ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ"

ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА

КАТЕДРА „КОМПЮТЪРНА ИНФОРМАТИКА"

Желязко Петров Терзийски

Използване на техники от изкуствения интелект за анализ и прогнозиране свойствата на пептиди

Автореферат

на дисертационен труд за присъждане на образователна и научна степен „Доктор"

по област на висшето образование 4. Природни науки, математика и информатика

в професионално направление 4.6. Информатика и компютърни науки

докторска програма „Информатика"

Научен ръководител: **доц. д-р Станка Хаджиколева**

Пловдив, 2024 г.

Дисертационният труд е обсъден и насочен за защита от катедрения съвет на катедра „Компютърна информатика“ на Факултета по математика и информатика при Пловдивския университет „Паисий Хилендарски“ на 16.01.2024 г.

Дисертационният труд съдържа 189 страници. Библиографията включва 150 източника. Броят на авторските публикации по темата на дисертационния труд е 5.

Защитата на дисертационния труд ще се състои на 11.03.2024 г. от …… ч. в Заседателна зала, Нова сграда на Пловдивския университет, бул. „България“ № 236.

Материалите по защитата са на разположение на интересуващите се в Деканата на ФМИ – каб. 330 в Нова сграда на ПУ „Паисий Хилендарски“, всеки работен ден от 8:30 до 17:00 часа.

Автор: Желязко Петров Терзийски

Заглавие: Използване на техники от изкуствения интелект за анализ и прогнозиране свойствата на пептиди

Пловдив, 2024 г.

Съдържание

ИЗПОЛЗВАНИ СЪКРАЩЕНИЯ 4
УВОД 5
ГЛАВА 1. ПЕПТИДИ. ИЗПОЛЗВАНЕ НА ИИ ЗА ПРОГНОЗИРАНЕ СВОЙСТВАТА НА ПЕПТИДИ .6
- **Използване на ИИ при прогнозиране свойствата на пептиди** 7
- **Процес на прогнозиране на биологичната активност на пептиди** 7
- **Изводи** 10

ГЛАВА 2. МОДЕЛ НА СОФТУЕРНО ПРИЛОЖЕНИЕ ЗА АНАЛИЗ И ПРОГНОЗИРАНЕ НА СВОЙСТВАТА НА ПЕПТИДИ ЧРЕЗ ИИ 10
- **Модул за съхранение и обновяване на набори от пептиди** 11
- **Модул за извличане на физикохимични характеристики и кодиране** 12
- **Модул за прогнозиране на биологична активност чрез ИИ** 13
- **Модул за входни данни** 15
- **Модул за изходни данни** 16
- **Изводи** 16

ГЛАВА 3. СОФТУЕРНО ПРИЛОЖЕНИЕ ЗА АНАЛИЗ И ПРОГНОЗИРАНЕ СВОЙСТВАТА НА ПЕПТИДИ ЧРЕЗ ИИ 16
- **Използвани софтуерни технологии и инструменти** 16
- **Проектиране на прототип на софтуерно приложение** 17
- **Реализация на прототипа** 18
- **Изводи** 20

ГЛАВА 4. ЕКСПЕРИМЕНТАЛНИ РЕЗУЛТАТИ 20
- **Структурна идентификация на моделите от ИИ** 21
- **Изследване влиянието на метода за избор на характеристики** 21
- **Изследване чрез статистически данни от БД** 23
- **Прогнозиране на биологична активност** 25
- **Изводи** 27

ЗАКЛЮЧЕНИЕ 27
- **Приноси** 28
- **Перспективи за бъдещо развитие** 28
- **Апробация** 29
- **Списък с публикации по темата на дисертационния труд** 29
- **Забелязани цитирания** 30
- **Участие в научноизследователски проекти** 30

ИЗПОЛЗВАНА ЛИТЕРАТУРА 31

Използвани съкращения

AAC	-	Amino Acid Composition Аминокиселинен състав
БАВ	-	Биологично активни вещества
БД	-	бази данни
ИИ	-	Изкуствен интелект
НМ	-	Невронна мрежа
AUC	-	Площ под ROC крива
FFNN	-	Feed Forward Neural Network
pI	-	Isoelectric point Изоелектрична точка
RF	-	Random forest
SVM	-	Support vector machines
QSAR	-	Quantitative structure property relationships

Увод

Пептидите са химически вещества с естествен или синтетичен произход, които се състоят от аминокиселини. Веригата им включва от 2 до около 50 аминокиселини. При наличие на повече от 50 аминокиселини, те се трансформират в протеини.

Пептидите притежават широк спектър от биологични свойства, оказващи влияние върху човешкия организъм. Използват се както при лечение на болести, така и като хранителни добавки или козметични средства. Първото пептидно лекарство е инсулина, открит още през 1921 г.

Интерес за медицината представляват пептидите, които имат положителен ефект върху човешкото здраве. Влиянието им се характеризира с конкретна насоченост. Това е т.нар. биологична активност, според която пептидите биват противоракови, антимикробни, за нормализиране на кръвното налягане, антиоксидантни, противодиабетни и др.

Свойствата на пептидите зависят пряко от тяхната структура. Методът, базиран на последователността, е един от най-използваните методи за изследване на количествена връзка структура-активност (QSAR - Quantitative Structure Property Relationships). Той може да се използва за in silico подбор на най-добрите кандидат-пептиди преди техния реален синтез. С оглед на това, тази дисертация е насочена към прогнозиране на биологичната активност на пептиди по метода QSAR, с помощта на методи на изкуствения интелект (ИИ).

Основна цел на дисертационното изследване **е да се проектира, разработи и апробира софтуерна система за прогнозиране на биологичните свойства на пептидите с помощта на различни методи на изкуствения интелект.**

Изпълнението на поставената цел предполага решаване на следните задачи:

Задача 1. Проучване на същността на пептидите, актуалните изследвания и научните достижения за използване на методи на изкуствения интелект за прогнозиране на биологичната им активност;

Задача 2. Проучване на публично достъпните бази данни и инструменти за кодиране на пептиди и създаване на база данни с известни пептиди и тяхната биологична активност;

Задача 3. Създаване на концептуален модел на софтуерно приложение за анализ и прогнозиране на свойствата на пептиди чрез различни методи на ИИ;

Задача 4. Разработване на софтуерен прототип, включващ база данни, модул за извличане на физикохимични характеристики на пептиди и модул за прогнозиране на биологична активност на пептиди, използващ различни методи на изкуствения интелект;

Задача 5. Тестване на софтуерния прототип и анализ на резултатите от проведените експерименти.

Структурата на дисертационният труд следва проведените изследвания, разработки и експерименти и включва Увод, 4 глави, Заключение, и три приложения, съдържащи описание на проектираната база данни за пептиди, статистическа информация за характеристиките на пептидите и изходни резултати от проведени експерименти за изчисляване на физикохимичните храктеристики на пептид.

В **Глава 1** е направено проучване на съвременни подходи и изследвания за прогнозиране на свойствата на пептидите с помощта на различни методи на ИИ. Представен е класическия алгоритъм за in silico прогнозиране на биологична им активност.

Глава 2 представя концептуален модел на софтуерно приложение за анализ на физикохимичните свойства и прогнозиране на биологичната активност на пептиди.

Софтуерно приложение Pep Lab е представено в **Глава 3**. То реализира описаните в модела функционалности и процеси.

В **глава 4** са представени резултатите от проведените с Pep Lab експерименти. Те включват генерирани различни набори от данни и прогнозиране на биологичната активност на пептиди чрез няколко модели от ИИ.

В заключението са обобщени резултатите от изпълнението на поставените задачи. Описани са научни, научно-приложни и приложни приноси, резултат от дисертационното изследване, както и направления за бъдещо развитие.

Списъкът на използваната литература включва **150** заглавия, от които **19** интернет-източника.

Резултатите от изследването са представени в **5 публикации**. **Две от тях са индексирани в международните бази от данни Scopus и Web of Science, и 1 е публикувана в издание с импакт фактор**.

Участието в **6 научно-изследователски проекта**, свързани с изкуствен интелект или изследването на пептиди даде допълнителни насоки за успешната реализация на дисертационното изследване.

Благодарности

Изказвам благодарности към научния ми ръководител доц. д-р Станка Хаджиколева, както и на проф. д-р Емил Хаджиколев за помощта, подкрепата, идеите, насоките и препоръките при провеждане на дисертационното изследване, както и възможността да участвам в научно-изследователски проекти. Искрени благодарности отправям и към доц. д-р Маргарита Терзийска и гл. ас. Ивелина Десева от УХТ-Пловдив за помощта, насоките и препоръките в областта на ИИ и биоинформатиката.

Глава 1. Пептиди. Използване на ИИ за прогнозиране свойствата на пептиди

Пептидите са вещества, чиито молекули са изградени от верига от 20-те основни аминокиселини, свързани чрез пептидни връзки. Най-често пептидната последователност се състои от 2 до 50 аминокиселини и поради това молекулната им маса е малка, обикновено е под 10kDa. За начало на разработването на пептидни медикаменти се приема 1921 г., когато е открит инсулина, първото пептидно лекарство.

Пептидите играят изключително разнообразна роля в човешкия организъм. Те вземат активно участие в растежа на тялото, поддържат работата на имунната система, използват се като антиинфекциозни средства и т.н. [1]. С други думи, техният ефект върху човешкото здраве е с конкретна насоченост [2].

Базираната на пептиди терапия преживява истински бум на световните пазари, като само през 2022 г. продажбите са за над 43 милиарда долара. Очаква се приходите на базирани на пептиди терапевтични средства да достигне 77 милиарда долара до 2032 година [3]. Това огромно търсене на пазара изисква по-гъвкави от традиционните *in vivo* и *in vitro* експериментални подходи за откриване на пептиди. Тези процеси отнемат много време, скъпи и трудоемки са [4].

Преодоляването на тези ограничения изисква прилагането на нови и ефективни изчислителни подходи за откриването и изследването на кандидат-пептиди. В практиката все по-често се използват т. нар. *in silico* методи, които служат като бързи и евтини инструменти за идентифицирането им в обширното пространство на комбинаторната последователност. Те използват знания от различни области като математика, информатика, биология, биотехнологии и аналитична химия за откриване на потенциални кандидати преди техния реален синтез, като по този начин се ускорява откриването на нови полезни пептиди. Изключително перспективен модел при *in silico* подхода е базиран на последователността за намирането на количествена връзка между структурата и активността на пептида (QSAR модел).

Използване на ИИ при прогнозиране свойствата на пептиди

Реализацията на QSAR модела посредством различни техники от МО позволява успешно да се идентифицира структурата на нови пептиди [5]. С различните методи се постига различна точност на прогнозите. Ето защо е актуален въпросът за необходимостта от системна оценка и подобряване на тези алгоритми по отношение на тяхната методология и способност за прогнозиране.

При моделирането на пептидни последователности чрез QSAR, предикторите се състоят от физико-химични характеристики или теоретични молекулярни дескриптори на пептиди, а категорийната стойност е конкретна биологична активност [6]. Моделите първо обобщават предполагаема връзка между химически структури с известна тяхна биологична активност в наличен набор от данни. След това QSAR моделът прогнозира активността на нови пептиди.

Проведеното проучване на научната литература показва, че са разработени множество алгоритми за прогнозиране на биологичната активност на пептиди, основаващи се на различни техники от МО - изкуствени невронни мрежи (НМ) [7], случайна гора (Random Forest - RF) [8], метода на опорните вектори (Support Vector Machines - SVM) [9], и др.

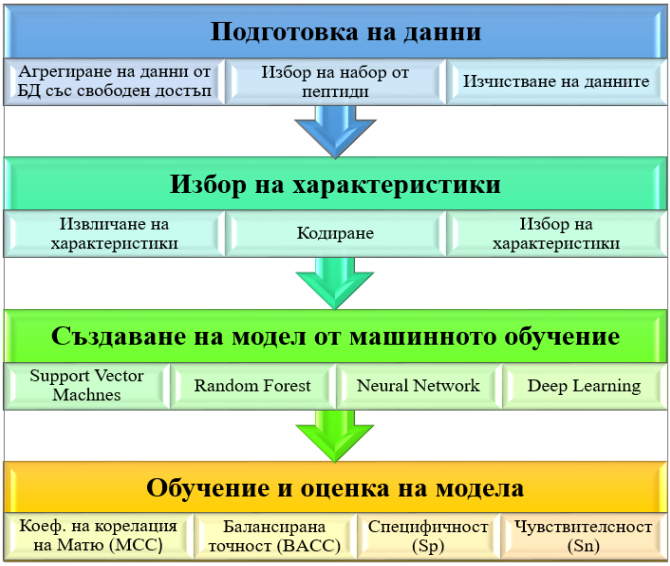
В заключение може да се каже, че МО се счита за най-добрия *in silico* инструмент в биоинформатиката за решаване на проблеми, включващи големи масиви от данни и множество променливи. Причина за това е възможността им да анализират и да се учат от наличните данни, и след като бъдат обучени, да правят относително точни прогнози за нови набори от данни [11].

Процес на прогнозиране на биологичната активност на пептиди

Класическата методологията за разработване на модели за прогнозиране на биологичната активност на пептиди включва 4 етапа [13]. Графично тя е представена на фиг. 1. Етапите са:

1. **Събиране и обработка на данни;**
2. **Избор на характеристики и кодиране;**
3. **Създаване на МО модел;**
4. **Обучение и оценка на модела.**

Трябва да се отбележи, че такава обща схема е широко прилагана и при прогнозиране на функциите в структури на ДНК и протеини [14].



Фигура 1. Методология за прогнозиране на биологичната активност на пептиди

При подготовката на данни за прогнозиране на биологична активност на пептиди следва да се подготвят два класа проби – положителен и отрицателен. Положителни са тези проби, за които експериментално е установено, че притежават конкретна активност. Отрицателни са съответно пептидите, които не притежават тази активност. В научната литература има много малко експериментално потвърдени отрицателни проби. Поради това, формирането на отрицателния набор от данни се прави по един от следните начини:

а) избират се произволни пептиди, с функция, различна от желаната;

б) избират се токсични пептиди;

в) генерират се случайни пептиди.

Тези подходи се основават на хипотезата, според която вероятността така генерираните проби да не са отрицателни, са минимални.

Предварителната обработка на данните е от съществено значение за ефективното използване на техниките от МО. Важно е да се изследват входните характеристики и да се оптимизират, за да се намали броят им. За тази цел, от даден входен вектор с определен брой характеристики се избира само част от тях. Това подмножество включва само характеристиките, които влияят най-силно върху процеса на прогнозиране. За получаване на значимите характеристики се използват различни подходи за избор на характеристики (Feature Selection). **Те имат някои важни предимства: премахва се незначимата информация; опростяват се моделите и стават по-лесни за осмисляне; намалява се значително изчислителната тежест на модела; подобрява се съвместимостта с модела на обучение**.

Основната предпоставка за използване на техника за избор на характеристики е, че данните съдържат някои характеристики, които са или излишни, или неподходящи, и по този начин могат да бъдат премахнати, без това да доведе до голяма загуба на информация. Процесът на извличане създава нови характеристики от оригиналния входен вектор с данни, докато изборът на характеристики връща подмножество на този вектор [15]. Техниката за избор на характеристики обикновено се използва в области, където има сравнително малко проби. Такъв е случаят и при изследването на пептиди.

Изборът на характеристики при оценка на биологичната активност на пептидите има смисъл само в контекста на конкретен модел за прогнозиране.

Таблица 1. Облачни услуги и софтуерни приложения за изчисляване на дескриптори на характеристики

Метод	Интернет адрес
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

Извличане и кодиране на характеристики е дефинирано в [16]. Софтуерни приложения за изчисляване на дескриптори на характеристиките се представени в Таблица 1.

Обучението с учител при МО успешно се използва за решаване на различни проблеми в биоинформатиката [10]. Моделът се обучава с известни входни и изходни данни, след което той може да прогнозира резултата при подаване на неизвестни данни [11]. Моделът, който генерира дискретни класове, се нарича алгоритъм за класификация. Този тип алгоритъм оценява прогнозата за клас, който е 0 (отрицателен) или 1 (положителен).

Моделите от ИИ, използвани за прогнозиране на свойствата на пептиди най-често включват - изкуствени невронни мрежи, SVM, RF и др. [12]. Вместо произволен избор на МО класификатор, е добре да се изследват един или повече класификатори и да се избере най-подходящия.

Преобучение на модела възниква, когато моделът не само се учи от входните сигнали, но също така интегрира ирелевантна инфор-мация за характеристиките от данни за обучението. За да се ограничи преобучението на модела, се използва кръстосано валидиране (cross-validation – CV).

За оценка на ефективността на прогнозирането на различните методи в МО най-често се използват следните показатели за оценка - класификационна точност (CA), специфичност (Sp), чувствителност (Sn), прецизност (Pr), средна хармонична стойност на чувствителност и прецизност (F1) и коефициент на корелация на Матю (MCC). Те се изчисляват с формулите:

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

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

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

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

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

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

където TP и TN е броят на положителните и отрицателни проби, класифицирани правилно от предсказващия модел. FP и FN представляват броя на положителните и отрицателните проби, класифицирани погрешно.

Изводи

Въз основа на направеното проучване на публично достъпните БД и набори от данни за БАП бе установено, че те имат следните недостатъци:

- наличие на дублирана информация за един или повече пептиди в дадена БД;

- някои набори от данни съдържат противоречива или грешна информация;

- в различните БД, даден пептид фигурира с различни биологични активности, но няма обобщена информация, която да дава пълна представа за пептида;

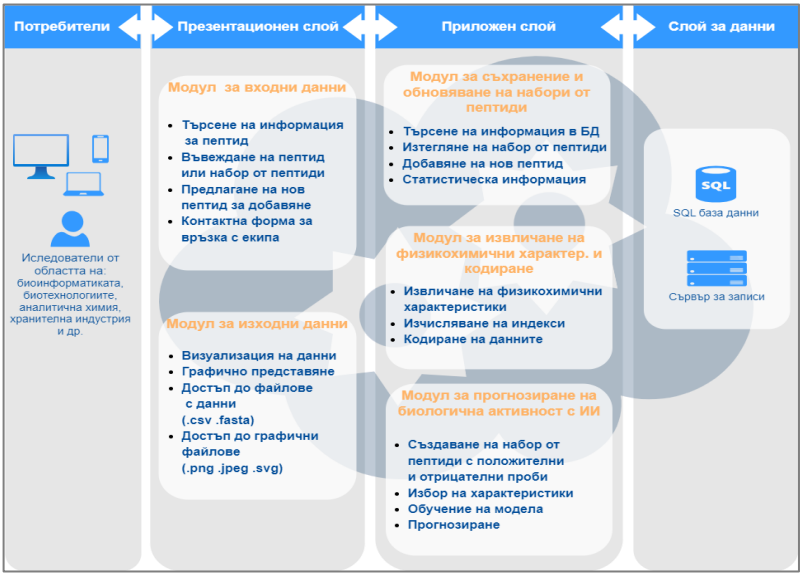
- в някои БД няма достатъчно информация за физикохимичните свойства на пептидите, а в наличната такава понякога има грешки;

- липсва (с малки изключения) възможност за извличане на информация за случайна пептидна последователност, както и представянето на резултата в графичен вид;

- съществуващите приложения на БД с пептиди не предлагат оптимизиране и респонсив дизайн.

За да се преодолеят тези ограничения, е необходимо да се разработи БД за пептиди, която съдържа информация за разнообразни биоактивни пептиди и включва отрицателни проби, което ще помогне за разработване на още по-добри модели от МО.

Глава 2. Модел на софтуерно приложение за анализ и прогнозиране на свойствата на пептиди чрез ИИ



Фигура 2. Архитектура на приложение за прогнозиране на биологична активност на пептиди с ИИ

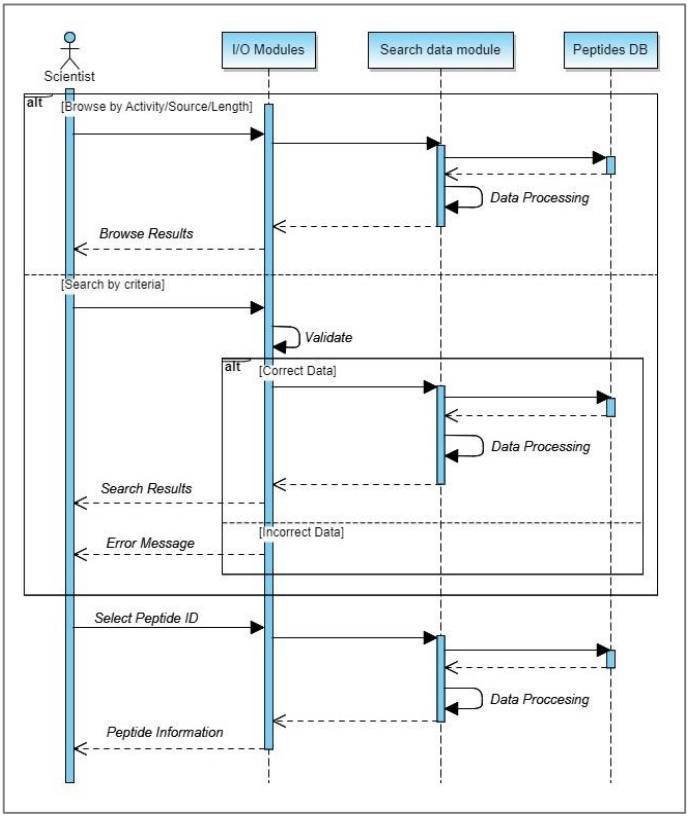
Приложението е изградено на базата на стандартна трислойна архитектура, съдържаща Презентационен слой, Приложен слой и Слой за данни (фиг. 2).

Включва няколко модула:

- **Модул за съхранение и обновяване на набори от пептиди**;
- **Модул за извличане на физикохимични характеристики и кодиране**;
- **Модул за прогнозиране на биологична активност чрез ИИ**;
- **Модул за входни данни**;
- **Модул за изходни данни.**

Модул за съхранение и обновяване на набори от пептиди

Този модул съдържа стандартните за уеб приложение функционалности като търсене на пептид по зададен параметър, визуализация по класификационни критерии и извеждане на подробна информация за даден пептид (фиг. 3). Поддържа възможност за извеждане на статистическа информация за наличните в базата данни пептидни последователности. От съществено значение е БД да съдържа само уникални пептиди, без дублиране, с коректна проверена информация за всеки пептид, с биологични активности, оповестени в научни публикации и експери-ментално проверени стойности.



Фигура 3. Диаграма на последователностите за извеждане на детайлна информация за пептид

Информацията за пептидите може да се визуализира според

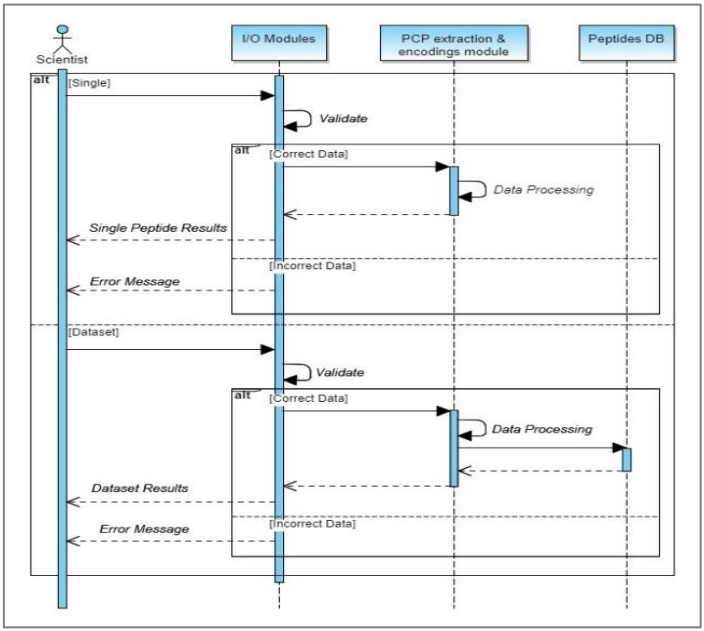
различни класифика-ционни критерии – активност, групи организми, дължина на пептида, идентификатор, амино-киселинна последователност и др.

В представения алгоритъм потребителя има две опции за търсене - Browse и Search. При опцията Browse, наличните пептиди са разделе-ни на три секции Activity, Source и Length, а при Search се избира определен критерии за търсене, за който потребителят въвежда информация. Тя се валидира и ако не е коректна, се връща съответно съобщение за грешка. И двете опции връщат като резултат списък с пептиди (Browse Results или Search Results), от които потребителят може да избере конкретен пептид (чрез ID на пептида) и информацията за него ще се визуализира.

Модул за извличане на физикохимични характеристики и кодиране

На база съдържащите се в пептида аминокиселини, модулът извлича информация за физикохимичните му характеристики и изчислява индекси, показващи определени свойства (фиг. 4). Модулът кодира пептида според най-популярните кодировки.

Аминокиселинен състав (Amino Acid Composition - AAC) е процедура за изчисляване на състава на аминокиселините в дадена пептидна последователност. Той показва честотното разпределение на всяка от 20-те основни аминокиселини.



Фигура 4. Диаграма на последователностите на модула за извличане на физикохимични характеристики и кодиране

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

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

Получава се вектор AAC 1x20, описан с израза (7), всяка стойност AA_i представя процентното съотношение на аминокиселината в пептида, изчислена чрез формула (8).

При **двоичното кодиране** (Binary Profile Features - BPF) на пептид, всяка аминокиселина се описва като единичен вектор с дължина 20, където всяка позиция, с изключение на една, има стойност 0. Крайният резултат е матрица L x 20 (9), където L е броя на аминокиселините, включени в пептида.

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

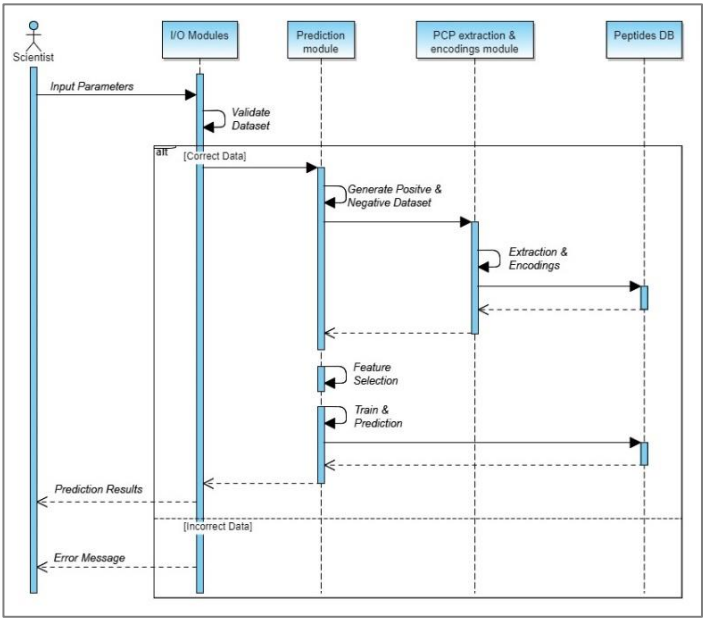
Групираният AAC (Grouped AAC - GAAC) се изчислява като процентното съотношение на различни групи аминокиселини в пептида според свойствата им. В резултат на GAAC се получава 10-размерен вектор (10).

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

Атомният състав (Atomic Composition - ATC) показва честотата на срещане на атомите в пептида. Сумата на всички атоми от съответния вид представят атомната композиция (11).

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

Модул за прогнозиране на биологична активност чрез ИИ



Фигура 5. Диаграма на последователностите на модула за прогнозиране на биологична активност чрез ИИ

Този модул извършва прогнозиране чрез ИИ и включва няколко основни етапа (фиг. 5):

- генериране на кодиран набор от данни с положителни и отрицателни проби;
- извличане на най-важните и информативни характеристики;
- обучение на модел от ИИ;
- тест и валидиране или прогнозиране.

Алгоритъмът на работа на този модул е представен на фиг. 6. Създаването на набор с данни от дадена биологична активност трябва да включва два класа пептиди – положителни и отрицателни. В базата от данни на приложението има пептиди с определен вид активност. Тези пептиди формират

положител-ните проби. Вторият клас проби може да бъде съставен от: *доказани отрицателни проби; пептиди с активност, различна от избраната; случайно генерирани пептиди*. Полученият набор с цифрови стойности се използва за **избор на най-важните и информативни характеристики**. Изборът на характеристики е една от най-важните техники в предварителната обработка на данните и е задължителна част от съвременните алгоритми за машинно обучение [18]. Правилното използване на алгоритми за избор на характеристики оптимизира обучението, като повишава скоростта на обучение и/или намалява сложността на модела. Съществуват три основни модела за избор на характеристики: **филтърният модел (***filter feature selection***)** [19]; **моделът на обвивката (***wrapper feature selection***)** [20]; **вграденият модел** *(embedded feature selection)* [21].

Софтуерното приложение предвижда използване на три модела от ИИ за **прогнозиране на биологична активност** – метода на опорните вектори (SVM), случайна гора (RF), класическа невронна мрежа с право разпространение на сигнала (FFNN).

Методът на опорните вектори е алгоритъм за машинно самообучение, който се използва за класификация и регресия. SVM решава оптимизационна задача с цел намиране на оптимална хиперравнина, която най-добре разделя данните на различните класове. Ако е даден набор от данни за обучение от *n* точки във формата: $(\overrightarrow{x_1}, y_1), \dots, (\overrightarrow{x_n}, y_n)$, където y_i приема стойност 1, или −1. Ако разширим SVM до случаи, в които данните не са линейно разделими, се използва функция на загубите:

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

Тук y_i е *i*-тата цел (т.е. в този случай 1 или −1). $\vec{w} \cdot \overrightarrow{x_i} - b$ е *i*-тият изход, $\vec{w}$ е нормален вектор към хиперравнината, а параметърът b определя отместването на хиперравнината от началото.

Тази функция е нула, ако ограничението в (13) е изпълнено, с други думи, ако $\overrightarrow{x_i}$ лежи от правилната страна на полето.

$$y_i(\vec{w} \cdot \overrightarrow{x_i} - b) \geq 1, \text{ за всяко } 1 \leq i \leq n \qquad (13)$$

Тогава целта на оптимизацията е да се минимизира израза:

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

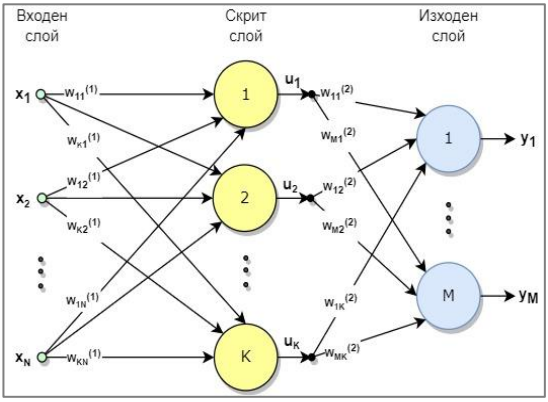
Основното предимство на SVM е ефективността при работа с голямо пространство от данни и възможността за класификация на нови наблюдения, чиито класове не са известни.

Случайна гора е алгоритъм за машинно самообучение основан на концепцията за ансамбли от дървета на решенията. RF комбинира множество дървета на решенията, като всяко от тях е обучено върху различни части от набора с данни. Ако броя на дърветата е B, тогава $b = 1, \dots, B$. От оригиналния набор с данни *D* се избира случайно подмножество *d,* а от всички *p* характеристики се избират *m*, където $m \leq p$. Извършва се генериране на дърво на решенията T_b. Тази операция се повтаря и се получава се ансамбъл от дървета $\{T_b\}_1^B$. Крайният резултат е вземане на решение на базата на гласуване чрез мнозинство:

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

RF е много популярен модел от ИИ, поради което намира различни приложения, като обработка на изображения, анализ на медицински данни и мн. др.

НМ с право разпространение на сигнала (FFNN) се състои от определен брой изкуствени неврони, които са свързани по между си (фиг. 6). Процесът на обучение на невронната мрежа представлява оптимизационна задача за минимизиране на грешката между предсказаните и действителните стойности. Обратно разпространение на грешката е градиентен метод метод за обучение от първи ред, като допуска възможността от попадане в локален екстремум, което се явява негов най-сериозен недостатък.



Фигура 6. НМ с право разпространение на сигнала

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

Целевата функция има вида (16). където M е броя на изходите, y_k изходния сигнал на k-ия неврон, а d_k е очакваната изходна стойност на неврона. Изчислява се производната на целевата функция по отношение на теглата на невроните от изходния слой:

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

Градиента за невроните от скрития слой се определят на същия принцип, само че при тях израза е по-сложен:

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

Определяне на вектора на градиента е много важно за последващия процес на промяна на теглата. При обратно разпространение на грешката тъй като се търси минимум, се задава посока на отрицателен градиент, поради което се изчислява:

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

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

където η – е скоростта на обучение, а изразът (20) е правилото, по което се актуализарат тегловните коефициенти в невронната мрежа. Най-често обучението на мрежата се прекратява, ако стойността на градиента падне под зададен праг ε, характеризиращ точността на процеса на обучение.

Избирайки някой от моделите от ИИ се реализира крайната цел на този модул, а именно прогнозиране на биологичната активност на даден пептид.

Модул за входни данни

Този модул има функцията да приеме въведените от потребителя данни и да ги предаде за обработка към модула, за който са предназначени. Информация се въвежда при използване на

функционалностите за търсене на информация в базата данни, анализ на пептидна последователност или набор от пептиди, прогнозиране на биологична активност, предлагане на нов пептид за добавяне, регистрация на потребител и контактната форма за връзка с екипа.

Модул за изходни данни

Модулът за изходни данни осигурява визуализация на резултатите от обработката в графичен или текстови вид, както и възможност за експортиране на данните в различни графични или таблични формати. Той приема данни от всички модули на слоя за бизнес логика и на базата на тях създава файлове за изтегляне в подходящи за последващи изследвания формати. Визуално резултатите се представят в текстов или графичен вид.

Изводи

В настоящата глава е представен модел на софтуерно приложение за анализ и прогнозиране на свойствата на пептиди чрез ИИ, като е акцентирано върху редица характерни особености на този процес. Разработен е обобщен модел на приложението. Представени са концептуална архитектура с основните модули, както и основните функционалности.

Създаден е модел на Модул за прогнозиране на биологичната активност на пептиди чрез ИИ. Чрез него се извършва избор на характеристики, чрез един от трите метода - филтърен, на обвивката и вграден. Прогнозирането може да се осъществява чрез един от 3 модела от ИИ. Това са SVM, Random Forest и FFNN. Те са избрани, защото представят отделни групи от моделите в ИИ – линеен класификатор, ансамблов модел, базиран на дървета на решенията и класическа невронна мрежа.

На базата на описания модел на софтуерно приложение е реализиран софтуерен прототип, представен в Глава 3.

Глава 3. Софтуерно приложение за анализ и прогнозиране свойствата на пептиди чрез ИИ

Разработването на приложение за анализ на пептиди чрез ИИ е сложен процес, изискващ познания от различни научни области - информатика, биотехнологии, биология, аналитична химия, ИИ и др. Процесът започва с проектиране на база данни и създаване на потребителски интерфейс. Базата данни трябва да е структурирана по начин, спомагащ за изпълнение на поставените в дисертационния труд задачи. В Глава 3 е представена програмната реализация на прототип на софтуерно приложение за анализ и прогнозиране на свойствата на пептиди чрез ИИ, следвайки модела, описан в Глава 2.

Използвани софтуерни технологии и инструменти

Софтуерното приложение ще се използва от изследователи в различни научни области, които могат да използват заедно и поотделно различните модули. За да бъде широко достъпно, приложението е уеб базирано.

За създаване на прототипа е използван пакета **WAMP** (Windows, Apache, MySQL, PHP). **Apache HTTP Server v2.4.51** е безплатен и отворен уеб сървър. **PHP v8.1.0** е един от най-популярните езици за програмиране в интернет. В дисертационният труд се използва и библиотеката с отворен код **BioPHP**.

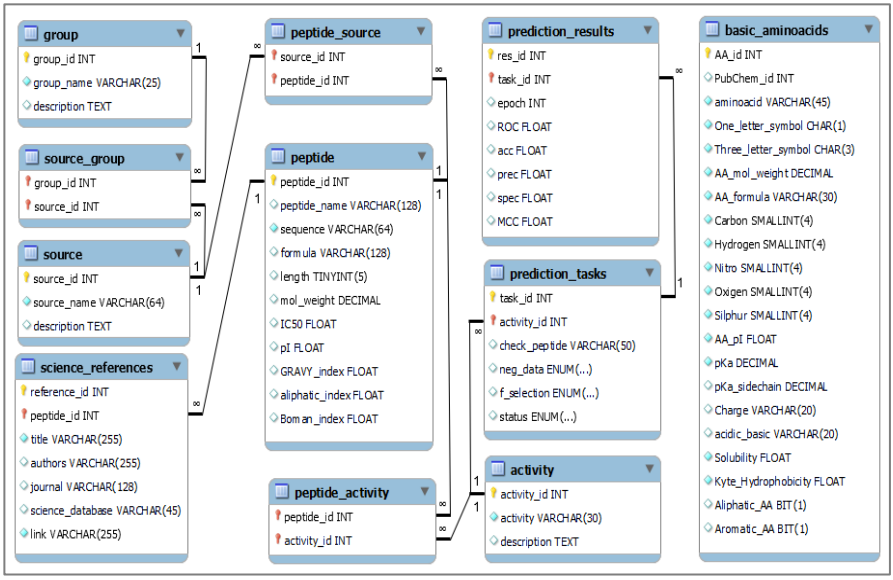
Тя включва класове за анализ на протеини, пептиди, ДНК и други инструменти. За системи за управление на бази данни е избран **MySQL v8.0.27**. Използван е **Bootstrap 5.3.0** framework, базовият за интернет език **HTML v5.0**, **CSS v3** и **JavaScript**. За графична реализиция са използвани библиотеки на **Highcharts**, реализирани с JavaScript. За имплементация на моделите от ИИ е използван езикът **Python v3.11**.

Проектиране на прототип на софтуерно приложение

Потребителският интерфейс е отговорен за изпращането на данни и заявки към слоя за бизнес логика и извеждане на резултата от обработката по подходящ начин. Дизайнът, реализиран с Bootstrap, е в респонсив вариант. За език на приложението е избран английски, за да може да бъде използвано широко от изследователи от цялата международна научна общност.

Приложният слой включва три модула. В зависимост от типа на задачата, тя може да се изпълни само от един модул или с помощта на няколко от тях.

Първият модул е този, чрез който се **въвежда и актуализира информацията за наборите от пептиди**. Той реализира функционалностите, описани в Глава 2. Вторият модул **извлича физикохимични характеристики и индекси и извършва кодиране за единичен пептид или набор от пептиди**. Извличат се и се изчисляват редица характеристики и индекси и се извършва кодиране. Третият модул е за **прогнозиране на биологичната активност на пептида**. Този модул не може да работи самостоятелно, а използва описаните вече два модула: за генериране на набор с положителен клас от проби и за извличане на физикохимични характеристики, изчисляване на индекси и кодиране на създадения набор от данни. Той има следните функционалности: създава набор с пептиди,



Фигура 7. Физически модел на БД на софтуерното приложение

включващ положителен и отрицателен клас; извършва избор на характеристики; прави обучение на избрания модел от ИИ; записва резултатите в базата данни; предава за визуализира резултата.

Базата данни играе ключова роля за работата на приложението. Тя съдържа информация за наличните пептидни последователности, структурата на аминокиселините, стойности по различни скали и свойства, резултати от предсказването на биологичната активност.

Физическият модел на цялостната базата данни е представен на фиг. 7. Той се състои от 11 таблици. Някои основни таблици, обслужващи функционалностите на приложението са: таблица *peptide* съдържа основната информация за пептидите; таблица *basic_aminoacids* съдържа информация за 20-те основни аминокиселини. В нея е включена химическа формула, броя атоми на съставляващите я химически елементи, константа на киселинна дисоциация, индекс на хидрофобност според скалата на Кайт-Дулитъл и др.; таблица *activity* съдържа информация за видовете биологична активност в базата данни. Подробна информация за всички таблици от БД е представена в Приложение 1 на дисертационния труд.

Реализация на прототипа

Този етап се състои в създаване на дизайна, навигацията, описание на процесите, включващо конкретни параметри и създаване на програмен код, използвайки избраните софтуерни технологии и инструменти, за получаване на работещ прототип.

Модулът за съхранение и обновяване на набори от пептиди е наречен **Pep Lab**. Той отговаря за извеждането на информация във всички секции, с изключение на "Tools" (фиг. 8). Модулът за извличане на физикохимични характери-стики и кодиране (**DMPep)**, заедно с модула за прогнозиране на биологична активност чрез ИИ (**PepAI)**, обслужват процесите в "Tools".

Фигура 8. Лента за навигация на софтурното приложение

В **Pep Lab** са въведени само уникални пептиди, без дублиране, които са преминали процес на валидация. Към 31.12.2023г. в нея има **2775 пептида**. Наличната информация за пептидни последователности в Pep Lab е достъпна на адрес *pep-lab.info*. Секция "Database" предлага две опции - търсене по избираеми критерии и търсене по параметър - активност, група или дължина.

Данните в Pep Pab са извлечени от няколко публични бази данни - BIOPEP-UWM [22], APD3 [23] и AHTPDB [24]. Източник на информация за нови пептиди са научните публикации в National Library of Medicine към National Center for Biotechnology Information на САЩ [25], публикции, индексирани в Scopus и Web of science по ключови думи – bioactive peptides, food-derived bioactive peptides, novel bioactive peptides и др.

Секцията „Statistics" показва обобщена информация за наличните в базата данни пептиди, а „Downloads" предоставя файлове за изтегляне за всяка активност в .csv и .fasta формат.

Аналитичният модул (**DMPep**) предоставя комплексен анализ на физикохимични характеристики на пептидна последователност. Използвайки аминокиселинна последователност, модулът може да предостави информация за редица характеристики:

Дължина на пептида се определя от броят на аминокиселините, които се съдържат в пептида.

Молекулна маса (MW) е сумата от масите на всички атоми в пептида. Тя се изчислява като се сумират молекулните тегла на съответните аминокиселини участващи в пептида, като се изваждат теглата на отделящите се молекули, при създаване на пептидните връзки (21).

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

Нетен заряд е сумата на зарядите на всички аминокиселини и крайните терминали. Нетният заряд може да се изчисли по формула (22), която е вариант на уравнението на Henderson–Hasselbalch, представен от Moore [26].

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

Чрез формули (23) и (24) се изчислява, съответно Q^- и Q^+:

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

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

В тези изрази pK_a отразява константата на киселинна дисоциация K_a като отрицателен десетичен логаритъм.

Изоелектричната точка (pI) е стойността на pH, в която нетният заряд е равен на нула т.е. пептида е електрически неутрален. DMpep изчислява нетния заряд при pH от 2 до 14.

Индекс на хидрофобност (GRAVY) се изчислява като сумата от хидрофобността на всяка аминокиселина, разделена на броя аминокиселини в пептида (25).

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

Алифатният индекс (AI) е мярка за съдържанието на алифатични аминокиселини в пептид или протеин. Индексът на алифатност се изчислява чрез формула (26), предложена от Ikai [28]. Коефициентите a и b са съответно относителната стойност на страничната верига на Valine и Isoleucine/Leucine.

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

***Индексът на киселинност* (ABI)** на пептида се определя по формула (27), където ABI_A и ABI_B са молните проценти на аминокиселините с киселинни и основни свойства в пептида.

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

Индексът за свързващия потенциал на пептида (индекс на Боман) [29] показва свързващия потенциал на пептида. Индексът на Боман се определя чрез формула (28), в която S_i е разтворимостта според скалата на Радзичка и Волфенден [30].

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

Модулът за прогнозиране на биологична активност с ИИ (PepAI), следвайки алгоритъма, описан в Глава 2, извърша следните основни действия:

1. Създава се набор от пептиди, съдържащ два класа – положителен и отрицателен. Информацията от набора се кодира в цифров вид и се нормализира.
2. Ако е избран метода за избор на характеристики, се определя оптималното подмножество.
3. Обучава се избрания модел от изкуствения интелект.
4. Прогнозира се дали зададения пептид е от избрания клас активност.
5. Резултатът се предава за визуализация.

Изводи

В Глава 3 е представена реализацията на прототип на софтуерно приложение за анализ и прогнозиране на свойствата на пептиди чрез ИИ. Описани са етапите на проектиране, използваните софтуерни технологии и инструменти. Представени са структурата на базата данни и функционалностите на всички модули.

Реализиран е прототип на приложението, включващ проектираните модули. Създадена е база данни с пептиди Pep Lab, съдържаща пептиди с известна биологична активност. Към 31.12.2023 г. в нея има **2775 пептида**. Pep Lab предлага възможност за представяне на пълните данни за всеки отделен пептид, възможност за изтегляне на пептидите по активности и статистическа информация за базата данни.

Реализираният аналитичен модул (DMPep) предоставя анализ на физикохимичните характеристики на пептиди и кодиране на пептидите според шест варианта: PCP, BPF, AAC, GAAC, DPC и ATC.

Модулът PepAI извършва прогнозиране на биологичната активност на произволен пептид, предоставяйки възможност за избор на различни набори от данни, различни видове кодиране на тези данни, както и избор на един от три модела от ИИ – SVM, Random Forest и HM. Описан е математическия апарат на всеки от моделите, както и програмната му реализация.

Глава 4. Експериментални резултати

Тази част от дисертационния труд представя получените експериментални резултати. Реализираните SVM, RF и FFNN модели са тествани при различни експерименти. За целта са генерирани четири набора от данни, включващи различен брой положителни проби: DS1, DS2, DS3, DS4 (Таблица 2). DS4 е получен като набора

Таблица 2. Набори с данни, използвани в дисертационното изследване

Име	***Положителен набор***	***Негативен набор***	***Общо***
DS1	71 невропептида	71 случайни проби от БД	142
DS2	232 противодиабетни пептида	232 случайни проби от БД	464
DS3	523 пептиди, пониж. кръвно налягане	523 случайни проби от БД	1046
DS4	1333 пептида пониж. кръвно налягане	1333 случайни проби от БД	2666

от данни с 523 пептида е допълнен с 810 пептида, взети от база данни AHTPD. Чрез тях са извършени следните експерименти:

- Чрез набор DS3 са тествани метриките на моделите, при промяна на структурните параметри;
- Изследвано е влиянието на метода за избор на характеристики върху модела;
- Изследвано е влиянието на различните набори върху параметрите на модела.

Използвани са следните метрики за оценка на моделите – площ под кривата на работната характеристики, точност на класификацията, прецизност, специфичност, претеглена хармонична средна стойност на прецизност.

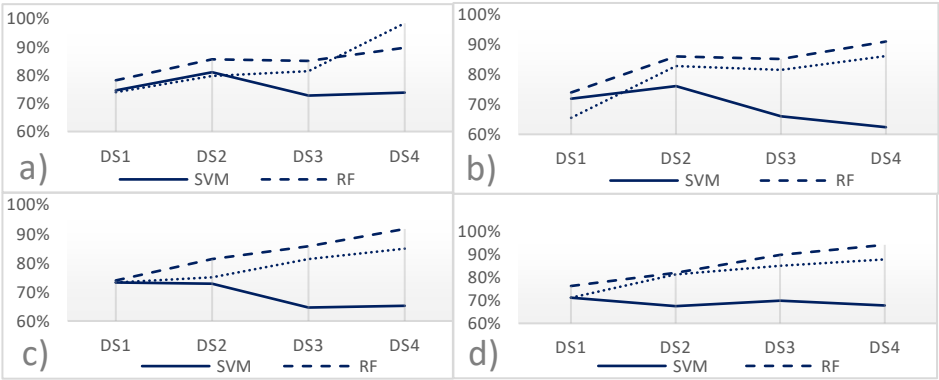
Структурна идентификация на моделите от ИИ

Тази експериментална постановка цели да се настроят структурните параметри на моделите, така че те да имат най-добри показатели при класификация на пептиди. При **SVM модела** е изследвано влиянието на различни видове ядра върху метриките. Тестовете са правени при параметър на регуляризация C=1 и толеранс 0,001. Резултатите от експеримента показват, че най-добри резултати се получават при SVM с **RBF ядро**. При **RF модела** се променя броя на дърветата и се наблюдава ефективността на модела. Моделът има най-добри показатели при **200 дървета,**. За да се определи архитектурата на **FFNN модела** са реализирани голям брой експерименти, като се променят броя на невроните от 5 до 100 със стъпка 5 и се проверяват показателите при 1, 2, 3 и 4 скрити слоя. Резултатите показват, че най-точно класифицира FFNN с **1 скрит слой с 10 неврона**.

От експерименталните резултати може да се направи извода, че *използването на НМ е подходящо при наличие на голям обем входни данни. В конкретния случай за класифициране на БАП, НМ следва да се прилага при наличие на поне 1000 положителни проби.*

Изследване влиянието на метода за избор на характеристики

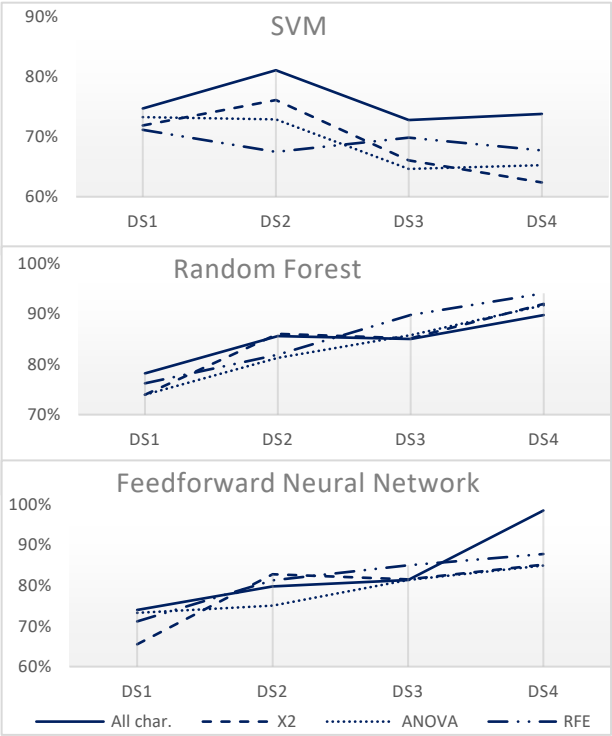
При тази група от експерименти, целта е да се изследва влиянието на различните варианти на избор на характеристики върху степента на изменение на изходните метрики.



Фигура 9. Изменение на точността на моделите при: a) без метод за избор на характеристики; b) χ^2; c) ANOVA; d) RFE

Всеки пептид в набора от данни съдържа вектор от 41 характеристики според следните кодировки: PCP (6 хар.), ATC (5 хар.), AAC (20 хар.), GAAC (10 хар.). Направени са изследвания при четири варианта – без избор на характеристики и 20-те най-добри характеристики селектирани, чрез χ^2, ANOVA и RFE. Графичното представяне на резултатите са показани на фиг. 9.

От представените графики се вижда, че изборът на характеристики оказва най-съществено влияние върху показателите на RF модела. Точността на класификация при този модел се увеличава при използване на избор на характеристики, като най-добра е при RFE. Използването на метод за избор на характеристики предизвиква понижаване на точността на FFNN модела, но пък води до намаляване на броя на входовете и съответно на изчислителната тежест. При този модел, изследователят трябва да прецени дали да използва избор на характеристики. При SVM модела резултатът се влошава спрямо този, без метод за избор на характеристики, така че е по-добре при него да не се прилагат тази техника.



Фигура 10. Изменение на точността на SVM, RF и FFNN взависимост от метода за избор на характеристики

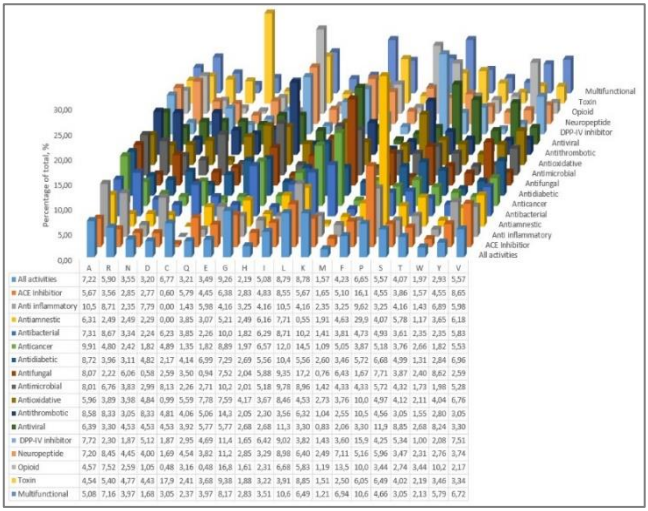
Подобни изводи могат да се направят и при сравнение на методите, по които се осъществява избор на характеристиките при различните модели(фиг. 10).

От тези графики става ясно, че SVM се представя най-добре без метод за избор на характеристики, при FFNN е добре да се използва RFE или χ^2, но само при набори с данни под 500 положителни проби. При повече от 1000 положителни проби, FFNN модела е напълно способен да определи най-важните характеристики и не е необходимо да се използва какъвто и да е метод. При RF модела, метода на RFE е най-добър, но също е важна големината на набора с данни.

Като цяло изследването показва, че при SVM не трябва да се използва избор на характеристики, при НМ може да се използват според преценката на изследователя, а при RF е задължително използването на тази техника.

Изследване чрез статистически данни от БД

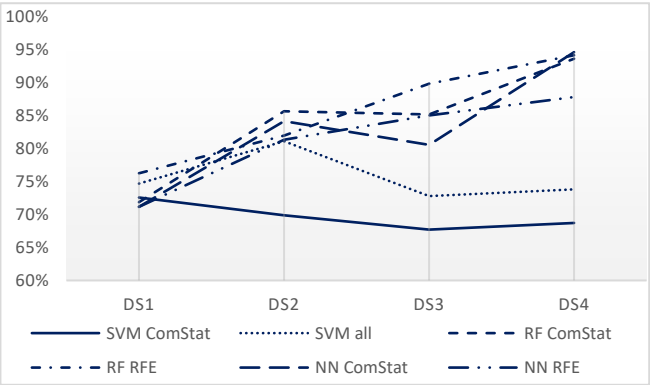
Статистическите данни от наличните в БД пептиди показват разпределение на аминокиселините в пептидите за всяка активност (фиг. 11). Вижда се, че най-често в пептидите се срещат аминокиселините: аланин (A), аргинин (R), цистеин (C), глицин (G), изолевцин (I), левцин (L), лизин (K), пролин (P), серин (S), валин (V). Това са аминокиселините, които имат **над 5% участие, като относителна честота на срещане на пептидите** в базата данни. Прави впечатление, че има много голям процент цистеин (C) в токсичните пептиди – 17,9%, и следователно той може да бъде разглеждан като показател за токсичност.



	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	3,20	6,77	3,21	3,49	9,26	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 inhibitior	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	5,98	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	3,14	2,24	6,23	3,85	2,26	10,0	1,82	6,29	8,71	10,2	1,41	3,81	4,73	4,93	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

Фигура 11. Честотно разпределение на аминокиселините общо и по активности.

Като се има предвид тази информация е предложен нов метод за избор на характеристики, наречен **ComStat** (Complex Statistics). При него за най-съществени характеристики се приемат посочените аминокиселини и физикохимичните характеристи-ки (дължина, изоелектрична точка, молекулна маса и трите индекса).



Фигура 12. Точност при избор на характеристики чрез статистически данни за четири набора с данни.

След корелационен анализ между характеристиките става ясно, че има висока степен на корелация между дължина и молекулна маса. Това показва, че едната от тях е излишна, освен това наличието на две характерстики с голяма корелация може да доведе до преобучение на модела. Избрана е молекулната маса да бъде премахната като излишна характеристика в

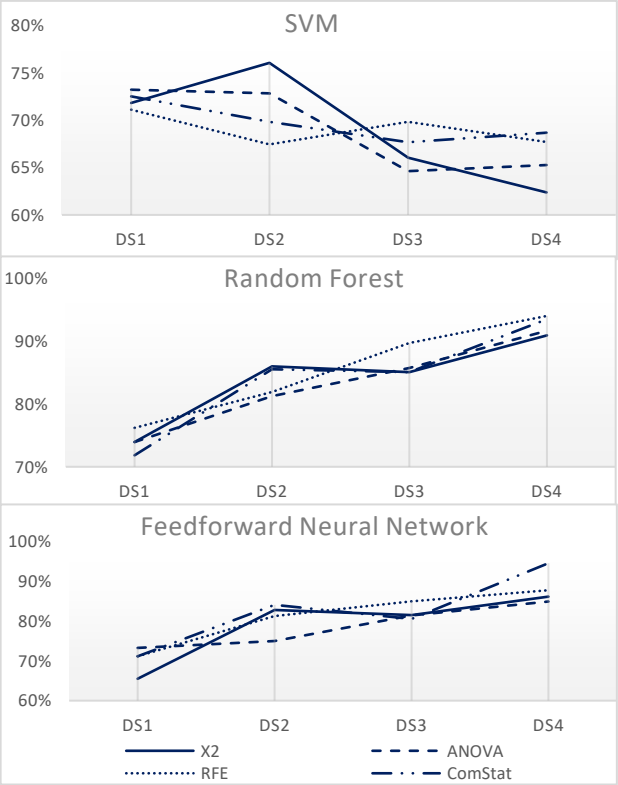
набора. С така подбрания набор от характеристики е извършено прогнозиране с трите модела от ИИ.

Резултатите от прилагането на ComStat метода са представени на фиг. 12. От графиките се вижда, че при SVM модела има значителна разлика, като графиката без селектиране е значително по-добра от тази след селектиране по метода ComStat. При другите два модела обаче графиките показват, че с и без селекция на характеристики, стойностите са близки, като при невронната мрежа даже има по-добри резултати. При двата по-големи набора с данни, резултатите са идентични.

Това показва, че предложения метод за редукция на входните характеристики ComStat може да се използва при RF и HM за набор с над 500 пептида без загуба на точност.

След добавяне на резултатите, получени по метода ComStat към графиките от фиг. 10, се получват сравнителни графики между всички описани методи за избор на храктеристики (фиг. 13).

Графиките показват, че метода ComStat е съизмерим с другите разгледани методи за избор на храктеристики.



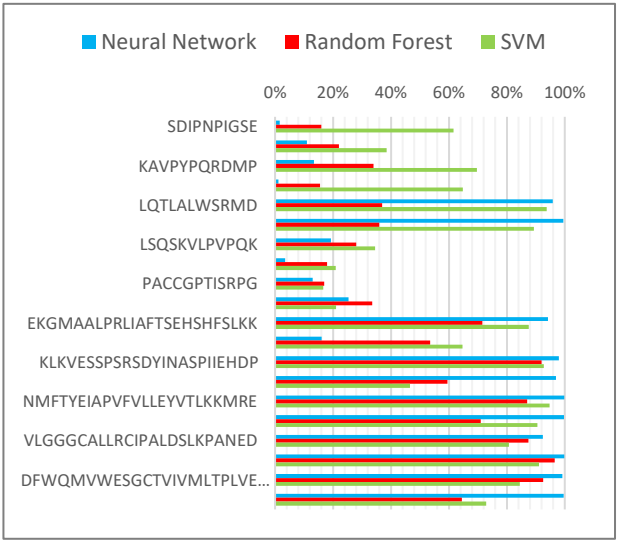
Фигура 13. Сравнение на точността на предсказване по метода ComStat и X^2, ANOVA и RFE

В заключение могат да се направят няколко изводи при прилагане на метода ComStat. Добре е да се използва като алгоритъм за избор на характеристики, при RF и HM, когато набора е над 500 положителни проби. При по-малки набори от проби той може да се приложи успешно за SVM модела.

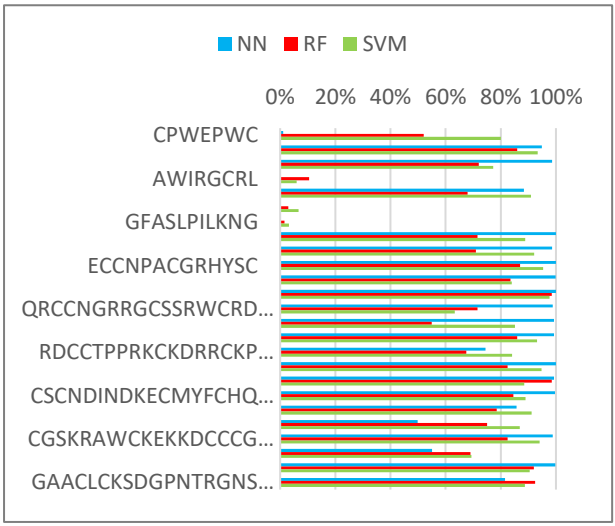
Предимство на метода е, че статистическия анализ се извършва предварително с наличните пептиди в БД и след това се използват избраните характеристики. Това води до значително по-малка изчислителна тежест в процеса на прогнозиране, съответно повишено бързодействие.

Трябва да се отбележи, че с предложения метод е експериментирано само с реализираната в рамките на дисертационното изследване БД с пептиди.

Прогнозиране на биологична активност



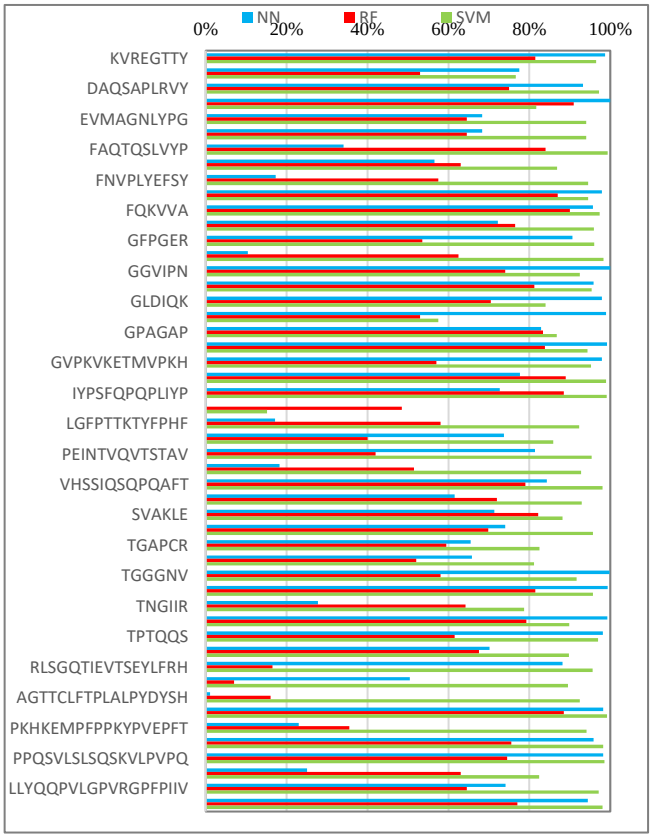
Фигура 14. Прогнозиране на противодиабетни пептиди



Фигура 15. Прогнозиране на понижаващи кръвното налягане пептиди.

За изследване на способността на моделите от ИИ да прогнозират е направен експеримент със следната постановка: избрани са наличните пептиди от три активности в БД – противодиабетни, понижаващи кръвното налягане и токсини за положителни проби. По такъв начин са получени три набора данни, по един за всяка активност, от които 10% са отделени за независим тест. С останалите проби са обучени с SVM, RF, FFNN моделите.

Според получените вече резултати, за SVM не е използван метод за избор на характеристики, а за RF и FFNN е приложен метода RFE. В наборите от данни има по 41 характеристики за пептид. Всеки от моделите е прогнозирал вероятността за принадлежността на пептидите от независимия набор към съответна активност. Резултатите от експеримента са представени в графичен вид на фиг. 14. От графичното представяне се вижда, че SVM се справя най-добре при прогнозирането, особено при пептиди с малка дължина. Това се дължи на малкия набор от данни (232 пептида), при който останалите два модела не могат да се обучат добре. Тяхната прогноза се по-добрява при пептиди с по-голяма дължина - над 20 аминокиселини. Прогнозните



Фигура 16. Прогнозиране на токсични пептиди

резултати са не са достатъчно точни. Ако се приеме, че пептид е разпознат при вероятност над 0.50, то SVM разпознава 75% от пептидите, FFNN – 65%, а RF едва 50%. Интересен факт е, че когато FFNN правилно класифицира даден пептид, го прави много категорично, с вероятност над 90%. FFNN показва много добро прогнозиране при пептидите с дължина над 20 аминокиселини - не е разпознат само един пептид.

За пептиди, понижаващи кръвното налягане, получените резултати са представени на фиг. 15.

При тази активност RF и FFNN се справят много по-добре, което се дължи на двойното увеличение на положителните проби в набора – 523. И трите модела имат висок процент на разпознаване на пробите: FFNN – 82%, RF – 88%, а SVM с над 94% разпознати пептиди. RF прогнозира с по-висока точност от FFNN. Трябва да се отбележи, че прогнозите, направени с RF, са с вероятност около 0.5. Над 40% от познатите пептиди са с вероятност в диапазона 0.50-0.70, което не дава категоричност на прогнозата.

При токсичните пептиди, независимо, че набора с пептиди е малък (250 проби), и трите модела ги разпознават много добре (фиг. 16). Предпоставка за това са техните „противоположни“ свойства (токсичност). FFNN има успешно прогнозиране на 85% от пептидите, защото поради малкия обем на данните не успява добре да се обучи, а RF и SVM разпознават 90% от токсичните пептиди.

Наблюдава се и категорично определяне на активността. При неразпознати проби прогнозираната вероятност е до 0.10, а при разпознатите почти изцяло над 0.70. Това води до извода, че този клас пептиди, поради специфичните си свойства, е лесно разпознаваем за изследваните модели от ИИ.

Изводи

Глава 4 представя експерименталните резултати, получени при класификация на пептиди с реализираните SVM, RF и FFNN модели и четири набора от данни. Установено е, че SVM има най-добри показатели при използване на нелинейно **RBF ядро**, RF при мрежа от **200 дървета**, а НМ с прави връзки трябва да има **1 скрит слой с 10 неврона** и **ReLu** активираща функция.

Изследвано е влиянието на метода за избор на характеристики, като са използвани 4 техники, както и влиянието на големината на набора с данни върху работата на модела. Резултатите показват, че SVM модела не се влияе от големината на набора с данни, а **прилагането на метод за избор на характеристики се отразява отрицателно на резултатите**. Изводът е, че **при малки набори от данни е подходящо да се използва SVM**.

RF е моделът, на който се отразява най-съществено техниката за избор на характеристики. Качеството на прогнозиране се подобрява, като най-добри резултати се постигат при използване на wrapper метода RFE.

Изследванията показват, че успешната класификация с **FFNN силно зависи от големината на набора от данни**. Това дава основание да бъде въведено следното правило: **НМ следва да се използва при набор от данни с поне 1000 пептида. Това се прави с цел да се обучи добре мрежата и да се постигнат максималните показатели**.

Предложен е и нов метод за избор на характеристики **ComStat**, основаващ се на статистически показатели. Като характеристики се използват аминокиселините с най-голяма относителна тежест от наличните в БД пептиди и физикохимичните характеристики. **Този метод подобрява точността на класификация на БАП при RF и НМ моделите за набор от данни с над 500 пептида**.

Резултатите от експериментите с независими набори потвърждават описаните изводи, но добавят и още един - когато трябва да се разпознават пептиди с **дължина до 20 аминокиселини най-добре е да се използва SVM модела**.

Заключение

Настоящият дисертационен труд представлява интердисциплинарно изследване. То е посветено на използването на техники от изкуствения интелект за анализ и прогнозиране на свойствата на пептиди.

От литературния обзор ясно проличава необходимостта от софтуерен продукт, който съчетава интуитивен потребителски интерфейс с надеждна база данни и инструменти за прогнозиране и анализ на пептиди. Всичко това доведе до формулиране на основната цел на дисертационния труд, а именно да се създаде прототип на софтуерна система, която съдържа:

- ✓ база данни, включваща пептиди с известна биологична активност;
- ✓ инструмент, чиято цел е комплексен анализ и извличане на физико-химични характеристики на пептиди и създаване на структури от данни, включващи основни показатели на пептида;
- ✓ създаване на модели, използващ методи от ИИ за прогнозиране на биологичната активност на пептиди.

За реализирането на тази цел са решени следните основни задачи:

1. Да се проучи опита на учените за прогнозиране на биологичната активност на пептиди с помощта на методи от изкуствения интелект.
2. Да се направи проучване на публично достъпните бази данни и инструменти за кодиране на пептиди.
3. Да се проучат методите от изкуствения интелект, които се използват за прогнозиране на биологичната активност на пептиди.
4. Да се разработи прототип на софтуерно приложение, включващо база данни, модул за извличане на физикохимични характеристики на пептиди и модул за прогнозиране на биологична активност на пептиди.
5. Да се създаде база данни от пептиди с доказани в научната литература активност и свойства.
6. Да се създадат, обучат и тестват модели от изкуствения интелект за прогнозиране на биологичната активност на пептиди.
7. Да се анализират получените резултати и да се направи оценка на всеки от моделите.

Приноси

Основните приноси на дисертацията могат да се характеризират като научни, научно-приложни и приложни.

Научни приноси на дисертационното изследване:

1. Разработени модели за прогнозиране на биологичната активност на пептиди – SVM, RF и изкуствени НМ.
2. Разработен нов метод за избор на характеристики ComStat, основаващ се на статистически показатели на пептиди.

Научно-приложни приноси на дисертационното изследване:

3. Разработен концептуален модел на софтуерно приложение за анализ и прогнозиране на свойствата на пептиди чрез различни методи на ИИ.
4. Имплементирани алгоритми за динамично изчисляване на характеристики на пептиди и кодиране на пептиди.
5. Имплементирани алгоритми на изкуствения интелект за прогнозиране на биологичната активност на пептиди, базирани на SVM, RF и изкуствени НМ.

Приложни приноси на дисертационното изследване:

6. Създадена е база данни за пептиди и е въведена информация за 2775 пептида с известни биологични активности.
7. Разработено е софтуерно приложение за анализ и прогнозиране на физикохимичните свойства на пептиди. То е свободно достъпно на адрес: www.pep-lab.info.

Перспективи за бъдещо развитие

В процеса на работа са генерирани нови интересни идеи, които да продължат работата по усъвършенстването на платформата Peplab. Сред тях са:

• Непрекъснато актуализиране на записите в базата данни. Това би позволило извършването на значително по-достоверни статистически анализи;
• Развитие на модула за прогнозиране на БАП с включването на deep learning алгоритми, за които също е необходима БД с повече записи;
• Създаване на модул за новини, чрез който изследователите ще бъдат по-добре информирани за новостите в платформата;
• Внедряване на система за регистрация на потребители с различни роли и с различни нива до достъп;
• Преход към структурна биоинформатика, т.е. включване в базата данни на изображения на пептидни последователности, разпознаването им с deep learning алгоритими и т.н.

Апробация

Основните резултати на изследването са докладвани на национални и международни научни форуми. Резултатите от изследването са представени в 5 публикации – 3 в специализирани списания, 1 сборник от международна конференция и 1 в сборник с трудове от българска конференция. Две от публикациите са индексирани в международни бази от данни Scopus и Web of Science, като едната от тях е публикувана в издание с импакт фактор: 2,7.

Резултати на дисертационното изследване са докладвани на 2 международни и 1 национална конференции, както и на 3 научно-изследователски семинари.

Доклади на конференции:

1. International Conference on Artificial Intelligence and E-Leadership, October, 15-16, 2020, Plovdiv, Bulgaira.
2. 69th Scientific Conference with International Participation "Food science, engineering and technology - 2022", September 29-30, 2022, Plovdiv, Bulgaria.
3. Научна конференция „Дни на науката 2021", 25-27 ноември, Пловдив, България.

Доклади на семинари:

1. „Използване на техники от изкуствения интелект за прогнозиране на биологична активност на пептиди", *Научен семинар "Моделиране на невронни мрежи с приложения в бизнеса",* 13.11.2020 г., гр. Пловдив.
2. „Класификация на аминокиселинни последователности чрез машинно обучение", *XVII Научен семинар на Школа за ИКТ иновации*, 29.11.2021 г., гр. Пловдив.
3. „In silico анализ на физикохимичните характеристики на пептиди", *XIX Научен семинар на Школа за ИКТ иновации*, 23.11.2022 г., гр. Пловдив.

Натрупаните знания по време на обучението в докторантура позволиха, в съавторство с колеги от катедра „Компютърна информатика", издаването на **учебник „Въведение в базите от данни"**. Освен това са разработени **материали за провеждане на занятия по дисциплините „Бази от данни"** и **„Програмиране в интернет с PHP/MySQL"**, които са използвани при провеждането на занятия в ПУ „Паисий Хилендарски" през 2019-2023г.

Списък с публикации по темата на дисертационния труд

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 и 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 & technology 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. **Терзийски, Ж.**, (2021) „Класифициране на антиоксидантни пептиди чрез машинно обучение", Научни трудове на Съюза на учените в България – Пловдив. Серия В. Техника и технологии. Том XIX, ISSN: 1311-9419 (Print); ISSN 2534-9384 (Online).

Забелязани цитирания

Публикация: **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.

Цитирана в: Gaffar, S., Hassan, M. T., Tayara, H., & Chong, K. T. (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.

Публиакция: **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.

Цитирана в: 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.

Участие в научноизследователски проекти

1. *СП19-ФМИ-012*, „Моделиране на невронни мрежи с приложения в бизнеса", финансиран от Фонд „Научни изследвания" при ПУ „П. Хилендарски" (2019-2020).
2. *СП19-ФМИ-004*, „Формиране на научни компетентности в Студентска школа за ИКТ иновации", финансиран от Фонд „Научни изследвания" при ПУ „П. Хилендарски" (2019-2020).
3. *КП-06-М36/2*, "Получаване, характеристика и пречистване на белтъчни ензимни хидролизати от лупина и прогнозиране на биологичната им активност чрез методи от изкуствения интелект", финансиран от Фонд „Научни изследвания" (2019-2022).
4. *МУ21-ФМИ-004*, „Подкрепа за научно изследване в школа за ИКТ иновации", финансиран от Фонд „Научни изследвания" при ПУ „П. Хилендарски" (2021-2022).
5. *08/21-Н*, "Разработване на облачно базиран модул за анализ на пептидни последователности, извлечени от храни", финансиран от фонд "Наука" на УХТ- Пловдив (2021-2022).

6. *МУПД23-ФМИ-021*, „Използване на методи на изкуствения интелект в бизнеса“, финансиран от Фонд „Научни изследвания“ при ПУ „П. Хилендарски“ (2023-2024).

Използвана литература

[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] T. K. 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, L. F. T. 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] U. D. R. 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] R. J. G. 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 researc, 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] D. S. Moore, "Amino acid and peptide net charges: A simple calculational 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] H. G. 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.