properties affecting the human organism. They are used both in the treatment of diseases and as food supplements or cosmetics. The first peptide drug was insulin, discovered as early as 1921.

Of interest to medicine are peptides that have a positive effect on human health. Their influence is characterized by a specific focus. This is the so-called biological activity, according to which peptides are anticancer, antimicrobial, blood pressure normalizing, antioxidant, antidiabetic, etc.

The properties of peptides depend directly on their structure. The sequence-based method is one of the most widely used methods for studying quantitative structure-activity relationships (QSAR - Quantitative Structure Property Relationships). It can be used for in silico selection of the best candidate peptides before their actual synthesis. Because of this, this thesis is aimed at predicting the biological activity of peptides by the QSAR method, using artificial intelligence (AI) methods.

The main goal of the dissertation research **is to design, develop, and validate a software system for predicting the biological properties of peptides using various artificial intelligence methods.**

The fulfillment of the set goal implies solving the following tasks:

Task 1. Study of the essence of peptides, current research, and scientific achievements for the use of artificial intelligence methods for predicting their biological activity;

Task 2. Study of publicly available databases and tools for coding peptides and creation of a database of known peptides and their biological activity;

Task 3. Create a conceptual model of a software application for the analysis and prediction of peptide properties by various AI methods;

Task 4. Development of a software prototype, including a database, a module for extracting physicochemical characteristics of peptides, and a module for predicting the biological activity of peptides, using different methods of artificial intelligence;

Task 5. Testing of the software prototype and analysis of the results of the conducted experiments.

The structure of the dissertation follows the conducted research, development and experiments and includes an Introduction, 4 chapters, a Conclusion, and three appendices containing a description of the designed database of peptides, statistical information on the characteristics of the peptides, and output results of experiments conducted to calculate the physicochemical characteristics of a peptide.

In **Chapter 1**, a survey of state-of-the-art approaches and research is made to predict peptide properties using various AI methods. The classic algorithm for in silico prediction of their biological activity is presented.

Chapter 2 presents a conceptual model of a software application for analyzing the physicochemical properties and predicting the biological activity of peptides.

A Pep Lab software application is presented in **Chapter 3**. It implements the functionalities and processes described in the model.

Chapter **4** presents the results of the Pep Lab experiments. These include generating various datasets and predicting the biological activity of peptides through several AI models.

In the conclusion, the results of the implementation of the set tasks are summarized. Scientific, scientific-applied and applied contributions resulting from the dissertation research are described, as well as directions for future development.

The list of used literature includes **150** titles, of which **19** are Internet sources.

The results of the research are presented in **5 publications**. **Two of them are indexed in the international databases Scopus and Web of Science, and 1 is published in a publication with an impact factor**.

Participation in **6 research projects** related to artificial intelligence or the study of peptides provided additional guidelines for the successful implementation of the dissertation research.

Acknowledgments

I express my gratitude to my supervisor, Assoc. Dr. Stanka Hadzhikoleva, as well as Prof. Dr. Emil Hadzhikolev for the help, support, ideas, guidelines, and recommendations during the dissertation research, as well as the opportunity to participate in research projects. I also send my sincere thanks to Assoc. Dr. Margarita Terziyska and Ch. assistant professor Ivelina Deseva from UHT-Plovdiv for the help, guidance, and recommendations in the field of AI and bioinformatics.

CHARTER 1. Peptides. Using AI to predict the properties of peptides

Peptides are substances whose molecules are made up of a chain of 20 basic amino acids connected by peptide bonds. Most often, the peptide sequence consists of 2 to 50 amino acids and therefore their molecular mass is small, usually below 10kDa. The beginning of the development of peptide drugs is considered to be in 1921, when insulin, the first peptide drug, was discovered.

Peptides play an extremely diverse role in the human body. They take an active part in the growth of the body, support the work of the immune system, are used as anti-infective agents, etc. [1]. In other words, their effect on human health is specific [2].

Peptide-based therapy is experiencing a real boom in global markets, with sales of over $43 billion in 2022 alone. Revenues from peptide-based therapeutics are expected to reach $77 billion by 2032 [3]. This huge market demand requires more flexibility than traditional *in vivo* and *in vitro* experimental approaches for peptide discovery. These processes are time-consuming, expensive, and labor-intensive [4].

Overcoming these limitations requires the application of novel and efficient computational approaches to the discovery and investigation of candidate peptides. In practice, so-called *in silico*

methods are increasingly used, which serve as fast and inexpensive tools for their identification in the vast space of combinatorial sequence. They use knowledge from various fields such as mathematics, informatics, biology, biotechnology, and analytical chemistry to discover potential candidates before their actual synthesis, thus accelerating the discovery of new useful peptides. An extremely promising model for *the in silico* approach is sequence-based for finding a quantitative relationship between the structure and activity of the peptide (QSAR model).

Using AI in predicting the properties of peptides

The realization of the QSAR model by means of various MO techniques allows to successfully identify the structure of new peptides [5]. Different methods achieve different accuracy of predictions. Therefore, the question of the need to systematically evaluate and improve these algorithms in terms of their methodology and predictive ability is relevant.

In the modeling of peptide sequences by QSAR, the predictors consist of physicochemical characteristics or theoretical molecular descriptors of peptides, and the categorical value is a specific biological activity [6]. Models first generalize a putative relationship between chemical structures with known biological activity in an available data set. The QSAR model then predicts the activity of new peptides.

The conducted research of the scientific literature shows that numerous algorithms have been developed for predicting the biological activity of peptides, based on various techniques from MO - artificial neural networks (NM) [7], random forest (Random Forest - RF) [8], the method of support vectors (Support Vector Machines - SVM) [9], etc.

In conclusion, MO is considered the best *in silico* tool in bioinformatics for solving problems involving large datasets and multiple variables. This is due to their ability to analyze and learn from available data, and once trained, make relatively accurate predictions on new data sets [11].

Process for predicting the biological activity of peptides

The classical methodology for developing models for predicting the biological activity of peptides includes 4 stages [13]. Graphically, it is presented in fig. 1. The stages are:

1. **Data collection and processing;**
2. **Feature selection and coding;**
3. **Creation of MO model;**
4. **Analysis and evaluation of the model.**

It should be noted that such a general scheme has also been widely applied in the prediction of functions in DNA and protein structures [14].

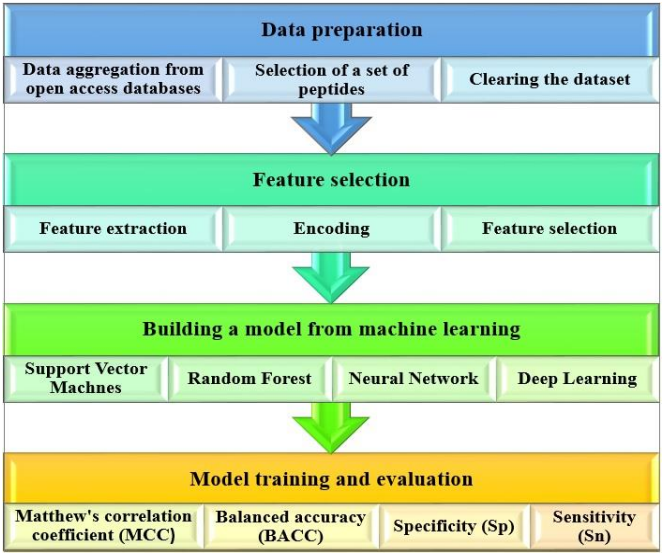


Figure 1. Methodology for predicting the biological activity of peptides

Therefore, the formation of the negative data set is done in one of the following ways:

a) random peptides are selected, with a function other than the desired one;

b) toxic peptides are selected;

(c) random peptides are generated.

These approaches are based on the hypothesis that the probability that the samples thus generated are not negative is minimal.

Data preprocessing is essential for the effective use of MO techniques. It is important to examine the input features and optimize them to reduce their number. For this purpose, from a given input vector with a certain number of features, only a part of them is selected. This subset includes only the features that most influence the prediction process. Different feature selection approaches are used to obtain the significant features. **They have some important advantages: trivial information is removed; models are simplified and become easier to understand; the computational burden of the model is significantly reduced; the compatibility with the learning model is improved**.

The basic premise of using a feature selection technique is that the data contains some features that are either redundant or irrelevant, and thus can be removed without causing much loss of information. The extraction process creates new features from the original input data vector, while feature selection returns a subset of that vector [15]. The feature selection technique is usually used in areas where there are relatively few samples. This is also the case in the study of peptides.

The selection of features when evaluating peptide biological activity makes sense only in the context of a specific prediction model.

Table 1. Cloud services and software applications for computing feature descriptors

Method	Internet address
Propy	code.google.com/p/propy/
Pse-in-One	bliulab.net/Pse-in-One/
Pse-in-One 2.0	bliulab.net/Pse-in-One2.0/
iFeature	ifeature.erc.monash.edu
Pfeature	webs.iiitd.edu.in/raghava/pfeature/
iLearn	ilearn.erc.monash.edu
Seq2Feature	www.iitm.ac.in/bioinfo/SBFE/
PyBioMed	github.com/gadsbyfly/PyBioMed
PyFeat	github.com/mrzResearchArena/PyFeat
VisFeature	github.com/wangjun1996/VisFeature

Feature extraction and encoding is defined in [16]. Software applications for calculating feature descriptors are presented in Table 1.

Tutoring with an ML teacher has been successfully used to solve various problems in bioinformatics [10]. The model is trained with known input and output data, and then it can predict the outcome when given unknown data [11]. A model that generates discrete classes is called a classification algorithm. This type of algorithm evaluates the prediction for a class that is 0 (negative) or 1 (positive).

AI models used to predict peptide properties most often include - artificial neural networks, SVM, RF, etc. [12]. Instead of randomly choosing an ML classifier, it is good to examine one or more classifiers and choose the most suitable one.

Model overtraining occurs when the model not only learns from the input signals but also integrates irrelevant feature information from the training data. To limit the overtraining of the model, cross-validation (CV) is used.

To evaluate the effectiveness of the prediction of the various methods in ML, the following evaluation indicators are most often used - classification accuracy (CA), specificity (Sp), sensitivity (Sn), precision (Pr), average harmonic value of sensitivity and precision (F1) and Matthew's Correlation Coefficient (MCC). They are calculated with the formulas:

$$CA = \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

$$Sp = \frac{TN}{TN+FP} \quad (2)$$

$$Sn = \frac{TP}{TP+FN} \quad (3)$$

$$Pr = \frac{TP}{TP+FP} \quad (4)$$

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (5)$$

$$MCC = \frac{TP.TN+FP.FN}{\sqrt{(TP+FN).(TP+FP).(TN+FP).(TN+FN)}} \quad (6)$$

where TP and TN are the number of positive and negative samples correctly classified by the predictive model. FP and FN represent the number of positive and negative samples misclassified.

Conclusions

Based on the research done on publicly available BAP databases and datasets, it was found that they have the following shortcomings:

- presence of duplicate information for one or more peptides in a given database;

- some data sets contain contradictory or wrong information;

- in the different databases, a given peptide appears with different biological activities, but there is no general information that gives a complete picture of the peptide;

- in some databases, there is not enough information about the physicochemical properties of the peptides, and in the available information there are sometimes errors;

- missing (with few exceptions) the possibility of extracting information about a random peptide sequence, as well as the presentation of the result in graphic form;

peptide database applications do not offer optimization and responsive design.

To overcome these limitations, it is necessary to develop a peptide database that contains information on diverse bioactive peptides and includes negative samples, which will help to develop even better MO models.

CHAPTER 2. Model of software application for the analysis and prediction of peptide properties by AI

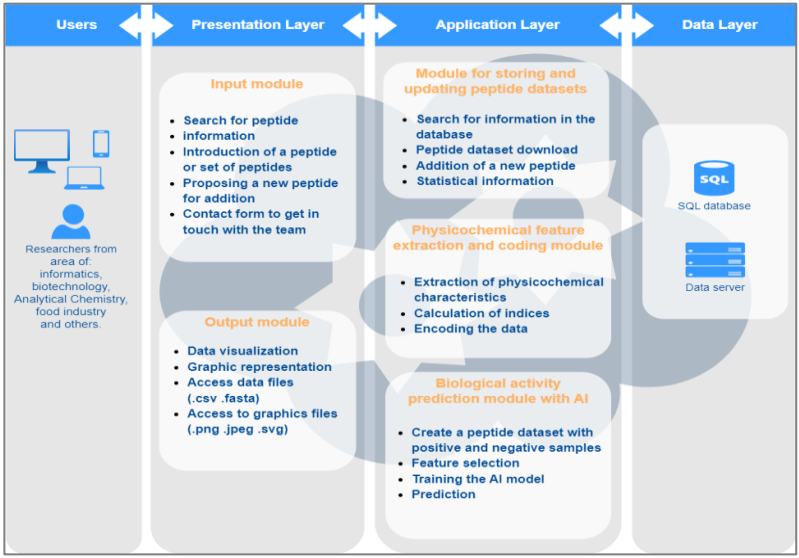


Figure 2. Architecture of an application for predicting biological activity of peptides with AI

The application is built based on a standard three-layer architecture containing a Presentation Layer, an Application Layer, and a Data Layer (Fig. 2).

It includes several modules:

- **Module for storing and updating sets of peptides**;
- **Module for extraction of physicochemical characteristics and coding**;
- **Module for prediction of biological activity through AI**;
- **Input Module**;
- **Output module.**

Module for storing and updating peptide datasets

This module contains the standard functionalities for a web application, such as searching for a peptide according to a set parameter, visualization according to classification criteria, and displaying detailed information about a given peptide (Fig. 3). Supports the possibility of outputting statistics - Czech information about the peptide sequences available in the database. The database must contain unique peptides, without duplication, with correct verified information for each peptide, with biological activities disclosed in scientific publications

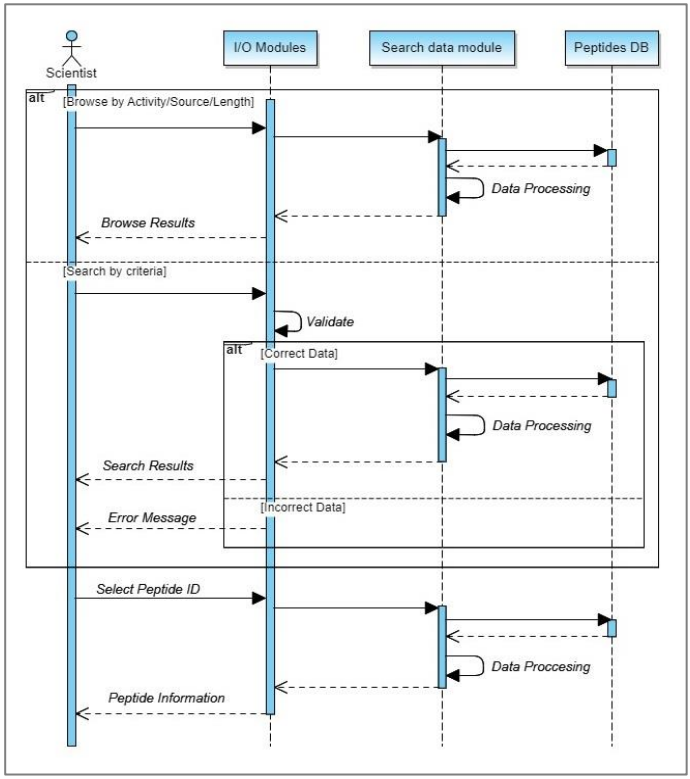


Figure 3. Sequence diagram for deriving detailed information about a peptide

and experimentally verified values.

Information about peptides can be visualized according to different classification criteria – activity, groups of organisms, peptide length, identifier, amino acid sequence, etc.

In the presented algorithm, the user has two search options - Browse and Search. With the Browse option, the available peptides are divided into three sections, Activity, Source and Length, and with Search, a certain search criteria is selected, for which the user enters information. It is validated and if it is not correct, a corresponding error message is returned. Both options return as a result a list of peptides (Browse Results or Search Results), from which the user can select a specific peptide (via the peptide ID) and the information about it will be visualized.

Module for extracting physicochemical characteristics and coding

Based on the amino acids contained in the peptide, the module extracts information about its physicochemical characteristics and calculates indices showing certain properties (Fig. 4). The module encodes the peptide according to the most popular encodings.

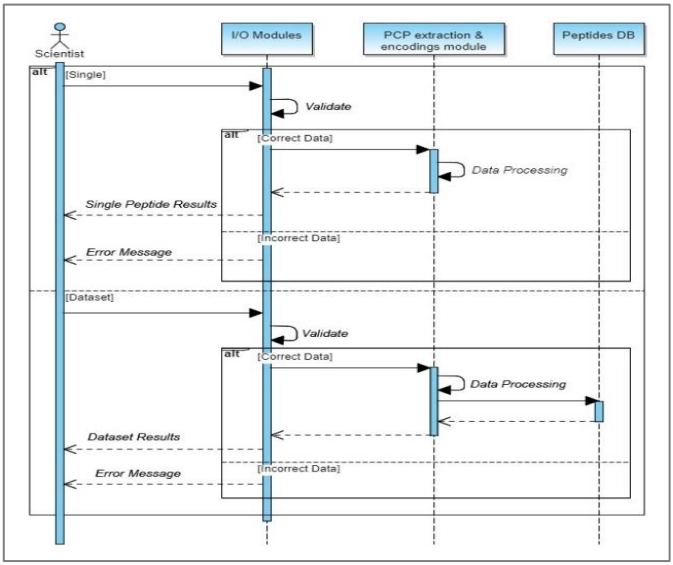


Figure 4. Sequence diagram of the physicochemical feature extraction and coding module

Acid Composition (AAC) is a procedure for calculating the composition of amino acids in a given peptide sequence. It shows the frequency distribution of each of the 20 basic amino acids.

$$AAC = [\mathrm{AA}_1\ \mathrm{AA}_2\ \ldots\ \mathrm{AA}_{20}] \quad (7)$$

$$AA_i = \frac{Frequency\ of\ AA_i}{L} \quad (8)$$

A vector AAC 1x20 described by the expression (7) is obtained, each value AA_i represents the percentage ratio of the amino acid in the peptide calculated by the formula (8).

In the Binary Profile Features (BPF) **encoding** of a peptide, each amino acid is described as a unit vector of length 20 where every position except one has the value 0. The end result is an L x 20 matrix (9) where L is the number of amino acids included in the peptide.

$$BPF = \begin{bmatrix} Binary\ AA_1(1x20) \\ \dots \\ Binary\ AA_L(1x20) \end{bmatrix} \quad (9)$$

Grouped AAC (Grouped AAC - GAAC) is calculated as the percentage ratio of different groups of amino acids in the peptide according to their properties. GAAC results in a 10-dimensional vector (10).

$$GAAC = [PCP_1 \ \ PCP_2 \ \ \dots \ PCP_{10}] \quad (10)$$

The atomic composition (Atomic Composition - ATC) shows the frequency of occurrence of the atoms in the peptide. The sum of all atoms of the corresponding species represents the atomic composition (11).

$$ATC = [\sum_{i=1}^{L} C_i \ \sum_{i=1}^{L} H_i \ \sum_{i=1}^{L} N_i \ \sum_{i=1}^{L} O_i \sum_{i=1}^{L} S_i] \quad (11)$$

Module for prediction of biological activity by AI

This module performs AI predictions and includes several basic steps (fig. 5):

- generating a coded data set with positive and negative samples;
- extraction of the most important and informative features;
- AI model training;
- test and validation or prediction.

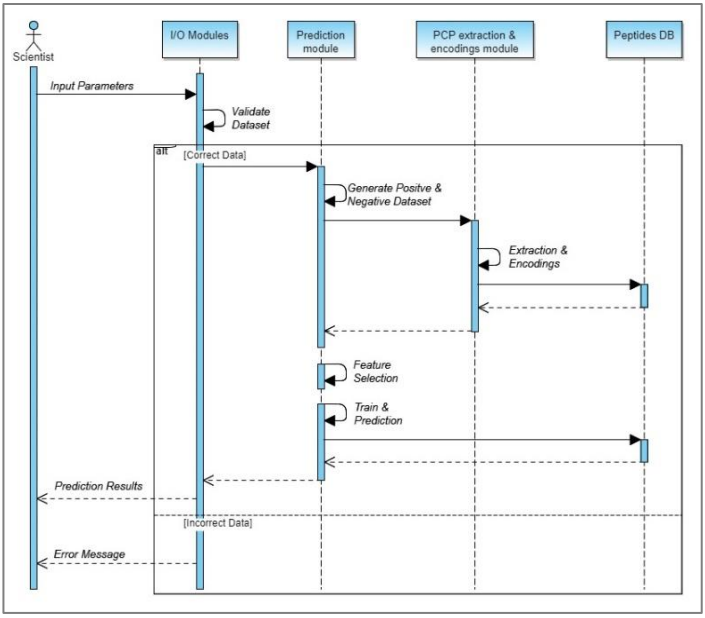


Figure 5. Sequence diagram of the biological activity prediction module by AI

The algorithm of operation of this module is presented in fig. 6. The creation of a data set of a given biological activity should include two classes of peptides - positive and negative. In the database of the application there are peptides with a certain type of activity. These peptides form the positive samples. The second class of samples may be composed of: *proven negative samples; peptides with activity other than the*

selected one; randomly generated peptides . The resulting set of numerical values is used to **select the most important and informative features** . Feature selection is one of the most important techniques in data preprocessing and is a mandatory part of modern machine learning algorithms [18]. Proper use of feature selection algorithms optimizes training by increasing training speed and/or reducing model complexity. There are three main models for feature selection: **the filter model (***filter feature selection* **)** [19]; the ***wrapper*** *feature selection* ***model*** **[** 20]; **the embedded model** *(embedded feature selection* **)** [21].

The software application envisages the use of three AI models to **predict biological activity** – the support vector method (SVM), random forest (RF), classical forward signal propagation neural network (FFNN).

The support vector method is a machine learning algorithm that is used for classification and regression. SVM solves an optimization problem in order to find an optimal hyperplane that best separates the data into the different classes. Given a training data set of *n* points of the form: $(\overrightarrow{x_1}, y_1), \ldots, (\overrightarrow{x_n}, y_n)$, where y_i takes the value 1, or −1. If we extend SVM to cases where the data are not linearly separable, a loss function is used:

$$max\ (0, 1 - y_i(\vec{w} \cdot \overrightarrow{x_i} - b)) \qquad (12)$$

Here y_i is *the i* -th target (ie, in this case 1 or −1). $\vec{w} \cdot \overrightarrow{x_i} - b$is *the ith* output, $\vec{w}$ is a normal vector to the hyperplane, and the parameter b odetermines the offset of the hyperplane from the origin.

This function is zero if the constraint in (13) is satisfied, in other words, if $\overrightarrow{x_i}$it lies on the right side of the field.

$$y_i(\vec{w} \cdot \overrightarrow{x_i} - b) \geq 1, \textit{ for each } 1 \leq i \leq n \qquad (13)$$

Then the goal of the optimization is to minimize the expression:

$$\left[\frac{1}{n}\sum_{i=1}^{n} max\ (0, 1 - y_i(\vec{w} \cdot \overrightarrow{x_i} - b))\right] + \lambda\|\vec{w}\|^2 \qquad (14)$$

The main advantage of SVM is the efficiency when working with a large data space and the ability to classify new observations whose classes are not known.

Random forest is a machine learning algorithm based on the concept of ensembles of decision trees. RF combines multiple decision trees, each trained on different parts of the data set. If the number of trees is B, then *b = 1, ... , B* . A random subset d is selected from the original data set *D,* and *m* is selected from all *p features*, where $m \leq p$. A decision tree T_b is generated. This operation is repeated and an ensemble of trees is obtained $\{T_b\}_1^B$. The end result is a decision based on majority voting:

$$\hat{f}_{RF}^{B}(x) = majority\ vote\ \{\hat{f}_b(x)\}_1^B \qquad (15)$$

RF is a very popular model of AI, which is why it finds various applications such as image processing, medical data analysis, and many more. etc.

A forward-propagation NM (FFNN) consists of a certain number of artificial neurons that are interconnected (Fig. 6). The neural network training process is an optimization task to minimize the error between predicted and actual values. Error backpropagation is a first-order gradient method learning method, admitting the possibility of falling into a local extrema, which is its most serious drawback.

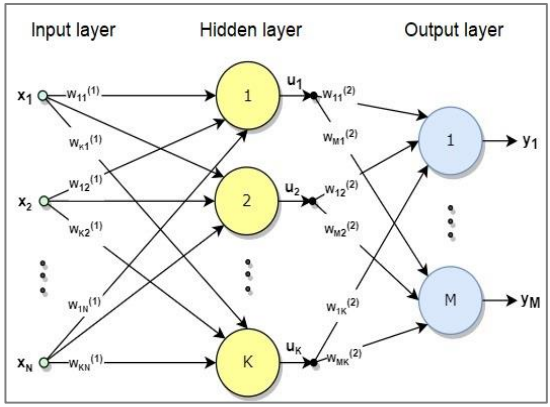


Figure 6. Feedforward neural network

$$E(w) = \frac{1}{2}\sum_{k=1}^{M}(y_k - d_k)^2 \qquad (16)$$

The objective function has the form (16). where M is the number of outputs, y_k the output signal of *the kth* neuron, a d_k is the expected output value of the neuron. The derivative of the objective function with respect to the weights of the output layer neurons is calculated:

$$\frac{\partial E}{\partial w_{ij}^{(2)}} = (y_i - d_i)\frac{df(u_i^{(2)})}{du_i^{(2)}}u_j, \qquad (17)$$

The gradient for the neurons of the hidden layer is determined according to the same principle, but for them the expression is more complicated:

$$\frac{\partial E}{\partial w_{ij}^{(1)}} = \sum_{k=1}^{K}(y_k - d_k)\frac{df(u_k^{(2)})}{du_k^{(2)}}w_{ki}^{(2)}\frac{df(u_i^{(1)})}{du_i^{(1)}}x_j, \qquad (18)$$

Determining the gradient vector is very important for the subsequent process of changing the weights. In backpropagation of the error, since a minimum is sought, a direction of negative gradient is set, therefore it is calculated:

$$\Delta w = -\eta\nabla \mathrm{E}(w) \qquad (19)$$

$$w(k+1) = w(k) + \Delta w, \qquad (20)$$

where η is the learning rate, and expression (20) is the rule by which the weight coefficients in the neural network are updated. Most often, the training of the network is terminated if the value of the gradient falls below a set threshold ε characterizing the accuracy of the training process.

By choosing one of the AI models, the ultimate goal of this module is realized, namely predicting the biological activity of a given peptide.

Input module

This module has the function of accepting user input and passing it to the module for which it is intended for processing. Information is entered when using the functionalities to search for

information in the database, analyze a peptide sequence or set of peptides, predict biological activity, suggest a new peptide for addition, register a user and contact the team.

Output module

The output data module provides a visualization of the processing results in graphical or textual form, as well as the ability to export the data in various graphical or tabular formats. It accepts data from all modules of the business logic layer and based on it creates downloadable files in formats suitable for further research. Visually, the results are presented in text or graphic form.

Conclusions

In this chapter, a model of a software application for the analysis and prediction of the properties of peptides through AI is presented, emphasizing a number of characteristic features of this process. A generalized application model is developed. A conceptual architecture with the main modules as well as the main functionalities are presented.

A module model was created for predicting the biological activity of peptides by AI. It selects features using one of three methods - filter, envelope and built-in. Forecasting can be done through one of 3 AI models. These are SVM, Random Forest and FFNN. They were chosen because they represent distinct groups of models in AI – linear classifier, ensemble model based on decision trees and classical neural network.

Based on the described software application model, a software prototype was implemented, presented in Chapter 3.

CHAPTER 3. Software application for the analysis and prediction of peptide properties by AI

Developing an application for AI peptide analysis is a complex process requiring knowledge from various scientific fields - informatics, biotechnology, biology, analytical chemistry, AI, etc. The process starts with designing a database and creating a user interface. The database must be structured in a way that helps to fulfill the tasks set in the dissertation work. Chapter 3 presents the software implementation of a prototype software application for the analysis and prediction of peptide properties by AI, following the model described in Chapter 2.

Used software technologies and tools

The software application will be used by researchers in different scientific fields who can use the different modules together and separately. To be widely accessible, the application is web-based.

WAMP package (Windows, Apache, MySQL, PHP) was used to create the prototype. **Apache HTTP Server v2.4.51** is a free and open-source web server. **PHP v8.1.0** is one of the most popular programming languages on the Internet. The open-source library **BioPHP is also used in the thesis work**. It includes classes for analysis of proteins, peptides, DNA and other tools. **MySQL v8.0.27** is

selected for database management systems. **Bootstrap 5.3.0** framework, the basic internet language **HTML v5.0**, **CSS v3** and **JavaScript** are used. For graphic implementation are used **Highcharts** libraries implemented with JavaScript. For the implementation of the AI models was used the Python language **v3.11**.

Designing a software application prototype

The user interface is responsible for sending data and requests to the business logic layer and displaying the result of the processing appropriately. The design implemented with Bootstrap is in a responsive version. English is chosen as the language of the application so that it can be widely used by researchers from all over the international scientific community.

The application layer includes three modules. Depending on the type of task, it can be performed by only one module or with the help of several of them.

The first module is the one through which **the information about the peptide sets is entered and updated**. It implements the functionalities described in Chapter 2. The second module **extracts physicochemical characteristics and indices and performs coding for a single peptide or set of peptides**. A number of features and indices are extracted and calculated and coding is performed. The third module is about **predicting the biological activity of the peptide**. This module cannot work independently, but uses the two modules already described: to generate a set of positive class of samples and to extract physicochemical characteristics, calculate indices and encode the created data set. It has the following functionalities: creates a peptide set including a positive and a negative class;

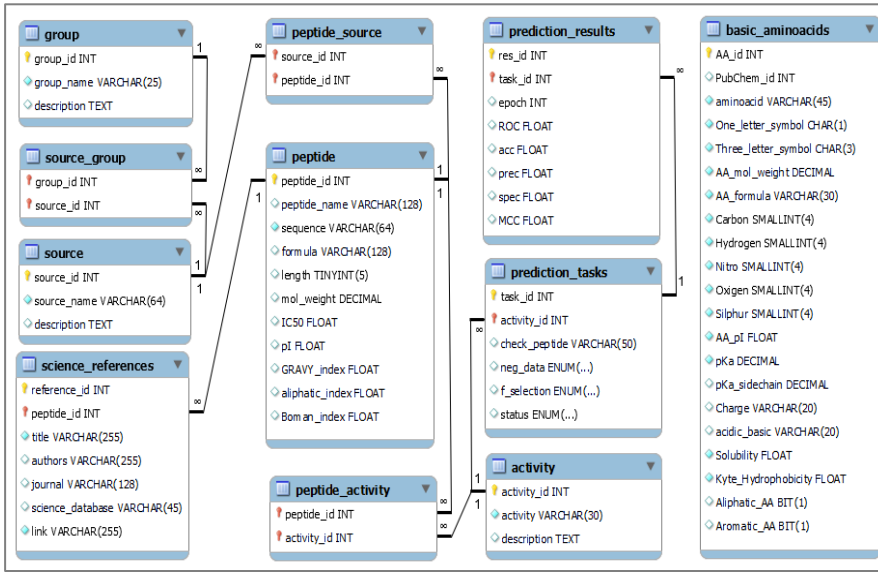


Figure 7. Physical model of the software application database

performs feature selection; does training of the selected model by AI; records the results in the database; transmit to visualize the result.

The database plays a key role in the operation of the application. It contains information about available peptide sequences, amino acid structure, values on different scales and properties, results of biological activity prediction.

The physical model of the complete database is presented in Fig. 7. It consists of 11 tables. Some main tables serving the functionality of the application are: table *peptide* contains the basic information about the peptides; table *basic_aminoacids* contains information about the 20 basic amino acids. It includes the chemical formula, the number of atoms of its constituent chemical elements, acid dissociation constant, hydrophobicity index according to the Kite-Doolittle scale, etc.; table *activity* contains information about the types of biological activity in the database. Detailed information about all tables from the database is presented in Appendix 1 of the dissertation.

Realization of the prototype

This stage consists of creating the design, navigation, description of the processes, including specific parameters, and creating the program code, using the selected software technologies and tools, to obtain a working prototype. The module for storing and updating sets of peptides is called **Pep Lab**. It is responsible for displaying information in all sections, except for “Tools” (Fig. 8).

Figure 8. Software application navigation bar

The module for extracting physicochemical characteristics and encoding (**DMPep)** , together with the module for predicting biological activity through AI (**PepAI)** , serve the processes in "Tools".

Only unique, duplicate-free peptides that have undergone a validation process are entered into **Pep Lab.** As of 31.12.2023 there are **2775 peptides** in it. Peptide sequence information available at Pep Lab is available at *pep-lab.info*. The "Database" section offers two options - search by selectable criteria and search by parameter - activity, group or length.

The data in Pep Pab are extracted from several public databases - BIOPEP-UWM [22], APD3 [23] and AHTPDB [24]. The source of information about new peptides is the scientific publications in the National Library of Medicine at the National Center for Biotechnology Information of the USA [25], publications indexed in Scopus and Web of science by keywords - bioactive peptides, food-derived bioactive peptides, novel bioactive peptides, etc.

The "Statistics" section shows summary information about the peptides available in the database, and "Downloads" provides downloads for each activity in .csv and .fasta format.

The analytical module (**DMPep**) provides a complex analysis of physicochemical characteristics of a peptide sequence. Using an amino acid sequence, the module can provide information on a number of characteristics:

Peptide length is determined by the number of amino acids that are contained in the peptide.

Molecular Weight (MW) is the sum of the masses of all atoms in the peptide. It is calculated by summing the molecular weights of the corresponding amino acids involved in the peptide, by subtracting the weights of the separated molecules when creating the peptide bonds (21).

$$MW = (\sum_{i=1}^{L} AAMW_i) - (L-1).MW_{H2O} \quad (21)$$

Net charge is the sum of the charges on all amino acids and the end terminals. The net charge can be calculated by formula (22), which is a variant of the Henderson–Hasselbalch equation presented by Moore [26].

$$Q = \sum Q^- + \sum Q^+ \quad (22)$$

Using formulas (23) and (24) is calculated, respectively Q^- and Q^+:

$$Q^- = \frac{(-1)}{1+10^{(pK_a - pH)}} \quad (23)$$

$$Q^+ = \frac{(+1)}{1+10^{(pH - pK_a)}} \quad (24)$$

In these expressions, pK_a reflects the acid dissociation constant K_a as a negative decimal logarithm.

The isoelectric point (pI) is the pH value at which the net charge is zero ie. the peptide is electrically neutral. DMpep calculates the net charge at pH 2 to 14.

Hydrophobicity index (GRAVY) is calculated as the sum of the hydrophobicity of each amino acid divided by the number of amino acids in the peptide (25).

$$GRAVY = \frac{\sum_{i=1}^{L} H_i}{L} \quad (25)$$

The aliphatic index (AI) is a measure of the aliphatic amino acid content of a peptide or protein. The aliphatic index was calculated by formula (26) proposed by Ikai [28]. Coefficients a and b are the relative side chain value of Valine and Isoleucine/Leucine, respectively.

$$AI = X_A + a.X_V + b.(X_I + X_L) \quad (26)$$

The acidity index **(ABI)** of the peptide is determined by formula (27) where ABI_A and ABI_B are the mole percentages of amino acids with acidic and basic properties in the peptide.

$$ABI = \begin{cases} Acidic, & ABI_A > ABI_B \\ Basic, & ABI_A < ABI_B \\ Neutral, & ABI_A = ABI_B \end{cases} \quad (27)$$

The peptide binding potential index (Bohmann index) [29] indicates the peptide binding potential. Bohmann's index is determined by formula (28), in which S_i is the solubility according to the scale of Radzicka and Wolfenden [30].

$$PBPI = \frac{\sum_{i=1}^{L} S_i}{L} \quad (28)$$

The AI Biological Activity Prediction (PepAI) module, following the algorithm described in Chapter 2, performed the following main actions:

1. A set of peptides is created, containing two classes - positive and negative. The information from the set is digitally encoded and normalized.
2. If the feature selection method is selected, the optimal subset is determined.
3. The selected model is trained by artificial intelligence.
4. It predicts whether the given peptide is of the selected activity class.
5. The result is passed for visualization.

Conclusions

In Chapter 3, the implementation of a prototype software application for the analysis and prediction of peptide properties by AI is presented. Design stages, software technologies and tools used are described. The database structure and functionalities of all modules are presented.

A prototype of the application was implemented, including the designed modules. A Pep Lab peptide database containing peptides with known biological activity has been created. As of 31.12.2023, there are **2775 peptides in it**. Pep Lab offers the possibility to present the complete data for each individual peptide, the possibility to download the peptides by activities and statistical information about the database.

The implemented analytical module (DMPep) provides analysis of peptide physicochemical characteristics and peptide coding according to six variants: PCP, BPF, AAC, GAAC, DPC and ATC.

The PepAI module performs prediction of the biological activity of an arbitrary peptide, providing the ability to select different data sets, different types of encoding of these data, as well as choosing one of three AI models - SVM, Random Forest and NM. The mathematical apparatus of each of the models is described, as well as its software implementation.

CHAPTER 4. Experimental results

This part of the dissertation presents the obtained experimental results. The implemented SVM, RF and FFNN models are tested in various experiments. For this purpose, four data sets were generated, including different numbers of positive samples: DS1, DS2, DS3, DS4 (Table 2). DS4 was

Table 2. Data sets used in the dissertation research

Name	*Positive set*	*Negative set*	*Total*
DS1	71 neuropeptides	71 random samples from the database	142
DS2	232 antidiabetic peptides	232 random samples from the database	464
DS3	523 peptides, lower blood pressure	523 random samples from the database	1046
DS4	1333 peptides below. Blood Pressure	1333 random samples from the database	2666

obtained by augmenting the 523 peptides dataset with 810 peptides taken from the AHTPD database. The following experiments were carried out through them:

- Using the DS3, the metrics of the models tested when the structural parameters were changed;
- The influence of the feature selection method on the model was investigated;
- The influence of different sets on the parameters of the model was studied.

The following metrics were used to evaluate the models – area under the performance curve, classification accuracy, precision, specificity, weighted harmonic mean precision.

Structural identification of AI models

This experimental setup aims to tune the structural parameters of the models so that they perform best in peptide classification. In **the SVM model,** the influence of different types of kernels on the metrics was investigated. The tests were done with a regularization parameter C=1 and a tolerance of 0.001. The experiment results show that the best results are obtained with SVM with **RBF kernel**. In **the RF model,** the number of trees is varied and the performance of the model is observed. The model has the best performance with **200 trees**. To determine the architecture of **the FFNN model**, a large number of experiments have been implemented, varying the number of neurons from 5 to 100 with a step of 5 and checking the performance at 1, 2, 3 and 4 hidden layers. The results show that it most accurately classifies the FFNN with **1 hidden layer with 10 neurons**.

From the experimental results, it can be concluded that *the use of NM is appropriate in the presence of a large volume of input data. In the specific case of classifying BAP, NM should be applied in the presence of at least 1000 positive samples.*

Investigating the influence of the feature selection method

In this group of experiments, the aim is to investigate the influence of different variants of feature selection on the rate of change of output metrics.

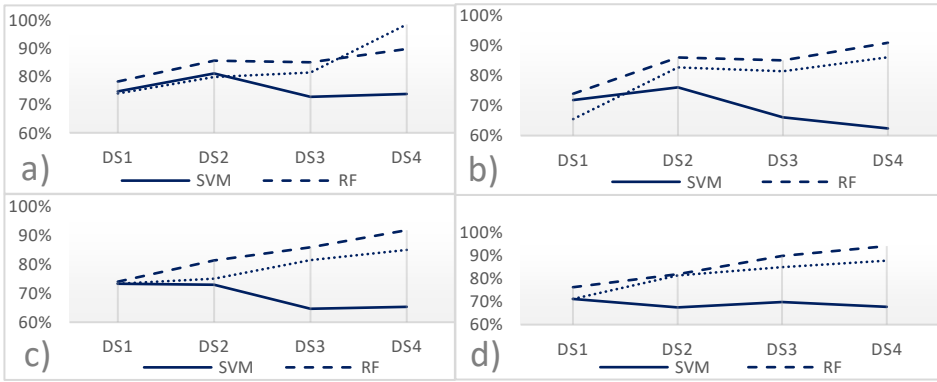


Figure 9. Change in the accuracy of the models with: a) no feature selection method; b) χ^2; c) ANOVA; (d) RFEs

Each peptide in the dataset contains a vector of 41 features according to the following encodings: PCP (6 characters), ATC (5 characters), AAC (20 characters), GAAC (10 characters). Four variants were studied – no feature selection and the 20 best features selected, by χ^2, ANOVA and RFE. The graphical representation of the results are shown in fig. 9.

From the graphs presented, it can be seen that the choice of characteristics has the most significant impact on the RF model's performance. The classification accuracy of this model increases when using feature selection, with RFE being the best. The use of a feature selection method causes a decrease in the accuracy of the FFNN model, but it also leads to a decrease in the number of inputs and, accordingly, the computational burden. In this model, the researcher must consider whether to use feature selection. In the SVM model, the result is worse than that without the feature selection method, so it is better not to apply this technique to it.

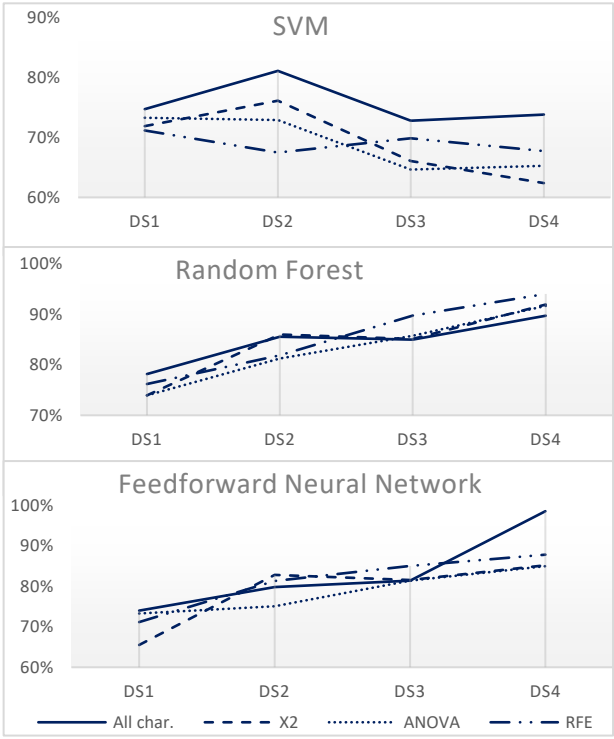


Figure 10. Change in accuracy of SVM, RF and FFNN depending on feature selection method

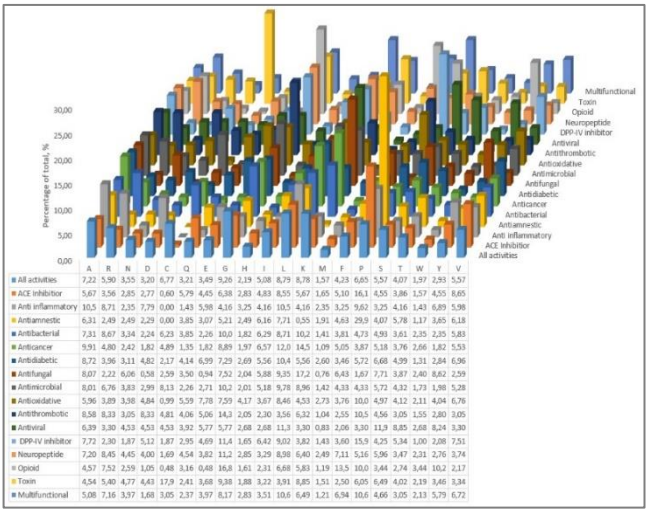
Similar conclusions can be drawn when comparing the methods used to select the characteristics of the different models (Fig. 10).

From these graphs it is clear that SVM performs best without a feature selection method, with FFNN it is good to use RFE or χ^2, but only on datasets below 500 positive samples. With more than 1000 positive samples, the FFNN model is fully capable of determining the most important features and does not need to use any method. In the RF model, the RFE method is best, but the size of the data set is also important.

Overall, the research shows that feature selection should not be used in SVM, in NM it can be used at the discretion of the researcher, and in RF it is mandatory to use this technique.

Research using statistical data from the database

The statistics of the peptides available in the database show the distribution of amino acids in the peptides for each activity (Fig. 11). It can be seen that the amino acids most often found in peptides are: alanine (A), arginine (R), cysteine (C), glycine (G), isoleucine (I), leucine (L), lysine (K), proline (P), serine (S), valine (V). These are the amino acids that have **more than 5% participation, as a relative frequency of occurrence of the peptides** in the database. It is noteworthy that there is a very high percentage of cysteine (C) in the toxic peptides – 17.9%, and therefore it can be considered as an indicator of toxicity.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
All activities	7,22	5,90	3,55	5,20	6,77	5,21	3,49	5,26	2,19	5,08	8,79	8,78	1,57	4,23	6,65	5,57	4,07	1,97	2,93	5,57
ACE inhibitor	5,67	3,56	2,85	2,77	0,60	5,79	4,45	6,38	2,83	4,83	8,55	5,67	1,65	5,10	16,1	4,55	3,86	1,57	4,55	8,65
Anti inflammatory	10,5	8,71	2,35	7,79	0,00	1,43	5,59	4,16	3,25	4,16	10,5	4,16	2,35	3,25	9,62	3,25	4,16	1,43	6,89	5,98
Antiamnestic	6,31	2,49	2,49	2,29	0,00	3,85	3,07	5,21	2,49	6,16	7,71	0,55	1,91	4,63	29,9	4,07	5,78	1,17	3,65	6,18
Antibacterial	7,31	8,67	3,34	2,24	6,23	3,85	2,26	10,0	1,82	6,29	8,71	10,2	1,41	3,81	4,73	4,91	3,61	2,35	2,35	5,83
Anticancer	9,91	4,80	2,42	1,82	4,89	1,35	1,82	8,89	1,97	6,57	12,0	14,5	1,09	5,05	3,87	5,18	3,76	2,66	1,82	5,53
Antidiabetic	8,72	3,96	3,11	4,82	2,17	4,14	6,99	7,29	2,69	5,56	10,4	5,56	2,60	3,46	5,72	6,68	4,99	1,31	2,84	6,96
Antifungal	8,07	2,22	6,06	0,58	2,59	3,50	0,94	7,52	2,04	5,88	9,35	17,2	0,76	6,43	1,67	7,71	3,87	2,40	8,62	2,59
Antimicrobial	8,01	6,76	3,83	2,99	8,13	2,26	2,71	10,2	2,01	5,18	9,78	8,96	1,42	4,33	4,33	5,72	4,32	1,73	1,98	5,28
Antioxidative	5,96	3,89	3,98	4,84	0,99	5,59	7,78	7,59	4,17	3,67	8,46	4,53	2,73	3,76	10,0	4,97	4,12	2,11	4,04	6,76
Antithrombotic	8,58	8,33	3,05	8,33	4,81	4,06	5,06	14,3	2,05	2,30	3,56	6,32	1,04	2,55	10,5	4,56	3,05	1,55	2,80	3,05
Antiviral	6,39	3,30	4,53	4,53	4,53	3,92	5,77	5,77	2,68	2,68	11,3	3,30	0,83	2,06	3,30	11,9	8,85	2,68	8,24	3,30
DPP-IV inhibitor	7,72	2,30	1,87	5,12	1,87	2,95	4,69	11,4	1,65	6,42	9,02	3,82	1,43	3,60	15,9	4,25	5,34	1,00	2,08	7,51
Neuropeptide	7,20	8,45	4,45	4,00	1,69	4,54	3,82	11,2	2,85	3,29	8,98	6,40	2,49	7,11	5,16	5,96	3,47	2,31	2,76	3,74
Opioid	4,57	7,52	2,59	1,05	0,48	3,16	0,48	16,8	1,61	2,31	6,68	5,83	1,19	13,5	10,0	3,44	2,74	3,44	10,2	2,17
Toxin	4,54	5,40	4,77	4,43	17,9	2,41	3,68	9,38	1,88	3,22	3,91	8,85	1,51	2,50	6,05	6,49	4,02	2,19	3,46	3,34
Multifunctional	5,08	7,16	3,97	1,68	3,05	2,37	3,97	8,17	2,83	3,51	10,6	6,49	1,21	6,94	10,6	4,66	3,05	2,13	5,79	6,72

Figure 11. Frequency distribution of amino acids in total and by activities.

Given this information, a new feature selection method called **ComStat** (Complex Statistics) is proposed. The most important characteristics are the specified amino acids and the physicochemical characteristics (length, isoelectric point, molecular mass and the three indices).

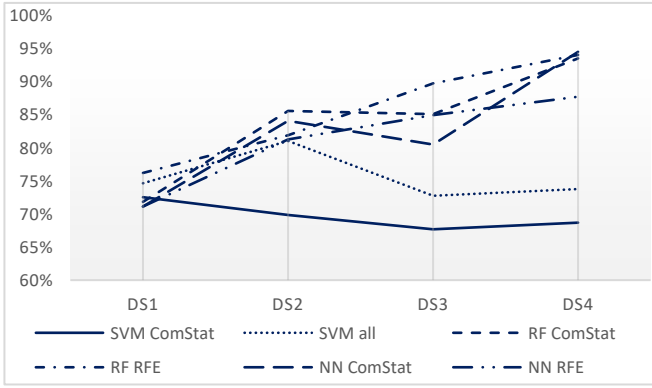


Figure 12. Accuracy of feature selection using statistics for four datasets.

After correlation analysis between the features, it is clear that there is a high degree of correlation between length and molecular weight. This indicates that one of them is redundant, and the presence of two highly correlated features may lead to overtraining of the model. Molecular mass was chosen to be removed as a redundant feature in the set.

With the feature set thus selected, prediction was performed with the three AI models.

The results of applying the ComStat method are presented in fig. 12. From the graphs, it can be seen that there is a significant difference with the SVM model, and the graph without selection is significantly better than the one after selection by the ComStat method. For the other two models, however, the graphs show that with and without feature selection, the values are close, and the neural network even has better results. For the two larger data sets, the results are identical.

This shows that the proposed ComStat input feature reduction method can be used in RF and NM for a set of over 500 peptides without loss of accuracy.

After adding the results obtained by the ComStat method to the graphs of fig. 10, comparative graphs are obtained between all described methods for selecting characteristics (Fig. 13).

The graphs show that the ComStat method is comparable to the other considered feature selection methods.

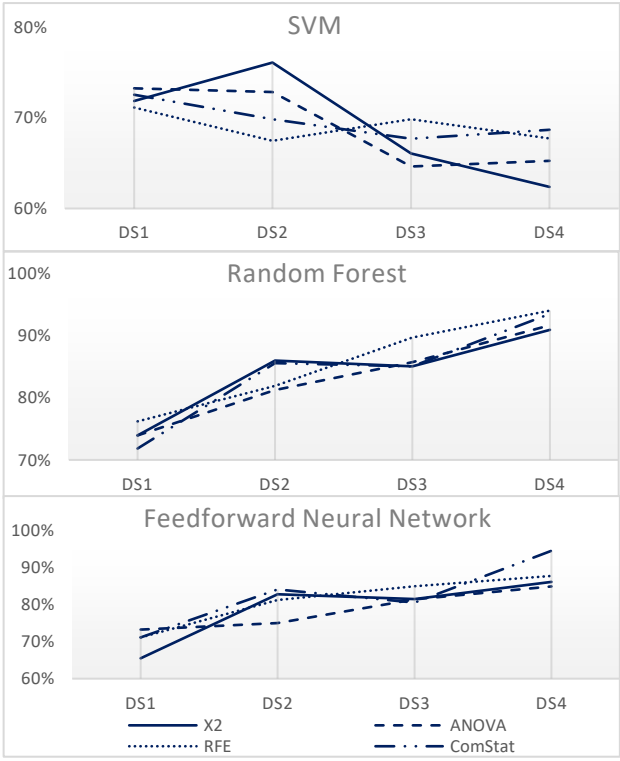


Figure 13. Comparison of ComStat and X^2 prediction accuracy, ANOVA and RFE

In conclusion, several conclusions can be drawn when applying the ComStat method. It is good to use as a feature selection algorithm, in RF and NM, when the set is over 500 positive samples. For smaller sample sets, it can be successfully applied to the SVM model.

An advantage of the method is that the statistical analysis is performed beforehand with the available peptides in the database and then the selected features are used. This leads to a significantly smaller computational burden in the forecasting process, correspondingly increased speed.

It should be noted that the proposed method was only experimented with the DB with peptides realized within the

framework of the dissertation research.

Prediction of biological activity

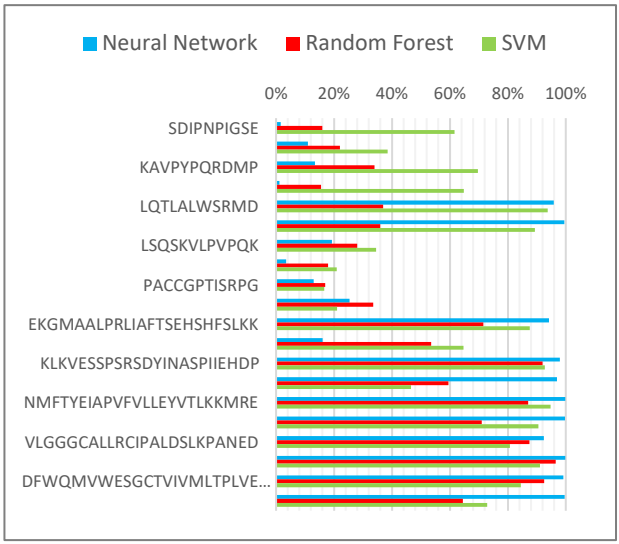


Figure 14. Prediction of antidiabetic peptides

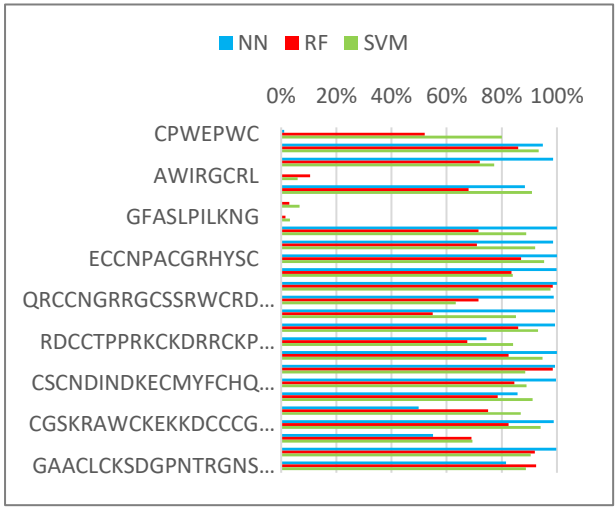


Figure 15. Prediction of blood pressure-lowering peptides.

To investigate the ability of AI models to predict, an experiment was performed with the following setup: the available peptides from three activities in BD - anti-diabetic, blood pressure-lowering and toxin-positive samples were selected. In this way, three sets of data were obtained, one for each activity, of which 10% were set aside for an independent test. With the remaining samples, the SVM, RF, FFNN models were trained.

According to the results already obtained, no feature selection method was used for SVM, and RFE method was applied for RF and FFNN. There are 41 features per peptide in the datasets. Each of the models predicted the probability of the peptides from the independent set belonging to a corresponding activity. The results of the experiment are presented graphically in fig. 14. From the graphical representation, it can be seen that SVM performs best in prediction, especially for short-length peptides. This is due to the small data set (232 peptides) where the other two models cannot be trained well. Their prediction improves for peptides with a longer length - over 20 amino acids. The predicted

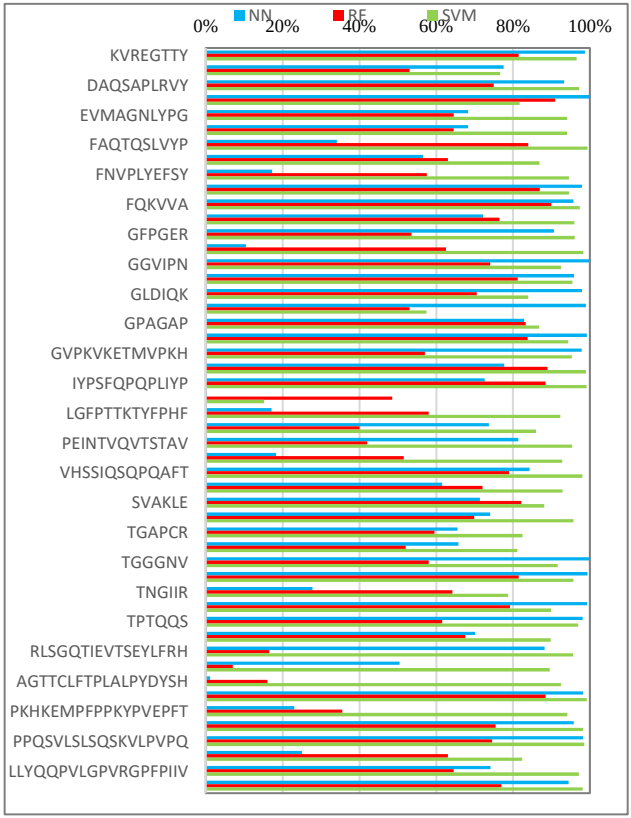


Figure 16. Prediction of toxic peptides

results are not accurate enough. Assuming that a peptide is recognized with a probability above 0.50, SVM recognizes 75% of peptides, FFNN – 65%, and RF only 50%. An interesting fact is that when FFNN correctly classifies a peptide, it does so very strongly, with a probability of over 90%. FFNN showed very good prediction for peptides longer than 20 amino acids - only one peptide was not recognized.

For blood pressure-lowering peptides, the results obtained are presented in Fig. 15.

In this activity, RF and FFNN do much better, which is due to the doubling of positive samples in the set - 523. All three models have a high sample recognition rate: FFNN - 82%, RF - 88%, a SVM with over 94% peptides recognized. RF predicts with higher accuracy than FFNN. It should be noted that predictions made with RF have a probability of about 0.5. More than 40% of the known peptides have a probability in the range 0.50-0.70, which does not give certainty to the prediction.

For toxic peptides, despite the fact that the peptide set is small (250 samples), all three models recognize them very well (Fig. 16). A prerequisite for this is their "opposite" properties (toxicity). FFNN has a successful prediction of 85% of the peptides because due to the small amount of data, it fails to train well, and RF and SVM recognize 90% of the toxic peptides.

A categorical determination of activity is also observed. For unrecognized samples, the predicted probability is up to 0.10, and for recognized samples almost entirely above 0.70. This leads to the conclusion that this class of peptides, due to their specific properties, is easily recognizable for the studied models by AI.

Conclusions

Chapter 4 presents the experimental results obtained in peptide classification with the implemented SVM, RF and FFNN models and four datasets. SVM is found to perform best when using a non-linear **RBF kernel**, RF with a **200-tree network**, and a straight-link NM should have **1 hidden layer with 10 neurons** and **a ReLu** activation function.

The influence of the feature selection method, using 4 techniques, as well as the influence of the data set size on the performance of the model were investigated. The results show that the SVM model is not affected by the size of the data set, and **the application of a feature selection method negatively affects the results**. The conclusion is that **for small data sets it is appropriate to use SVM**.

RF is the model most significantly affected by the feature selection technique. The prediction quality is improved, with the best results being achieved when using the RFE wrapper method.

Research shows that successful classification with **FFNN strongly depends on the size of the dataset**. This gives reason to introduce the following rule: **NM should be used for a dataset with at least 1000 peptides. This is done in order to train the network well and achieve maximum performance**.

A new feature selection method, **ComStat**, based on statistical indicators, is also proposed. The amino acids with the greatest relative weight of the peptides available in the DB and the physicochemical characteristics are used as characteristics. **This method improves the classification accuracy of BAP at The RF and NM models for a data set with over 500 peptides**.

The results of the experiments with independent sets confirm the described conclusions, but add one more - when peptides with a length of up to 20 amino acids are to be recognized, **it is best to use the SVM model**.

Conclusions

This dissertation is an interdisciplinary study. It is dedicated to the use of techniques from artificial intelligence to analyze and predict the properties of peptides.

From the literature review, the need for a software product that combines an intuitive user interface with a reliable database and tools for peptide prediction and analysis is clear. All this led to the formulation of the main goal of the dissertation work, namely to create a prototype of a software system that contains:

- ✓ a database including peptides with known biological activity;
- ✓ a tool aimed at complex analysis and extraction of physico-chemical characteristics of peptides and creation of data structures including basic parameters of the peptide;
- ✓ model creation using AI methods to predict the biological activity of peptides.

For the realization of this goal, the following main tasks have been solved:

1. To explore the experience of scientists in predicting the biological activity of peptides using artificial intelligence methods.

2. To survey publicly available peptide coding databases and tools.
3. To explore artificial intelligence methods that are used to predict the biological activity of peptides.
4. To develop a prototype software application including a database, a module for extraction of physicochemical characteristics of peptides and a peptide biological activity prediction module.
5. To create a database of peptides with activity and properties proven in the scientific literature.
6. To build, train and test artificial intelligence models to predict the biological activity of peptides.
7. Analyze the obtained results and evaluate each of the models.

Contributions

The main contributions of the dissertation can be characterized as scientific, scientific-applied and applied.

Scientific contributions of the dissertation research:

1. Developed models for predicting the biological activity of peptides - SVM, RF and NNs.
2. ComStat feature selection method based on peptide statistical metrics was developed.

Scientific and applied contributions of the dissertation research:

3. Developed a conceptual model of a software application for the analysis and prediction of peptide properties by various AI methods.
4. Implemented algorithms for dynamic computation of peptide features and peptide coding.
5. Implemented artificial intelligence algorithms to predict the biological activity of peptides based on SVM, RF and artificial NNs.

Applied contributions of the dissertation research:

6. A database of peptides was created and information was entered for 2775 peptides with known biological activities.
7. A software application was developed for the analysis and prediction of the physicochemical properties of peptides. It is freely available at: www.pep-lab.info.

Prospects for future development

In the process of work, new interesting ideas have been generated to continue the work on improving the Peplab platform. Among them are:

• Continuous updating of database records. This would enable significantly more reliable statistical analyses;

• Development of the BAP prediction module with the inclusion of deep learning algorithms, which also requires a database with more records;

• Creation of a news module, through which researchers will be better informed about novelties in the platform;

• Implementation of a system for registering users with different roles and with different levels of access;

• Transition to structural bioinformatics, ie. inclusion in the database of images of peptide sequences, their recognition with deep learning algorithms, etc.

Approbation

The main results of the research have been reported at national and international scientific forums. The results of the research are presented in 5 publications - 3 in specialized journals, 1 collection of international conference and 1 collection of papers from a Bulgarian conference. Two of the publications are indexed in international databases Scopus and Web of Science, and one of them was published in an edition with an impact factor: 2.7.

Results of the dissertation research were reported at 2 international and 1 national conferences, as well as at 3 research seminars.

Conference reports:

1. International Conference on Artificial Intelligence and E-Leadership, October, 15-16, 2020, Plovdiv, Bulgaria.
2. 69th Scientific Conference with International Participation "Food science, engineering and technology - 2022", September 29-30, 2022, Plovdiv, Bulgaria.
3. Science Conference "Science Days 2021", November 25-27, Plovdiv, Bulgaria.

Workshop reports:

1. "Using artificial intelligence techniques for predicting biological activity of peptides" , *Scientific seminar "Neural network modeling with applications in business",* 11/13/2020 , Plovdiv.
2. "Classification of amino acid sequences by machine learning", *XVII Scientific Seminar of the School of ICT Innovations*, 29.11.2021, Plovdiv.
3. "In silico analysis of the physicochemical characteristics of peptides", *XIX Scientific seminar of the ICT Innovation School*, 23.11.2022, Plovdiv.

The accumulated knowledge during doctoral studies allowed, in co-authorship with colleagues from the Department of Computer Informatics, the publication of **the textbook "Introduction to Databases"** . In addition, **materials were developed for conducting classes on the disciplines "Databases"** and **"Programming on the Internet with PHP/MySQL"** , which were used in conducting classes at PU "Paisiy Hilendarski" in 2019-2023.

List of publications on the topic of the dissertation

1. **Terziyski, Z.**, Terziyska, M., Deseva, I., Hadzhikoleva, S., Krastanov, A., Mihaylova, D., Hadzhikolev, E., (2023) Pep Lab Platform: Database and Software Tools for Analysis of Food-Derived Bioactive Peptides. Applied Sciences. 13(2): 961. https://doi.org/10.3390/app13020961 (**Scopus and Web of Science, Q2, IF:2.7**)

2. **Terziyski, Z.**, Terziyska, M., Hadzhikoleva, S., & Desseva, I., (2023) A software tool for data mining of physicochemical properties of peptides. In BIO Web of Conferences (Vol. 58, p. 03007). EDP Sciences. https://doi.org/10.1051/bioconf/20235803007 (**Scopus**)
3. Terziyska, M., Desseva, I., **Terziyski, Z.**, (2021) "Food-Derived Bioactive Peptides And Artificial Intelligence Techniques For Their Prediction: A Brief Review", International journal of scientific & technological research vol. 10 (08), ISSN 2277-8616.
4. Terziyska, M., Desseva, I., **Terziyski, Z.**, (2020) "Deep learning algorithm for food-derived antioxidative peptides bioactivity prediction", Journal of informatics and innovative technologies 4 (2), ISSN: 2682–9517 (print) ISSN: 2683–0930 (online).
5. **Terziyski, Zh.**, (2021) "Classification of antioxidant peptides using machine learning", Scientific papers of the Union of Scientists in Bulgaria - Plovdiv. Series C. Technique and technologies. Volume XIX, ISSN: 1311-9419 (Print); ISSN 2534-9384 (Online).

Noted citations

Publication: **Terziyski, Z.** , Terziyska, M., Deseva, I., Hadzhikoleva, S., Krastanov, A., Mihaylova, D., Hadzhikolev, E., (2023) Pep Lab Platform: Database and Software Tools for Analysis of Food-Derived Bioactive Peptides. Applied Sciences. 13(2): 961. https://doi.org/10.3390/app13020961.

Cited in: Gaffar, S., Hassan, MT, Tayara, H., & Chong, KT (2023). IF-AIP: A machine learning method for the identification of anti-inflammatory peptides using multi-feature fusion strategy. Computers in Biology and Medicine, 107724.

Publication: **Terziyski, Z.** , Terziyska, M., Hadzhikoleva, S., & Desseva, I., (2023) A software tool for data mining of physicochemical properties of peptides. In BIO Web of Conferences (Vol. 58, p. 03007). EDP Sciences.

Cited in: Sergeeva, I., Permyakova, L., Markov, A., Ryabokoneva, L., Atuchin, V., Anshukov, A., ... & Proskuryakova, L. (2023). Peptides of Yeast Saccharomyces cerevisiae Activated by the Aquatic Extract of Atriplex sibirica L. ACS Food Science & Technology.

Participation in scientific research projects

1. *SP19-FMI-012* , "Neural network modeling with applications in business", financed by the "Scientific Research" Fund at the PU "P. Hilendarski" (2019-2020).
2. *SP19-FMI-004* , "Formation of scientific competences in the Student School for ICT Innovations", financed by the "Scientific Research" Fund at the PU "P. Hilendarski" (2019-2020).
3. *KP-06-M36/2* , " Preparation, characterization and purification of protein enzyme hydrolysates from lupine and prediction of their biological activity using artificial intelligence methods", financed by the "Scientific Research" Fund (2019-2022).
4. *MU21-FMI-004* , "Support for scientific research in a school for ICT innovations", financed by the "Scientific Research" Fund at PU "P. Hilendarski" (2021-2022).
5. *08/21-H* , " Development of a cloud-based module for the analysis of peptide sequences extracted from food " , financed by the "Science" fund of the UHT-Plovdiv (2021-2022).

6. *MUPD23-FMI-021* , "Use of artificial intelligence methods in business", financed by the "Scientific Research" Fund at the PU "P. Hilendarski" (2023-2024).

References

[1] A. Padhi, M. Sengupta, S. Sengupta, K. Roehm and A. Sonawane, "Antimicrobial peptides and proteins in mycobacterial therapy: current status and future prospects," Tuberculosis, vol. 94, no. 4, pp. 363-373, 2014.

[2] A. Sanchez and A. Vazquez, "Bioactive peptides: A review," Food quality and safety, vol. 1, no. 1, pp. 29-46, 2017.

[3] PrecedenceResearch, "Peptide Therapeutics Market," 2023. [Online]. Available: https://www.precedenceresearch.com/peptide-therapeutics-market. [Accessed 31. 12. 2023].

[4] B. Manavalan, S. Basith, T. Shin, S. Choi, M. Kim and G. Lee, "MLACP: machine-learning-based prediction of anticancer peptides," Oncotarget, vol. 8, no. 44, p. 77121, 2017.

[5] B. Manavalan, S. Basith, T. Shin, L. Wei and G. Lee, "AtbPpred: a robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees," Computational and Structural Biotechnology Journal, vol. 17, pp. 972-981, 2019.

[6] T. Simonson, Computational Peptide Science, New York: Springer, 2022.

[7] J. Zurada, Introduction to artificial neural systems, West Publishing Co., 1992.

[8] TK Ho, "Random decision forests," IEEE In Proceedings of 3rd international conference on document analysis and recognition, vol. 1, pp. 278-282, 1995.

[9] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, pp. 273-297, 1995.

[10] Y. Xiong, J. Liu, W. Zhang and T. Zeng, "Prediction of heme binding residues from protein sequences with integrative sequence profiles," Proteome science, vol. 10, no. 1, pp. 1-8, 2012.

[11] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee and S. ... Zhao, "Applications of machine learning in drug discovery and development," Nature reviews Drug discovery, vol. 18, no. 6, pp. 463-477, 2019.

[12] B. Manavalan, J. Lee and J. Lee, "Random forest-based protein model quality assessment (RFMQA) using structural features and potential energy terms," PloS one, vol. 9, no. 9, p. e106542, 2014.

[13] M. Zhang, LFT Marquez-Lago, A. Leier, C. Fan, C. Kwoh and C. Jia, "MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters," Bioinformatics, vol. 35, no. 17, pp. 2957-2965, 2019.

[14] 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 Proceedings of the 2017 4th International Conference on Biomedical and Bioinformatics Engineering, pp. 91-96, 2017.

[15] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. 3, pp. 1157-1182, 2003.

[16] Z. Chen, P. Zhao, F. Li, T. Marquez-Lago, A. Leier, J. Revote 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 in bioinformatics, vol. 21, no. 3, pp. 1047-1057, 2020.

[18] UDR Khaire, "Stability of feature selection algorithm: A review," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 4, pp. 1060-1073, 2022.

[19] A. Bommert, T. Welchowski, M. Schmid and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," Briefings in Bioinformatics, vol. 23, no. 1, p. 354, 2022.

[20] RJG Kohavi, "Wrappers for feature subset selection," Artificial intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.

[21] T. Lal, O. Chapelle, J. Weston and A. Elisseeff, "Embedded methods," Feature Extraction: Foundations and Applications, pp. 137-165, 2006.

[22] P. Minkiewicz, A. Iwaniak and M. Darewicz, "BIOPEP-UWM Virtual—A Novel Database of Food-Derived Peptides with In Silico-Predicted Biological Activity," Applied Sciences, vol. 14, no. 7204, p. 12, 2022.

[23] G. Wang, X. Li and Z. Wang, "APD3: the antimicrobial peptide database as a tool for research and education," Nucleic acids research, vol. D1, no. D1087-D1093, p. 44, 2016.

[24] R. Kumar, K. Chaudhary, M. Sharma, G. Nagpal, J. Chauhan, S. Singh, A. Gautam and G. Raghava, "AHTPDB: a comprehensive platform for analysis and presentation of antihypertensive peptides," Nucleic acids research , vol. 43, no. D1, pp. D956-D962, 2015.

[25] PubMed, "National Library of Medicine," 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/. [Accessed 31. 12. 2023].

[26] DS Moore, "Amino acid and peptide net charges: A simple computational procedure," Biochemical Education, vol. 13, no. 1, pp. 10-11, 1985.

[28] A. Ikai, "Thermostability and aliphatic index of globular proteins," The Journal of Biochemistry, vol. 88, no. 6, pp. 1895-1898, 1980.

[29] HG Boman, "Antibacterial peptides: basic facts and emerging concepts," Journal of internal medicine, vol. 254, no. 3, pp. 197-215, 2003.

[30] A. Radzicka and R. Wolfenden, "Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution," Biochemistry, vol. 27, no. 5, pp. 1664-1670, 1988.