Софийски университет “Св. Климент Охридски”
Факултет по математика и информатика

Интелигентни информационни системи в биоинформатиката: семантично интегриране, анализ и класификация на биомедицински данни

Илиян Недков Михайлов

Автореферат на дисертация

за присъждане на образователната и научна степен “доктор”
в професионално направление 4.6 “Информатика и компютърни науки”,
докторска програма „Информационни технологии – Био и медицинска информатика“

Научен ръководител: доц. д-р Димитър Иванов Василев

София, 2021 г.

Където не е указано, всички препратки в текста на автореферата към страници, глави, секции, цитирана литература, таблици и фигури, се отнасят за текста на дисертацията. Всички фигури, таблици и цитирана литература в автореферата използват същата номерация както в текста на дисертацията.

Съдържание

Обща характеристика на дисертацията 4

Структура на дисертационния труд 6

Глава 1: Увод 6

Описание на проблема 6

Сложност на проблема 8

Области и средства на изследванията в дисертацията 8

Глава 2: Теоретични основи и анализ на състоянието по проблемите за интегриране, анализ и класификация на биомедицински данни. 8

Съхранение на данни 9

Трансфер на данни 9

Гъвкаво управление на данни 10

Анализ на данни 10

Архитектура на данните 10

Интеграция, ориентирана към услуги 11

Семантична интеграция 11

Стандарти за биомедицински данни 11

Базирани на софтуер като услуга подходи за интегриране на данни 12

Интегриране на биомедицински данни и откриване на знания 13

Глава 3: Формализация и методи за интелигентно интегриране, анализ и класификация на биомедицински данни. 13

Семантично интегриране на био-медицнски данни. 15

Създадена методология за семантичното интегриране на биомедицински данни от различни заболявания. 17

Основни характеристики и нововъведения при изграждане на методологията за семантично интегриране. 17

Модел за прогнозиране на преживяемост на пациенти болни от рак 20

Използване на машинно обучение за оценка точността на предсказване на протеинови структури 21

Свързани данни на базата на онтологии с цел създаване на съветстваща система в областта на здравеопазването. 22

Предсказване на антимикробна резистентност в метагеномни данни. 23

Компресия на омикс данни 23

Глава 4: Софтуерна реализация на интелигентни системи за интегриране, анализ и класификация на биомедицински данни 24

Софтуерен модул за интегриране на хетерогенни данни и машинно обучение за предвиждане преживяемостта на пациенти болни от рак 25

Софтуерни решения и резултати за предсказване на протеинови структури и оценка на точността . 29

Софтуерни решения и резултати с цел създаване на съветваща система за консултиране на диети при болни от диабет. 30

Реализиран модел и софтуерно решение за интегриране, класификация и анализ на метагеномни данни 31
Реализиран модел и софтуерно решение за компресиране на секвенционни данни 33

Глава 5: Приноси и перспективи **34**
Научни 34
Приложни 35
Перспективи за бъдещо развитие 35
Декларация за оригиналност 36
Публикации по темата на дисертационният труд **36**

Обща характеристика на дисертацията

Съдържание, цели и структура на дисертацията

Дисертационната работа е написана върху 186 страници, в които се включват 50 фигури, 21 таблици, списък на литература, речник на термините, списък на съкращенията, списък с публикации по дисертацията. Използваната литература включва 151 заглавия, списъкът с авторските публикации включва 10 статии.

Глава 1 въвежда в проблема относно интегрирането, анализа и класификацията на биомедицински данни, както и създаването на интелигентни информационни системи, включващи този кръг от проблеми. В тази глава са дефинирани значението и актуалността на работата, както и целите и задачите на дисертацията.

Глава 2 представя обстоен преглед и коментари на теоретичната обосновка и анализ на състоянието на проблемите за интегриране, анализ и класификация на биомедицински данни. Основните акценти са поставени върху подходите за съхранение на данни , методи за интегриране на данни, семантично интегриране на данни, софтуерни аспекти на интеграция на данни ориентирана към услуги, интегриране на биомедицински данни и извличане на знания.

Глава 3 представя методологическите аспекти на подробно описаните задачи на дисертационния труд, като главно внимание се обръща на формализацията и методите за интелигентнно интегриране, анализ и класификация на биомедицински данни. Разглеждата се подробно свойствата на изградената система за семантично интегриране на данни от ракови заболявания, прогнозиране на развитието на заболяването с помощта на методи на машинно обучение. Представени с методите за прогнозиране на протеинови структури и анализ и класификация на антимикроабната резистентност на метагеномни данни. Представени са методи за компресия на секвенционни данни основана на шумозащитеност при кодирането. Представено е и интегриране на данни и методика за създаване на съветваща система базирана на онтологии за предлагане на хранителен режим при болни от диабет.

Глава 4 е посветена главно на резултатите от представените в глава 3 подходи за решаване на поставените задачи в дисертационния труд. Главно внимание в глава 4 е обърнато върху представянето на софтуерните решения за практическо изпозлване на разработените подходи. Особено важно е да се подчертае, че в е представен един холистичен подход за създаване на платформа за предоставяне на софтуер като услуга, използвана за реализация на всички системи в дисертацията.

Глава 5 представя приносите на дисертационния труд, както и възможностите за бъдещо развитие.

Цели и задачи на дисертацията

Основната цел на дисертационния труд е създаването на методология и практическата й реализация за интелигентно интегриране на биомедицински данни и техния анализ, използвайки средства на информатиката, информационните технологии, биоинформатиката и изкуствения интелект.

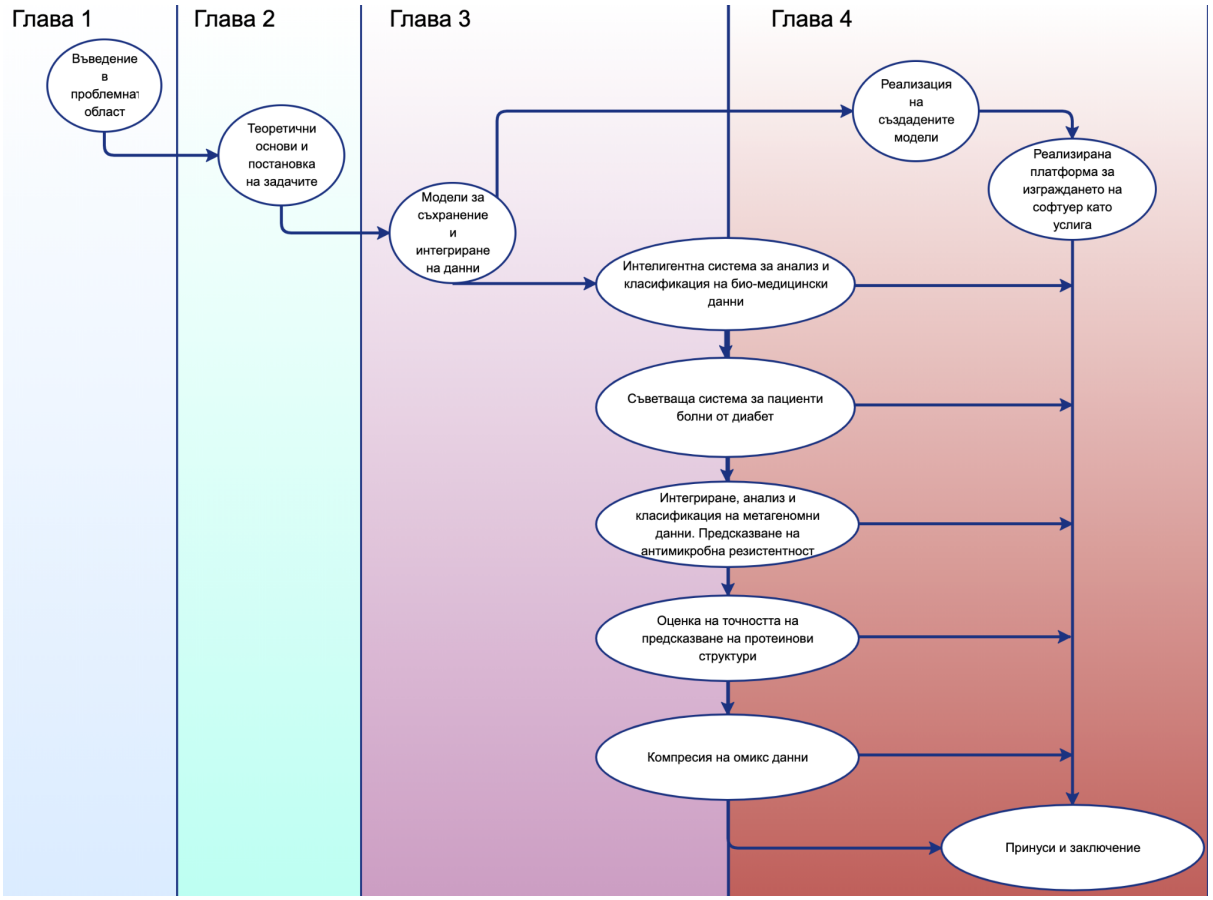
Основните задачи свързани с целта на дисертацията биха могли да се систематизират както следва:

1) Задачи свързани със съвместно използване на данни с клиничен и лабораторен, молекулярен профил:
   a) Разработване на модел и софтуерна реализация за интегриране на хетерогенни био-медицискни данни;
   b) Разработване на модел и софтуерна реализация за семантично интегриране на данни от ракови заболявания;
   c) Разработване на модел и софтуерна реализация за прогнозиране на преживяемостта на пациенти с ракови заболявания, използвайки инструменти на машинно обучение
2) Задачи, свързани със създаването на система за вземане на решения при извънклична терапия на болни от диабет:
   a) Разработване на модел и софтуерна реализация за извличане на знания от семантично интегрирани данни за целите на основана на онтологии система за вземане на решения с приложение при лечение на диабетоболни;
   b) Разработване на модел и софтуерна реализация за генериране на съвети за формиране на диети на базата на извлечени закономерности при болни от диабет.
3) Задачи, свързани с анализ на метагеномни данни с цел класификация по микробиомна резистентност:
   a) Разработване на модел и софтуерна реализация за семантично интегриране на данни от метагенмни изследвания, свързани със замърсявания в градска среда;
   b) Разработване на модел и софтуерна реализация за определяне на произход, анализ и класификация на метагеномни данни за микробиомна резистентност с използване на методи на машинно обучение.
4) Задачи, свързани с алтернативни методи за прогнозиране структурата на протеини:
   a) Използване на методи на машинно обучение и софтуерна реализация за целите на определяне точността на нагъване на протеинови структури
   b) Оценка точността на разработената методология за предсказване нагъването на протеинови структури с използване методи на машинно обучение
5) Задачи, свързани с компресиране на данни от паралелно секвениране (Next Generation Sequencing, NGS):
   a) Разработване на модел и софтуерна реализация за компресиране на големи масиви секвенционни данни посредством алгоритми на шумозащитно кодиране;
   b) Оценка на точността на работа на създадения модел за компресиране на данни от паралелно секвениране.

6) Разработване на платформа за предоставяне на софтуер като услуга, използвана за реализация на всички системи в дисертацията.

Структура на дисертационния труд

На фигура 1.2. е представена структурата на дисертацията като семантично свързани дейности чрез отделните глави на труда: описани и постановка на проблема (глава 2), методология на развитие на поставените задачи (глава 3) и реализация и обсъждане на поставените задачи (глава 4). Самата смислова част на развитие на идеята за семантично интегриране е представена посредством отделните дейности спрямо съответните глави, като всичко произлиза от основната цел и съответното предложение за създаване на модели за съхранение и интегриране на данни на базата на реализирани разпределени софтуерни решения като услуга. Основната идея е да бъда постигната целта за създаването на методология и практическата й реализация за интелигентно интегриране на биомедицински данни и техния анализ, използвайки средства на информатиката, информационните технологии, биоинформатиката и изкуствения интелект.



Фигура 1.1. Структура на дисертационният труд

Глава 1: Увод

Описание на проблема

С бързото развитие на високо производителните технологии, генериращи така наречените “-омикс” данни (данни от всички изследвания и анализи в областта на геномиката, протеомиката,

метаболомиката, транскриптомиката, метагеномиката и др. молекулярни науки) в областите на биологията и медицината, особено тези от масивно паралелно секвениране (Next Generation Sequencing (NGS))[1], и вследствие на това - бързо нарастващия обем на такива масиви от биологични данни, бяха създадени разнообразни средства и хранилища (бази от данни и уеб сървъри), за да се улесни управлението на данните, достъпността и последващия им анализ. Предпоставка за изследване на биоинформатиката е възможността за търсене и намиране, анализиране, както и достъп до данни, съхранени в хранилища или различни ресурси, съдържащи данни. За дадена биоинформатична задача изследователите често трябва да имат значителен опит при работа с различни източници на данни, свързан с търсенето на информация, извличането на знания и анализ на тези данни. Безспорно, интегрирането на данни е времеемък и ресурсоемък процес, особено що се отнася до импорта и експорта на огромни масиви от данни свързани с високопроизводителни изследователски и диагностични технологии в областта на биологията и медицината. В този контекст интегрирането на големи масиви разпределени, хетерогенни и различни по формат и времеви произход данни се оказва съществен проблем за пълното използване на богатството от биологични данни [2]. В този контекст, значението на интегрирането на данни от биомедицински изследвания и практики базирани на високопроизводителни технологии (като генериране на -*омикс* данни) има два основни компонента [3]:

(1) поради голямото ниво на автоматизация на действителните експериментални процедури, усилията за получаване на експерименталните данни отнемат само около 20% или по-малко от общите изследователски усилия в проект високопроизводителни технологии за генериране на омикс данни; като приблизително четири пети от ресурсите отиват за интегриране и анализ на тези данни [4];

(2) отговорите на най-важните, сложни биологични въпроси днес рядко се предоставят директно чрез експериментални резултати, като за предоставяне на потенциални отговори, анализът често включва интегрирането на разнообразни данни от множество източници на данни.

Настоящият дисертационен труд е посветен на кръг от проблеми съпътстващи големите масиви от данни и тяхното интегриране в приложната област на данни от медицински и биологични изследвания и практики. Несъмнено основна роля в дисертационната работа играят начините за съхранение на данни, използването на нови технологии от NoSQL бази от данни, както и последващия анализ, класификация, извличане на знания от големи масиви биомедицински данни. Съществен принос за дисертационния труд имат и работите за създаване на интегративна, основана на онтологии съветваща система за хранителен режим при болни от диабет, анализ и класификация на метагеномни данни по отношение на антимикробна резистентност, и методи за компресиране на големи масиви секвенционни данни. Отличителна черта на дисертацията е разработването на различни приложения, имащи нови методологически аспекти и предлагащи готови софтуерни решения, които извън предложените примери имат до много голяма степен универсален характер.

Основните параметри за значимост и актуалност на дисертационния труд биха могли да се систематизират като:

- Разработване на модел и софтуерна реализация за семантично интегриране, анализ и класификация на биомедицински данни (лабораторни и клинични) за ракови заболявания;
- Разработване на модел и софтуерна реализация за прогнозиране преживяемостта при пациенти с ракови заболявания с помощта на методи на машинно обучение;
- Разработване на модел и софтуерна реализация на основана на онтологии система за вземане на решения за съставяне на хранителен режим при пациенти, болни от диабет;
- Разработване на модел на базата на машинно обучение и софтуерна реализация за предсказване на протеинови структури, тяхната класификация и анализ за оценка на точността;
- Разработен е модел и софтуерна реализация за компресиране на данни от паралелно секвениране посредством алгоритми за шумозащитно кодиране;

- Всички софтуерни реализации в дисертационния труд са направени на базата на създадена платформа за представяне на софтуер като услуга.

Сложност на проблема

Като основни характеристики на сложността на проблемите пред дисертационния труд биха могли да се представят:

- Хетерогенната структура и произход на данните;
- Проблемите пред единна теоретична основа, отчитаща спецификата на интегрирането на биомедицински данни;
- Относително недостатъчната класификация на биомедицински данни;
- Не интензивното развитие и използване на предметни онтологии в медицината и биологията;
- Подходите за извличане на знания и създаване на услуги на тази основа са слабо интегративни;
- Изследванията , свързани с използване на значителни по размер големи масиви от данни са свързани с нови подходи в съхранението, трансфера, компресията на данни.

Области и средства на изследванията в дисертацията

Информатика и компютърни науки. Главната задача на работата - интегриране на данни е един от най-нуждаещите се от бързи и универсални решения проблеми в областта на информатиката и компютърните науки. Използването на NoSQL бази от данни за съхранение и интегриране на данните използвани в работата е също бързо развиваща се област на информатиката, Изпозлването на методи на машинното обучение почти във всички задачи в дисертацията допринася несъмнено за получаването на оптимални решения. Софтуерните реализациии на отделните задачи в дисертацията са разработени в съвременна среда на програмни езици, библиотеки, платформи и потребителски интерфейс.

Биоинформатика. Биоинформатиката е интердисциплинарна научна област, която се занимава с изследването на данни от различни части на биологията и медицината със средствата на информатиката. Биоинформатиката се обуславя изцяло от разработване на алгоритми и софтуер, които да могат да бъдат използвани както за интегриране и анализ на постъпващата информация, така и за извършване на числови експерименти, класификация и извличане на знания. С развитието на информационните технологии, както и с развитието на технологиите в областта на медицинските и биологични изследвания, обемът генерирани данни нараства с много бързи темпове, понякога по-бързо и от самите изчислителни ресурси, с което силно нарастват и възможностите за разширяване на изследванията за постигане на нови много по-значими резултати, както и извличане на нови знания.

Изкуствен интелект. Изкуственият интелект е област, която се използва за изграждане на интелигентни софтуерни системи, които могат да решават все по-голям кръг от проблеми, които досега са правени с конвенционални изчислителни средства. Този кръг от проблеми може да включва всякакви задачи, които не се поддават на решаване с формално описани алгоритми - анализ на данни, разпознаване на образи и реч, роботика, самообучаващи се програми. Изкуственият интелект е област, която се развива изключително бързо и с успешни приложения във всички области. Особена роля има в работата с биомедицински данни и въобще, като приложение в медицината и съвременните биологически изследвания.

Глава 2: Теоретични основи и анализ на състоянието по проблемите за интегриране, анализ и класификация на биомедицински данни.

Системите за интегриране на данни, се характеризират с архитектура, базирана на глобална схема и набор от източници. Източниците съдържат реални данни, докато глобалната схема предоставя съгласуван, интегриран и виртуален изглед на основните източници. Следователно моделирането на връзката между източниците и глобалната схема е ключов аспект. За тази цел има няколко основни подхода за интегриране на данни, които могат грубо да бъдат класифицирани в пет групи [23],[24]: съхранение на данни, обединение на база данни,

ориентирана към услуги интеграция, семантична интеграция и wiki- базирана интеграция. Във всички тези групи в значителна степен все по-важен компонент на интеграцията на данните заемат дейностите за разработване на разнообразни онтологии, за по-специфично справяне с техническите предизвикателства за определянето на дескрипторите и идентификаторите на информация, която трябва да се споделят и интегрират от различни ресурси [19], [25].

В днешно време се събират огромни обеми от данни от много разнородни източници, които генерират данни в реално време с различни качества - което се нарича *big data* или големи масиви от данни (ГМД). Интегрирането на големи масиви от данни е широка област с много предизвикателства, особено след като традиционните техники за интегриране на данни не успяват да се справят с това.

Можем да кажем, че интеграцията на големи масиви от данни се различава от традиционната интеграция на данни в много измерения: обем, скорост, разнообразие и достоверност, които са основните характеристики на големите данни: Обемът на данните е оригиналният атрибут на така наречените големи масиви от данни. В днешно време броят на свързаните устройства и хора е по-висок от преди, което силно повлиява на броя на източниците на данни и количеството данни по целия свят и във всяка една сфера. Разнообразието от източници на данни предполага, че имаме повече разнообразие във форматите, в които се съхраняват данните. Имаме структурирани и неструктурирани данни на високо ниво. Във всеки тип имаме огромен брой формати: текст, изображения, звуци, документи, пространствени данни и други. Фактически имаме различно качество на данните което води до проблем с тяхната достоверност. Можем да намерим несигурни или неточни данни във всички области от социалните мрежи до биоинформатичните системи, които интегрират данни.

Съхранение на данни

Подходите за съхранение на данни предлага решение как това да се прави (условно на едно място) с цел улесняване на достъпа и управлението на голямо разнообразие от данни от различни източници. Хранилищата за данни се фокусират върху транслацията на данни, събирането на всички достъпни данни от много различни източници, трансформирането и добавянето им в самото хранилище. Хранилището за данни има много аспекти и може да бъде поместено на различни физически места, това зависи от обема, скоростта на генериране и разнообразието на данните, които трябва да се интегрират. Едно хранилище на данни е съвкупност от множество различни бази от данни, които могат да взаимодействат помежду си. Основните свойства на хранилищата на данни се свързват от многомерни анализи до изисквания за статистически данни и извличане на данни до възможности за проучване, както и въвеждането на адаптивни аналитични приложения, като тези технологии са част от стабилна и проверена среда.

Основните свойства на хранилището за данни са: да не се съдържат излишни данни (nonredundant), хранилището да бъде винаги достъпно (stable), записаните данни в хранилището винаги да са точно тези, които потребителят е избрал (consistent). Всички тези свойства могат да бъдат удовлетворени от облачните услуги, където ресурсите, които са на разположение са много по - големи и възможностите за изграждане на йерархично хранилище на данни е възможно.

Хранилища за данни. Предизвикателствата пред съхранението са свързани предимно с обема, скоростта и разнообразието от големи масиви от данни. Съхраняването на големи масиви от данни като традиционно физическо съхранение е проблематично, тъй като твърдите дискове (HDD) често се повреждат, а традиционните механизми за защита на данните (напр. RAID или масив от независими дискове) не са ефективни при съхранение в петабайт мащаб [26]. Необходимо е да се разработят принципи и алгоритмични решения, като се вземат предвид пространствено-времевите модели на използване на данните, за да се определи аналитичната стойност на данните, както и съответните данни за тяхното запазване чрез балансиране на разходите за съхранение и предаване на данни с бързото натрупване на големи данни [28].

Трансфер на данни

Трансферът на данни протича на различни етапи от жизнения цикъл на данните, както следва: (i) събиране на данни от сензори, биологични инструменти; (ii) интеграция на данни от множество центрове за данни; (iii) управление на данните за прехвърляне на интегрираните данни към платформи за обработка (напр. платформи в облак) и (iv) анализ на данни за преместване на данни от хранилище към анализиращ хост (напр. клъстери с висока производителност (High Performance Computing)). Прехвърлянето на големи обеми данни поставя очевидни предизвикателства на всеки от тези етапи. Следователно са необходими интелигентни техники за предварителна обработка и алгоритми за компресиране на данни, за да се намали ефективно размерът на данните преди прехвърлянето на данните [29].

Гъвкаво управление на данни

За компютрите е трудно ефективно да управляват, анализират и визуализират големи, неструктурирани и разнородни данни. Разнообразието и достоверността на големите данни предефинират парадигмата за управление на данните, изисквайки нови технологии (напр. Hadoop, NoSQL) за почистване, съхраняване и организиране на неструктурирани данни [30]. Докато метаданните са от съществено значение за целостта на произхода на данните [31], предизвикателството остава да се генерират автоматично метаданни за описание на ГМД и съответните процеси. Генерирането на метаданни за геопространствени или биологични данни е дори предизвикателство поради присъщите характеристики на данните с висока размерност и сложност (например корелация на пространството и времето и зависимост). Освен генерирането на метаданни, големите масиви данни също създават предизвикателства пред системите за управление на бази данни (СУБД), тъй като на традиционните RDBMS липсва скалируемост за управление и съхранение на неструктурирани големи данни [32],[33]. Докато нерелационните (NoSQL) бази данни като MongoDB и HBase са предназначени за големи масиви от данни [34], моделите на гъвкаво адаптиране на NoSQL бази данни за обработка на геопространствени или биологични ГМД чрез разработване на ефективно пространствено-времево индексиране и алгоритми за заявки все още са предизвикателство [35]. Решенията NoSQL се основават на три основни теореми: CAP, BASE и евентуална последователност.

Анализ на данни

Анализът на данни е важна фаза в моделите и алгоритмите, използващи големи масиви данни за извличане на информация и закономерности [45]. Анализът на големите данни от своя страна поставя следните проблеми за сложност и мащабируемост на основните алгоритми [46]. Анализът на големи данни изисква сложни мащабируеми и оперативно съвместими алгоритми [47] и е адресиран чрез програми за анализ на и паралелни платформи за обработка (например Hadoop), за да се използва силата на разпределената обработка. Тази стратегия „разделяй и владей“ обаче не работи с многослоеви и многомащабни итерации [27], които са необходими за повечето алгоритми за анализ на биологични данни. Освен това повечето съществуващи аналитични алгоритми изискват структурирани хомогенни данни и имат затруднения при обработката, отчитайки хетерогенността на ГМД [29]. Този отворен въпрос изисква или нови алгоритми, които се справят с разнородни данни, или нови инструменти за предварителна обработка на данни, за да ги направят структурирани така, че да отговарят на съществуващите алгоритми. В биоинформатиката оптимизирането на съществуващите алгоритми за пространствен анализ чрез интегриране на пространствено-времеви принципи за ускоряване на откриването на биологично знание е предизвикателство и се превърна във високоприоритетно изследователско поле на „биоинформатичното мислене, изчисления и приложения“ [27].

Архитектура на данните

Големите данни постепенно трансформират начина, по който се провеждат научните изследвания, както се доказва от все по-насочения към данните и отворен научен подход [47]. Подобни трансформации създават предизвикателства пред системната архитектура. Например, безпроблемното интегриране на различни инструменти, геопространствени услуги както и биологични изследвания остава основен приоритет. Допълнителните приоритетни въпроси

включват интегриране на тези инструменти в работни процеси за многократна употреба, включване на данни в инструментите за насърчаване на функционалността [45] и споделяне на данни и анализи между общностите. Идеалната архитектура безпроблемно ще синтезира и споделя данни, изчислителни ресурси, мрежа, инструменти, модели и най-важното - хора [50]. Биоинформатичната кибер-инфраструктура се използва активно в биомедицинските науки [51]. NCBI, макар и все още в разработка, е добър пример за такава киберинфраструктура в биоинформатичната област.

Интеграция, ориентирана към услуги

Съхранението на данни и обединението на данни се фокусират върху централизирането на достъпа до данни, съответно чрез превод на данни и превод на заявки. Те се сблъскват с някои проблеми, произтичащи от съхранението и обработката на данни, честите актуализации и високите разходи за обмен на данни и/или поддръжка. Отчасти за избягване на тези проблеми е усъвършенстван и децентрализиран подход, при който отделни източници на данни се съгласяват да отворят своите данни чрез уеб услуги (WS). WS са предназначени за комуникация между компютри през мрежата и са описани от езика за описание на уеб услугите (WSDL). Има няколко различни протокола за WS, например SOAP (Simple Object Access Protocol; протокол за обмен на XML-базирани съобщения през компютърни мрежи), REST (REpresentational State Transfer; прост протокол, реализиран с помощта на HTTP методи). WS поддържа взаимодействие компютър-компютър чрез интерфейс за програмиране на уеб приложения (Web API) и може да изпълнява заявка за база данни или изчисления. В контекста на интеграцията на данни данните могат да бъдат достъпни програмно чрез WS и източниците на данни служат като доставчици на услуги. Следователно този подход може да се разглежда като подход, ориентиран към услугата.

Подходът, ориентиран към услуги, включва интеграция на данни чрез комуникация между компютри чрез уеб API и актуално извличане на данни от различни източници на данни. Този подход остава предизвикателство, най-вече защото успехът му в хетерогенната интеграция на данни изисква много източници на данни да станат доставчици на услуги чрез отваряне на данните им чрез WS и чрез стандартизиране на идентичностите на данните и номенклатурата, за да се улесни обменът и анализът на данните.

Семантична интеграция

Повечето уеб страници в биомедицинските източници на данни са предназначени за четене от хора (например HTML). Семантичната мрежа [65],[66] има за цел да опише данните по начин, който компютрите могат да се разбират и да се изгради взаимосвързана мрежа, чрез която компютрите могат да лесно и недвусмислено да взаимодействат и анализират. Според декларацията за дефиниция от World Wide Web Consortium (W3C), целта на семантичната мрежа е да създаде универсален носител за обмен на данни, като се използват няколко стандарта, включително Resource Description Framework (RDF; http: // www. w3.org/RDF), RDF схема (RDFS - език за описание на речника на RDF; http://www.w3.org/TR/rdfschema), уеб онтологичен език (OWL; http://www.w3.org/owl) и стандартен език за уеб заявки SPARQL (http://www.w3.org/TR/rdf-sparql-query) за RDF. RDF предоставя стандартни формати (напр. XML формат) за обмен на данни и описва данните като просто изявление, съдържащо набор от тройки: субект, предикат и обект. Всякакви две твърдения могат да бъдат свързани от еднакъв предмет или обект. OWL се основава на RDF и Uniform Resource Identifier (URI) и описва структурата и значението на данните въз основа на онтологията, което дава възможност за автоматизирано мислене на данни и изводи от компютри. Семантичната мрежа осигурява машинно четим начин за представяне на данни и оперативна съвместимост [67],[68].

Прилагането на семантичните уеб технологии за интеграция на биологични данни е значителен напредък за биоинформатиката, позволяващ автоматизирана обработка на данни и знания. Семантичната интеграция използва онтологии за описание на данни и по този начин представлява интеграция, базирана на онтологии [68]. Семантичната мрежа продължава да се развива и нейното приложение при интеграцията на биологични данни има няколко ограничения. Семантичната интеграция локално съхранява голяма колекция от RDF документи,

чрез копиране на данни от множество източници на данни и преобразуване на данни в RDF формат.

Стандарти за биомедицински данни

Високопроизводителните технологични платформи, като например платформите за паралелно геномно секвениране (NGS) в био-медицината могат да генерират огромни количества данни за относително кратък период. За да бъдат в крак с революцията на технологиите за секвениране, проектите за секвениране на геноми постепенно преминават от класически моделни организми (напр. плодова мушица (*Drosofila M*) , мишка, дрожди). От друга страна невъзможно е да се интегрират толкова големи количества данни в една среда (например хранилище за данни). Източниците на данни са разработени за различни цели и изпълняват различни функции. Следователно, перспективно е да се създаде ефективен начин за обмен на данни между тези разпределени и разнородни източници на данни. Дузина източници на данни обаче са предназначени само за съхранение на данни, но не и за обмен на данни. Нарастващият обем на биомедицински данни също изисква „компютърно четими“ подходи за интегриране на данни. За да се улесни интегрирането на данни, източниците на данни трябва да се превърнат в доставчици на услуги. С други думи, източниците на данни трябва не само да служат като доставчици на данни, които предоставят данни за четене от човека с уеб интерфейси (например HTML), но също така да функционират като доставчици на услуги, които предоставят данни за компютърна оперативна съвместимост чрез WS. Доставчиците на услуги предоставят данни като WS, улеснявайки взаимодействието между компютри и по този начин позволявайки автоматизирано интегриране на данни от множество източници на данни [74]. Както споменахме, има няколко различни протокола, които могат да се използват за създаване на WS. Сред тях SOAP и REST са широко възприети (Фигура 2.9.). SOAP е добре дефиниран стандарт с XML-структурирани съобщения за заявка и отговор, докато REST е относително лек, разчитайки на HTTP методи (а именно POST, GET, PUT или DELETE). Повечето търговски приложения излагат своите услуги като RESTful Web API (Фигура 1), до голяма степен благодарение на неговата простота и лесна реализация.

Фигура 2.1. Статистика на различните Web API протоколи

Също толкова важно е, че интегрирането на данни също изисква стандартизиране на номенклатурата и онтологиите за биомедицински данни.

Базирани на софтуер като услуга подходи за интегриране на данни

Целта на интеграцията на данни е да даде възможност за автоматично комбиниране на информация от различни ресурси без човешка намеса, така че да се справи с нарастващото натрупване на биомедицински данни. Към тази цел данните, които трябва да бъдат интегрирани, трябва да бъдат предефинирани по-широко, което включва не само последователности и други необработени данни, но и методи, инструменти, алгоритми, анализирани резултати, открити знания (вж. Статия за интеграция на знания; , 2007) и дори връзки между хората [77]. Всички видове данни могат да се предоставят като услуга. Тоест суровите данни трябва да бъдат достъпни чрез WS, методите, инструментите и алгоритмите, които се използват за анализ на данни, трябва да се предлагат като WS (т.е. SaaS, Софтуер като услуга), а анализираните резултати и откритите знания също трябва да се доставят като WS [77]. В резултат на това WS извършва различни манипулации на данни, включително извличане, интегриране, анализ, визуализация и споделяне на данни.

Конвейер с комбинация от множество WS може да постигне интеграция на данни. Такива WS-базирани конвейри намаляват технологичните бариери и осигуряват на потребителите по-лека среда за програмиране. Базираните на WS конвейри включващи обмен на данни между компютър и компютър, опростяват интеграцията и анализа на данни, увеличават максимално обхвата на споделянето и повторната употреба, и функционират като среда за свързване на потребители, разположени навсякъде със сходни изследователски интереси, и накрая за формиране на научна социална и проектна общност.

Интегриране на биомедицински данни и откриване на знания

Интегрирането на мулти -*omics* данни от един и същи набор от обекти се очаква да увеличи точността и скоростта на прогнозирането на резултатите, например ранното откриване на ракови заболявания въз основа на данни от много платформи. Технологиите за анализ на „*omics*“ данни се характеризират с високопроизводителни интерфейси, които улесняват изследването на генома, епигенома, транскриптома, протеома и метаболома по глобално безпристрастен начин. Подходи за анализ на -*omics* данни сега се използват за разбиране на сложните биологични системи и за разкриване на молекулярните сигнатури, които стоят в основата на сложните фенотипове [80]. При интегрирането на мулти OMICS данни, които не са индексирани се използват други подходи, които са част от моделите на машинното обучение с цел откриване на нови групи и подгрупи на обектите. Определяне на вида на раковото заболяване или типизирането на тумор е често срещан проблем, които се решава посредством машинното самообучение без учител. Ефективността на клъстерирането може да бъде изчислена количествено, като се използва симулационно проучване, разширения на няколко изгледа на критериите за индекс и анализ на обогатяването [81].

Интегрирането данни за ракови заболявания, които са взимани от различни клинични лаборатории, имат различни формати, свойства и структура, но описват едно и също раково заболяване. В случай се използват методите на машинното обучение, за извличане на закономерности и сходни свойства между отделните обекти. По този начин могат да се изградят връзки между хетерогенните данни.

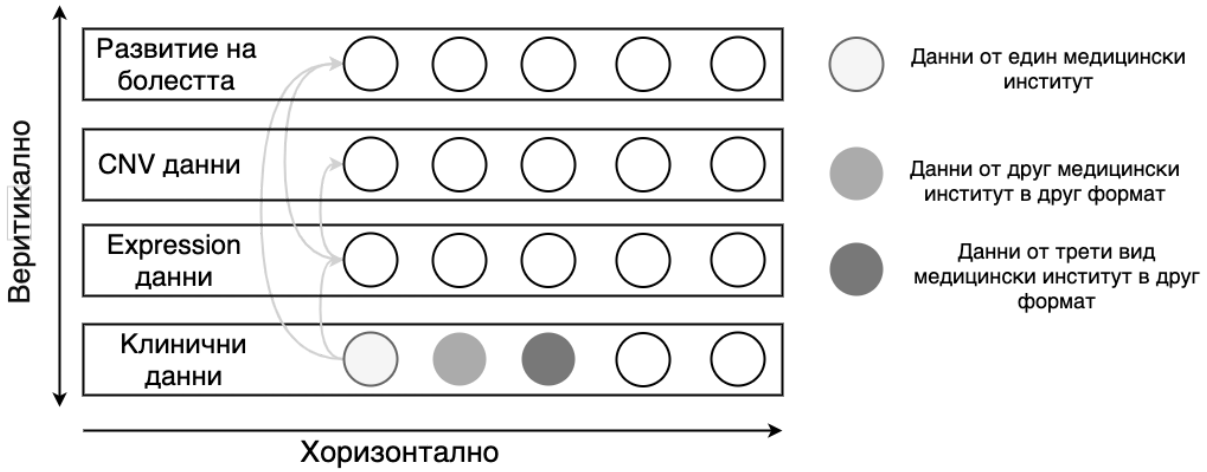
При анализа на големите биологични данни става задължително да се изследват основните принципи на интегрирането на данни от множество източници, за да се предостави изглед от по - високо ниво с цел извеждане на нови знания, основани на използване на методи от машинното самообучение при интегрирането на данни [82] През последното десетилетие технологиите с висока производителност се използват масово заедно с клинични тестове за изследване на различни заболявания, за да се дешифрират основните биологични механизми и да се разработят нови терапевтични стратегии. Генерираните данни с висока производителност често съответстват на измервания на различни биологични параметри (напр. генна експресия, РНК транскрипти, протеини), представляват различни възгледи за един и същ обект (напр. генетични, епигенетични) и се създават чрез различни технологии (напр. микрочипове, секвениране следващо поколение и др.). Данните са разнородни, от различни видове и формати.

Глава 3: Формализация и методи за интелигентно интегриране, анализ и класификация на биомедицински данни.

Хетерогенност на биомедицинските данни. През последното десетилетие технологиите с висока производителност се използват масово заедно с клинични тестове за изследване на различни заболявания, за да се дешифрират основните биологични механизми и да се разработят нови терапевтични стратегии. Генерираните данни с висока производителност често съответстват на измервания на различни биологични единици (напр. транскрипти, протеини), представляват различни възгледи за един и същ обект (напр. генетични, епигенетични) и се създават чрез различни технологии (например микрочипове, РНК-секвениране). Данните са разнородни, от различни видове и формати. Очевидна е необходимостта да се интегрират данните, за да се съхраняват, достъпват, свързват, анализират и добиват лесно [86].

Обобщено погледнато, от методологична гледна точка, интегрирането на данни в контекста на биоинформатиката се разбира обединяване на данни от различни източници и цел създаване на единна представа, форма, начин на постъпване и възможност за извод на знания. Всеки източник на данни в биоинформатиката има свой собствен подход за структуриране на данните, което увеличава сложността за интеграция. Интегрирането на данни и анализа на биомедицински данни са отделни дисциплини, които са се развили относително изолирано. Съществува общо съгласие, че обединяването на двете дисциплини с цел разработване на по-устойчиви методи за анализ е необходимо [87],[88]. Интеграцията на данни основно включва заявки към различни източници на данни. Тези източници на данни могат да бъдат, но не се ограничават до отделни бази от данни или полуструктурирани източници на данни, разпределени в мрежа. Интеграцията на данни улеснява разделянето на цялото пространство от данни на две основни измерения, отнасящи се до мястото, където се намират данните или знанията за метаданните, и до представянето на данни и модели на данни. Биомедицинските експерименти ползват голям брой различни аналитични методи, които улесняват извличането на релевантни данни от разпределена информация.

Методи за интегриране на биомедицински данни. Хетерогенността на биомедицинските данни прави всеки интегративен анализ силно предизвикателен. Данните, генерирани с различни технологии, включват различни набори от атрибути. Когато данните са силно разнородни и слабо свързани, се прилагат два взаимосвързани интегративни подхода: хоризонтална и вертикална интеграция (Фигура 3.1.). Хоризонталната интеграция на данни обединява информация от същия тип, но от различни източници на данни и, потенциално, в различни формати. Това улеснява обединяването на разнородни данни, като клинична информация, от много различни източници в един модел данни. Вертикалната интеграция на данни, от друга страна, означава свързване на различни по тип данни за да се постигне по добър анализ и извеждане на знания за множество типове данни. Този подход помага за управление на връзките между генетичната експресия на пациента, клиничната информация, наличните химически знания и съществуващите онтологии. Повечето съществуващи подходи за интегриране на данни се фокусират върху един тип данни или едно заболяване и не могат да улеснят интеграцията на кръстосан тип или заболяване [91],[92].



Фигура 3.1. Вертикална и хоризонтална интеграция на данни от различни източници и с различни формати.

Основните проблеми, които стоят пред вертикалното и хоризонталното интегриране на данни са свързани с хетерогенността на данните. Хетерогенността на данните е общо понятие, което в контекста на интегрирането на данни може да се определи като съвкупност от проблеми. Един от основните проблеми, които водят до хетерогенност е използването на различните операционни системи, платформи и хардуерни конфигурации. В контекста на биомедицинските данни съществува и много силно изразена хетерогенност във използваните формати за представяне на данни. Различните формати на едни и същи данни водят до хетерогенност и в структурата и изводите, които могат да се направят от тези данни. Разликите в знанията, които могат да се изведат се нарича семантична хетерогенност. Характеризира се с използването на много ресурси, които анализирайки от различни входни точки се извличат противоречиви знания. Един от разпространените варианти за решение на този проблем е използването на онтологии.

Семантично интегриране на био-медицнски данни.

Основната идея на семантичната мрежа е да се добавят машинно четими метаданни към ресурси в глобалната мрежа, за да се дефинират и опишат отношенията между тях. Семантичните уеб технологии са в състояние да усвоят тази придобита информация. Освен това те не изграждат отделна мрежа, а функционират като разширение на текущата мрежа. Технологията Semantic Web се състои от йерархично използване на различни стандарти и технологии, при които всеки слой използва възможностите на слоевете по-долу.

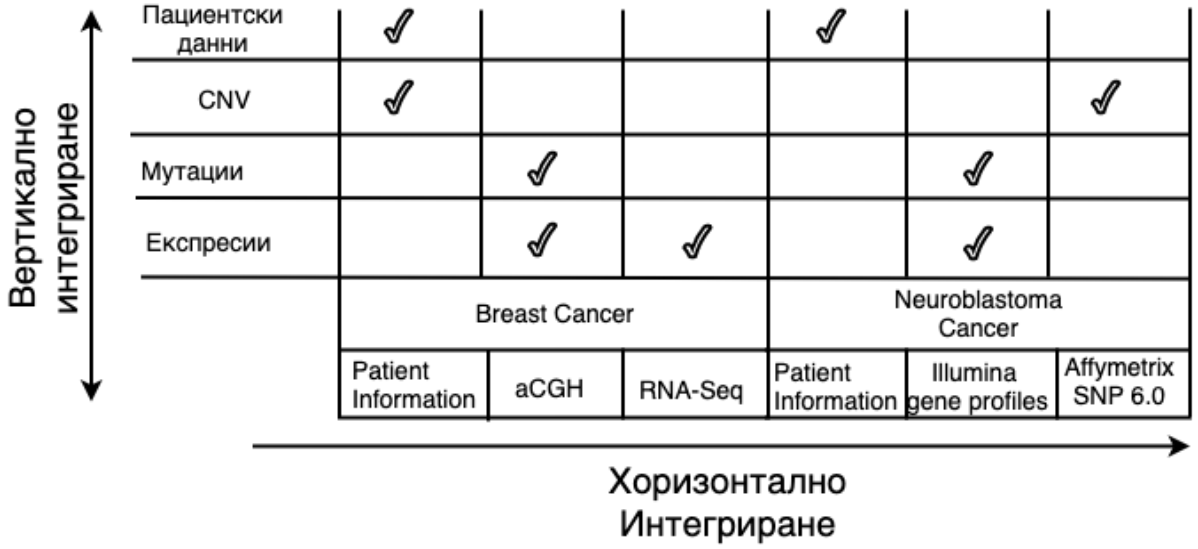
Проблемът при използването на подход с чисто семантично интегриране е въвеждането на множество мета данни. Тези мета данни трябва да имат също общ формат, което е голям проблем при работата с биологични данни. Тези мета данни трябва да могат да описват достатъчно добре както например в клиничен аспект ракови заболявания така и характеристики на бактерии и други свързани заболявания. За тази цел е нужен по динамичен модел за свързване на данните, който може да се адаптира спрямо типа на данните, който се интегрират. За тази цел можем да използваме така наречение свързани данни *(Linked data)*. Свързаните данни са най-усъвършенстваната форма за публикуване на данни в мрежата според спецификациите на W3C. W3C описва множество добри практики за публикуване на данни в мрежата [94]. Този подход има смисъл само когато данните са в общодостъпни. Тази мрежа за данни понякога се нарича „Мрежа от данни“, термин с по-практичен акцент от по-старата, но еквивалентна „Семантична мрежа“.

Основната идея на свързаните данни е, че всеки обект - анимиран или неодушевен, частен или абстрактен - може да има идентификатор: универсален идентификатор на ресурс или URI (*Uniform Resource Identifier)*. Данните в мрежата от данни се отнасят до обекти, които са много точно идентифицирани. URI са последователности от символи с няколко части, разделени

с точки и наклонени черти. Например URL (*Uniform Resource Locator*) адресите (универсален локатор на ресурси), които са уеб адресите, въведени в уеб браузър, за да се получи страница, също са вид URI. Това съвпадение, което превръща URI-тата в надмножество на URL адресите, не е случайно: очаквано поведение при въвеждане на URI в уеб браузър е извличането на информация за идентифицирания обект. Низът от символи, използван за идентифициране на нещо, извлича повече информация за това нещо. URI имат за цел да именуват всички обекти в света по еднакъв начин.

Втората ключова идея на свързаните данни е, че може да се даде информация за всеки идентификатор. Данните за избраният обект се извличат от посоченият URI. На база на него може да се направи разделение на използваните източници от данни. По този начин се постига разпределение на данните и организация на данните, които са общи.

Интегриране на данни от множество различни източници на примера на ракови заболявания. В дисертацията е предложен подход, за хоризонтално и вертикално интегриране на данни се обединяват цели набори от данни, като семантичната цялост на данните се запазва и обогатява. Чрез комбиниране на данни от множество ракови заболявания по този начин се създава мрежа от данни, където обекти, като протеини, клинични характеристики и експресивни характеристики, са свързани помежду си [95]. Данните често могат да бъдат представени като мрежи, където възлите показват биологично значими обекти (обикновено гени или протеини), а дъгите представляват връзки между тези обекти (например регулиране, взаимодействие). В генерираната мрежа възлите представляват пациенти, а дъгите представляват прилики между профилите на пациентите, състоящи се от клинични данни, експресионни профили и информация за броя на копията на варирането на гените (*CNV - copy number variation*). Такава мрежа може да се използва за групиране на подобни пациенти и за асоцииране на тези групи с различни характеристики [96]. Основните предизвикателства тук са: (1) изграждане на подходяща свързана мрежа от данни, откриване на полуструктура на модела на данните [97] и картографиране на твърдения от прилагания модел за интегриране на данни [98]; и (2) почистване на данни, комбинирано в официален работен поток за интегриране на данни Фигура 3.3.



Фигура 3.3. Интегриране на хетерогенни данни. По хоризонталната ос са показани различни типове данни и източници, интегрирани за конкретен пациент. По вертикалната ос са дадени интегрирани типове данни, свързани с изследваните ракови заболявания и свързани с определен пациент.

За хоризонтална и вертикална интеграция на данни се изискват различни бази от данни, тъй като всеки от тези подходи се занимава с различни аспекти на проблема с интеграцията. Хоризонталната интеграция на данни се занимава с неструктурирани и разнородни данни. По

този начин в работата е използвана базирана на документи база данни (като MongoDB), която може да обработва различни типове и формати данни. За вертикална интеграция на данни се прилага база данни, базирана на графи, тъй като е подходяща за представяне на релации, които са от решаващо значение в този случай. В това проучване всички връзки се установяват между съществуващите записи за всеки обект и са представени от полуструктура (semi-structure).

Предложената интегративна рамка улеснява директния анализ на данните. Първо се фокусираме върху конкретно клинично значимо приложение: моделиране и прогнозиране на времето за оцеляване на пациенти с рак. Това се състои в прилагането както на конвенционални методи за класификация, така и на алгоритми за машинно обучение. Чрез интеграция на данни се въвежда нова интегрирана и универсална, т.е.приложима и за двата вида рак, функция за прогнозиране на времето за оцеляване. Тази характеристика е изградена от три клинични характеристики, които са най-свързани с преживяемостта. Освен това тази интегрирана функция осигурява връзка с новоразработената свързана мрежа за данни. Тази характеристика се използва в метода на конвенционалната класификация k-съседи, за да се намерят пациенти, които са най-тясно свързани с изследвания. След това чрез свързаните данни намираме други пациенти, които може да нямат новата интегративна функция, но все пак са свързани с различни видове данни, като генна експресия или CNV. След това се използват модели за машинно обучение, базирани на регресия на вектор за подкрепа и дърво за решения, за прогнозиране на времето за оцеляване и кръстосано валидиране.

Създадена методология за семантичното интегриране на биомедицински данни от различни заболявания.

Задачата е да се осигури метод, базиран на свързани данни и технологии с отворен код, който да комбинира знания от много съществуващи отворени източници за ефективно интегриране на сурови либрационни данни. Суровите данни за либрацията, които в крайна сметка могат да бъдат интегрирани за цялостно изясняване на сложни фенотипове, включват функционални генни анотации, профили на генна експресия, протеомни профили, ДНК полиморфизми, вариации на броя на ДНК копия, епигенетични модификации и др. [101].

Специфично предизвикателство в проучването е да се интегрират и анализират набори от небалансирани и неструктурирани данни. Молекулните данни са в суров формат с всички полета и атрибути, генерирани от технологията за секвениране или микрочипове. Преди да се започне процеса на интеграция, е необходимо да се извършат някои операции за предварителна обработка на суровите формати и да се генерира подходяща нова структура от данни. В случая се работи с набори от данни, богати на връзки, и е от съществено значение да може да се намерят много анотации за съществуващите връзки, които ще помогнат за подобряване на набора от връзки чрез подходящи ресурси от наличните изтоници на знания.

В дисертационния труд е избран подход, основан на семантична интеграция, тъй като повечето характеристики на изпозлваните данни имат различна семантика за всеки пациент, което е съществен фон за персонализираната медицина. Очакваните резултати от такъв тип подход включват идентифициране на скрити протеинови подтипове, отличаващи се с общи модели на мрежова промяна и предсказващ модел за развитие на рак, базиран на знанията за съединените протеини.

Описание на проблемната област и данните. Данните включват функционални генни анотации, профили на генна експресия, протеомни профили, ДНК полиморфизми, вариации на броя на ДНК копия, епигенетични модификации и др. [19]. Суровите данни във всеки изследван набор от данни са в определен формат и имат специфична семантика. Полето (атрибут) във всеки набор от данни има различни значения поради технологиите и последващия запис. Предоставените данни сами по себе си също съдържат информация за мутирали протеини, експресия и CNV.

Първоначалната точка за трансформация, групиране и интегриране са пациентските файлове. Генерираният запис за всеки конкретен пациент съдържа атрибути като възраст, пол, националност и др. В това проучване се използват два набора от данни - невробластом (NB) и рак на гърдата (BC). Наборът на невробластома съдържа RNA-Seq генни експресионни профили на 498 пациенти, както и експресия на Agilent микрочипове и данни за aCGH за съвпадна

подгрупа от 145 пациенти и съответна клинична информация. Наборът данни за рак на гърдата съдържа профили за микрочипове и данни за номера на копията и клинична информация (време за оцеляване, множество прогностични маркери, данни за терапията) за около 2000 пациенти.

Основни характеристики и нововъведения при изграждане на методологията за семантично интегриране.

Подходът за интегриране на данни, който е разработен, първоначално беше ориентиран към конкретно приложение - за обединяване на данни от реални проучвания и лечения на невробластома и рак на гърдата - но неговите дизайнерски характеристики го правят достатъчно генеричен и приложим в широк кръг от тематични области. В резултат на прилагането му се обединяват различни набори от данни, като семантичната цялост на данните се запазва и обогатява. В конкретен случай, чрез комбиниране на данни от множество източници (фиг. 3.4), е създадена нова мрежа от данни, където обекти, като протеини, клинични характеристики и експресивни характеристики, са свързани помежду си [90]. В тази мрежа възлите представляват пациенти, а дъгите представляват прилики между профилите на пациентите, състоящи се от клинични данни, профили на експресия и данни от CNV. Такава мрежа може да се използва за групиране на пациенти и за асоцииране на тези групи с различни характеристики. Основните предизвикателства тук са: (1) изграждане на подходяща свързана мрежа от данни, откриване на полуструктура на модела на данни и картографиране на твърдения от прилагания модел за интеграция на данни [106]; и (2) почистване на данни, комбинирано в официален работен поток за интегриране на данни.

Използвахме различни бази данни за хоризонтална и за вертикална интеграция на данни. Тези различни бази от данни са необходими, защото хоризонталната и вертикалната интеграция на данни се отнасят до различни аспекти на проблема с интеграцията. Данните за хоризонталната интеграция на данни са неструктурирани и разнородни. По този начин се използва ориентирана към документи база данни, която може да обработва различни типове данни и формати. За вертикална интеграция на данни се използва графика база данни, тъй като е подходяща за представяне на релации - от решаващо значение в този случай. В това проучване всички връзки се установяват между съществуващите записи за всеки обект и са представени от полуструктура.

Интеграционен модел върху база данни NoSQL може потенциално да обедини данните от медицински изследвания, алтернативно на най-често използваните методи за статистически анализ и машинно обучение. Повечето системи за NoSQL бази данни споделят общи характеристики, поддържащи мащабируемост, наличност, гъвкавост и осигуряващи бързо време за достъп за съхранение, извличане и анализ на данни [107],[108]. Освен това може да се разшири потенциала на модела, като се използват множество набори от данни, независимо от нивото на хетерогенност, конкретни формати, видове данни и т.н. - всичко това е много специфично за изследванията на рака.

Методологията за интегриране на неструктурирани данни от изследваните проби и пациенти се основава на правилното използване на нерелационни бази данни и онтологии на домейни като генната онтология (GO) [112]. Данните, които са включени в работата, съдържат скрити връзки между протеините, предоставени от различни пациенти в рамките на проучвания на двете заболявания (BC и NB). Използва се цялата налична информация за вече изградени взаимоотношения в източниците на данни, като се търси и намира допълнителна информация в някои източници на трети страни, за постигане на семантична интеграция на данните. По този начин, стъпка по стъпка, се разработва мрежа, която съчетава протеиновите връзки между пациентите и заболяванията. Предизвикателството тук е да се съхраняват всички взаимоотношения с техните циклични зависимости. Последните са възможни, тъй като един пациент има връзка с мутирал протеин (и), мутирали протеин (и) има / имат препратка (и) към експресията, а други пациенти имат препратки към същия протеин (и).

При всяко заболяване (BC и NB) всеки пациент има различен набор от мутирали или експресирани протеини. Само малки набори от мутирали протеини са равни и съществуват при всеки пациент. Всички протеини принадлежат към семейства, които съдържат много свързани протеини. Чрез прилагане на семантични анотации и техники за търсене се откриват и

комбинират всички протеини, които са семантично свързани с изследваните заболявания. По този начин може да се открие цялата достъпна и необходима информация за подобен брой свързани протеини.

По време на анализа на сурови данни се създава един вид „полуструктура“ на данните - структура, съдържаща само атрибути, съществуващи във всеки запис. В т.н. полуструктурирани данни обектите, принадлежащи към един и същ клас (протеинови мутиации, експресирали транскрипти и CNV), могат да имат различни атрибути, въпреки че са групирани заедно, а редът на атрибутите не е важен. Полуструктурираните данни стават все по-разпространени, напр. в структурирани документи и при извършване на просто интегриране на данни от множество източници. Традиционните модели на данни и езиците за заявки са неподходящи, тъй като полуструктурираните данни често са нередовни: липсват някои данни, подобни концепции се представят с помощта на различни типове, присъстват хетерогенни набори или обектната структура не е напълно известна [103].

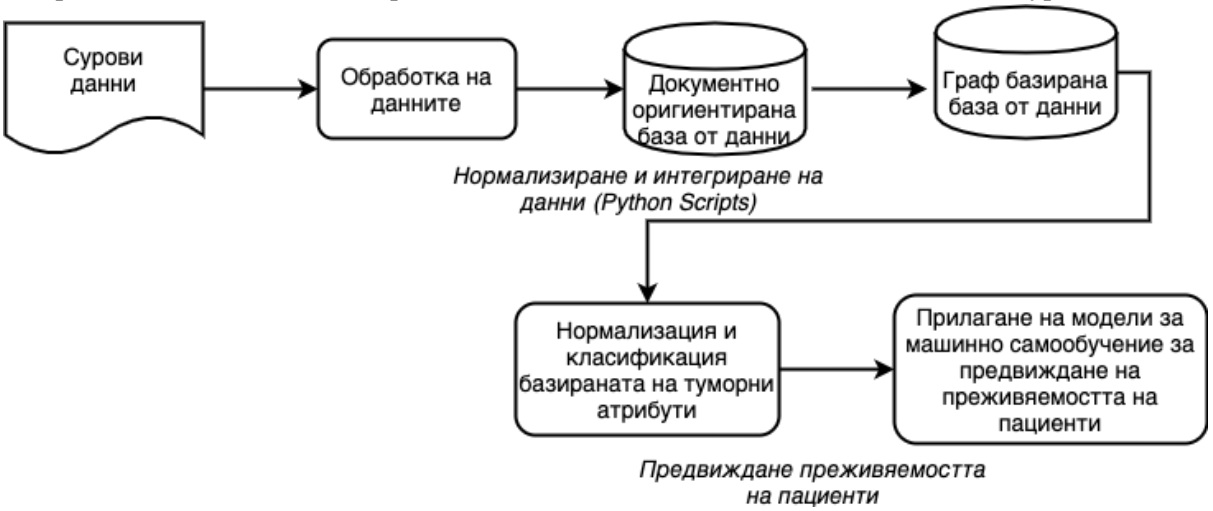
Модели за извличане на знания от биомедицински данни. Данните, използвани в дисертационния труд, се основават на клинични записи, включително: възраст на пациента, стадий на развитие на тумора, размер на тумора и жизнения статус на съответния пациент. Съществуват и данни относно различни видове терапии и хирургични интервенции, като по този начин са включени много характеристики, свързани с развитието на заболяването - размер на тумора, възраст при диагностициране, туморен стадий, информация за приложени операции и приложени лечения като химиотерапия, хормонална терапия и др. в суровите данни. В настоящата работа се използват два набора от данни от проучвания за рак на гърдата. Първият набор от данни съдържа профили на 498 пациенти, както и съответната клинична информация. Вторият набор от данни за рак на гърдата също съдържа данни за геномни профили и клинична информация за 2000 пациенти. Различните видове налични данни и източници на информация са показани на фигура 3.6. Един и същ тип информация се предоставя от различни източници в различни формати. Ние интегрираме всички данни както хоризонтално, така и вертикално.

Първична обработка на данни. Базата данни се състои от два слоя: първо, нерелационна ориентирана към документи база данни - клас бази данни, които съхраняват данните си под формата на документи. Тези бази данни са хоризонтално мащабируеми и много по-гъвкави от релационните бази данни. В допълнение, вторият слой е графична база данни - клас бази данни, които съхраняват данни под формата на графика и използва техника, наречена смисъл без индекс. В графичната базата данни основният акцент е върху връзката между данните. В документално-ориентираната база данни се прилага ограничение (наречено „схема на данни“) въз основа на генерираната полуструктура. Приложената схема за данни за всеки запис за всеки вид данни обединява данни в различни формати и от различни източници. За всеки вид данни, тази схема на данни винаги съдържа ID и проба ID (представляваща името на субекта, както е предвидено в клиничната информация).

Подходът основан на полуструктура се използва за интегриране на всички разнородни данни. По този начин се извършва хоризонтална интеграция на данните. За вертикална интеграция на данни се използват два слоя от полуструктурата - за всеки вид данни (съдържащи само атрибути, които съществуват във всеки запис) и за всички видове данни (съдържащи ID и Sample ID). По този начин се създава мрежа от връзки между всички видове данни, за да ги управляваме.

Интегриране на данните. Както беше отбелязано по-горе, по дефиниция интеграцията на данни е процес на комбиниране на данни от различен тип и от различни източници и обединяването им в значима и ценна информация. За интеграция на данни използваме новосъздадената мрежа от отношения. В тези мрежи възлите представляват пациенти, а ръбовете представляват прилики между профилите на пациентите. Сходството означава, че двама пациенти са свързани помежду си от множество протеини. Тези мрежи от връзки могат да бъдат приложени към групи пациенти и да се свържат тези групи с различни клинични характеристики. Мрежата има два слоя. Първият слой, обхващащ вътрешни взаимоотношения, е изграден от отношения, генерирани от сурови данни. Суровите данни съдържат описание на всеки пациент, със свързани данни за експресията, варианти на брой копия и клинична информация. Тази информация се трансформира във взаимоотношения между пациентите и

експресираните протеини. Вторият слой се основава на семантично свързани данни от външни източници на знания. Тези източници предоставят информация за допълнителни протеини, свързани със съществуващите в набора от данни. Тези нови отношения се съхраняват в създадената графична база данни. За да се използва допълнителната информация от външните източници на знания, те се свързват в мрежата чрез хипервръзки (URL адреси). По този начин може да избегне визуална неразбираемост, която би била причинена от излишък на информация. Тези два слоя се комбинират в една мрежа с различни тегла за всяка връзка. Представеният в дисертацията подход към интеграцията на данни се състои от няколко стъпки (Фигура 3.7.).



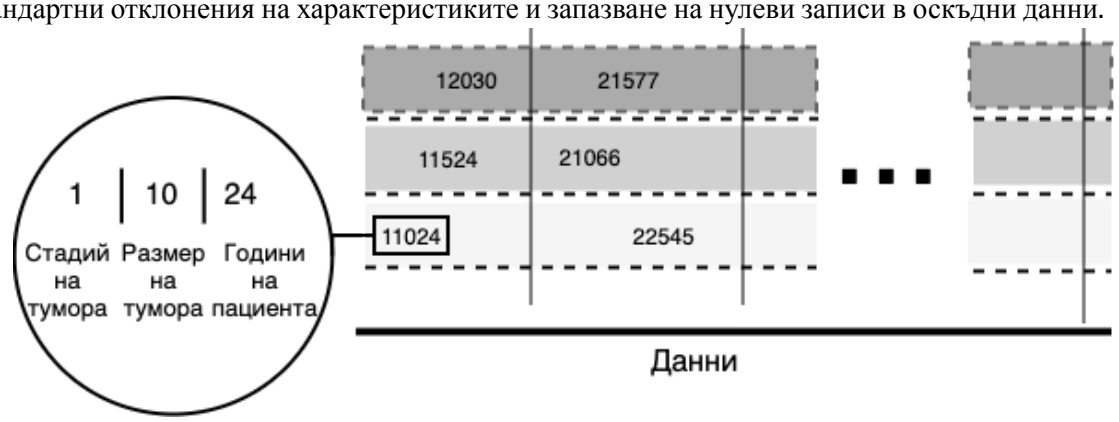
Фигура 3.7. Архитектура на предложеният подход за интегриране на данни

Модел за прогнозиране на преживяемост на пациенти болни от рак

Интелигентните системи за интегриране на биомедицински данни имат два основни компонента, отразени в дисертацията: семантично интегриране на данни и модели за извличане на знания от тези интегрирани данни. В тази връзка бе разработен модел, който използва семантично интегрираните данни и прогнозира развитието на заболявания, на примера на рак на гърдата.

За целите на настоящето проучване е разработен нов и универсален прогнозен параметър - интегрирана с тумор клинична характеристика (TICF). Тази характеристика се изгражда чрез числено обединяване на туморния стадий, размера на тумора и възрастта при поставяне на диагнозата (Фигура 3.8) в точно този ред. Редът на обединяване на тези клинични параметри е важен поради класирането на клиничната информация за развитието на тумора и неговото значение за степента на преживяемост на пациента. По-конкретно, пациент с тумор в етап четири, ще има по-кратко време за оцеляване в сравнение с пациенти с тумор в етап два. Следващата характеристика - размерът на тумора - се добавя на второ място, защото с увеличаване на размера на тумора степента на оцеляване на пациента се намалява. Размерът на тумора е втората характеристика и защото той е по-малко важен за времето на оцеляване от етапа на тумора. Третата използвана характеристика е възрастта към момента на поставяне на диагнозата, при която по-възрастните пациенти имат по-нисък процент на оцеляване. Ако редът на обединяване на тези компоненти, съставящи TICF, ще се различава, пациентите с отдалечени характеристики, свързани с оцеляването, ще бъдат неправилно групирани. По този начин се осигурява нормализирано разстояние между пациентите, което е от съществено значение при следващите ни подходи за ML за прогнозиране на времето за оцеляване.Приложеният подход за нормализиране се основава на стандартното отклонение на тренировъчния набор, за да може по-късно да приложи същата трансформация върху тестовия комплект. Алтернативна стандартизация се основава на мащабиране на характеристиките, целящи да ги получат между дадена минимална и максимална стойност, често между нула и единица, или така, че максималната абсолютна стойност на всяка характеристика да се мащабира до размера на

единицата. Мотивацията да се използва това мащабиране включва устойчивост към много малки стандартни отклонения на характеристиките и запазване на нулеви записи в оскъдни данни.



Фигура 3.8. TICF прогнозен параметър

Следващият етап в тази методология използването на ML модели, за предсказване на времето за оцеляване и тяхното оценяване. Използваните модели ML включват поддръжка на регресия на базата на опорни вектори (SVR) с различни ядра: RBF, линейна и полиномиална, както и Lasso регресия, Kernel Ridge регресия, К-квартална регресия, дърво на решенията и многослойна рецепция (MLP) регресия. По-формално, машинно-базиран метод на опорен вектор конструира хиперплан или набор от хиперплоскости в пространство с висока или безкрайна размерност, които могат да се използват за класификация, регресия или други задачи като откриване на извънредни точки. Интуитивно, добро разделяне се постига от хиперплана, който има най-голямото разстояние до най-близката точка от данните за обучение от всеки клас (т.нар. Функционален марж), който носи по-ниска грешка при обобщаването на класификатора. Използваме също модел на стохастичен градиент спускане (SGD) - алгоритъм за обучение на широка гама от модели в ML, включително (линейна) SVM, логистична регресия и графични модели. Когато се комбинира с алгоритъма за обратно размножаване, това е де факто стандартният алгоритъм за обучение на ANN.

Използване на машинно обучение за оценка точността на предсказване на протеинови структури

В тази работа са приложени ML модели за оценка на точността на двете свойства, КВ енергия и вероятност, които могат да бъдат използвани като функции за точкуване. Имайки предвид, че вероятността е по-точна мярка за приспособяването на структурата на последователността. Целта на това проучване е да потвърди това с помощта на ML модели. Като идеята е да се провери дали моделните прогнози за стойностите на вероятността ще бъдат по-точни от тези на енергията на КВ. Това ще покаже, че вероятността предоставя възможност за по-добро използване на структурна информация при прогнозиране от енергията на КВ.

Кръстосаното валидиране разделя данните за обучението на няколко несвързани кохорти с приблизително еднакъв размер. Всяка кохорта се използва от своя страна като данни за тестване, докато останалите кохорти се използват като данни за обучение. След това се прилага моделът за прогнозиране, изграден върху данните за обучение, за да се предскажат етикетите на класовете на данните от теста. Този процес се повтаря, докато всички кохорти не бъдат използвани като данни за тестване веднъж, а след това точността на прогнозите на всички слепи тестове се комбинират, за да се получи обща оценка на ефективността.

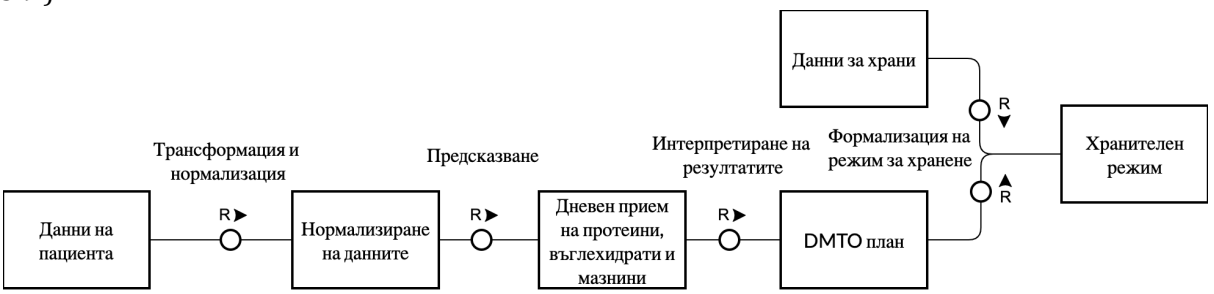
Методология за оценка на точността на енергията. За целите на това проучване е разработен ML-управляван подход за оценка на точността на енергията (E) и вероятността за базиране на честотата (L) за прогнозиране на структурата на протеина, основана на знанието. И двата подхода се основават на статистически данни за скритите / изложени свойства на остатъците. Последователността и моделът на всеки един от 245 протеинови обекта се трансформират в числови стойности, които могат да се използват като параметри в ML модели.

За да се предскажат стойностите на енергията и вероятността на КВ, бяха използвани три контролирани ML модела на предсказване. Избраните модели са от пакета на python scikit-learn: 1) Lasso - linear_model (алфа = 0,1), 2) Регресия на най-близките съседи (NNR) - kNeighborsRegressor (n_neighbors = 5, алгоритъм = 'kd_tree'), 3) Регресия на дървото на решенията (DTR) - DecisionTreeRegressor (max_depth = k). За всеки един от моделите се използва k-кратна кръстосана проверка за разделяне на набора на k по-малки набори за по-добра оценка.

Свързани данни на базата на онтологии с цел създаване на съветстваща система в областта на здравеопазването.

В настоящото проучване принципите на свързаните данни се прилагат в предложената методология, насочена към използване на цялата съществуваща информация, заедно с наличните онтологии, разработени специално за болестите, които представляват интерес, като DMTO, който е новосъздадена онтология на OWL 2, съдържаща усложнения, свързани с диабета , симптоми, лекарства, лабораторни тестове. Новите актуализации на DMTO се отнасят до опциите за персонализиране на плановете за лечение на пациенти с диабет тип 2. Данните за пациентите, използвани в проучването, са от записи на клинични лабораторни тестове и представеният резултат може да служи като неразделна част от оперативно съвместима информационна система за медицинска документация.

Създадена методология. Използва се онтологията DMTO за управление и съхраняване на данни за пациента, както и за подпомагане създаването и предлагането на диетични планове въз основа на лабораторните тестове на пациента. Базата от знания по предмета се прилага с помощта на DMTO и SWRL правила. Добавят се специфични правила за изчисляване на количеството и пропорциите на макронутриенти, които пациентът трябва да приема въз основа на резултатите от лабораторните тестове. Тези правила се използват за изготвяне на планове за диетично лечение чрез официална информация и знания. След като се импортира набор от данни за пациента, се извършва подходяща формална аргументация и се генерира план за лечение (Фиг. 3.9.).



Фигура 3.9. Процес на генериране на план за диета.

Първата и най-важна модификация, специфична за пациента, е промяната на типа от RDF / XML на OWL / XML, за да може да се създадат свойства на обект has_lifestyle_participant и has_breakfast_meal. RDF / XML е сериализационен синтаксис за RDF графиките. OWL / XML е синтаксис за сериализация за структурна спецификация на OWL 2. RDF / XML онтологиите не могат да се представят правилно с помощта на стандартни XML инструменти. Освен това е имало желание за по-редовен и по-прострен XML формат. Ето защо е изобретен OWL / XML.

Следваща значителна модификация е свързана с разширяването на „профила на пациента", за да има повече от един лабораторен тест. DMTO е изграден като набор от модули. Тези модули бяха внедрени от нулата или внесени от други добре познати онтологии. За да се постигне персонализирана диета, различни пропорции между мазнини, въглехидрати и протеини се задават като диетични параметри за всяко хранене (в момента само за закуска). За да се установи дали пациентът има резултати от лабораторни тестове в рамките на нормалните или извън нормалните граници, се определя набор от правила, които проверяват това и също така

определят необходимото съотношение между мазнини, протеини и въглехидрати според резултатите от лабораторните тестове. Съотношението между мазнини, въглехидрати и протеини се определя за храненето в диетата на пациента.

Предсказване на антимикробна резистентност в метагеномни данни.

Прогнозиране произхода на пробите. Мащабни метагеномични изследвания [128]-[131] са част от глобална инициатива за изследване и разбиране на микробиомното разнообразие. Скрининг с висока производителност, като например последователности от цели геноми идентифицира генетичната информация до по-подробни нива като нивото на видовете и може допълнително да открие изобилие от еукариоти, гъби и вируси. Повечето методи за анализ на данни за метагеномни последователности се основават на контролирани техники за машинно обучение [132],[133]. Повечето от тези модели се ограничават до предсказване на проби от места.

Класификацията на пробите по произход обикновено се извършва чрез контролирани методи за машинно обучение, които включват разделяне на пробите в комплекти за обучение и тестване. В настоящата работа е направен предварителен преглед на някои от добре познатите методи след което бе решено изследването да се съсредоточи върху три от тях, които не включват много параметри и са по-леки за работа, но достатъчно информативни в рамките на R-project. По-специално бяха използвани Gradient Boosting Machine (GBM) [137], Случайна гора (Random forest) [138] и Невронна мрежа (NNet) [139]. Приложените модели за машинно обучение, бяха използвани за предсказване на кой континент и на кой град принадлежат пробите.

Оценка на относителния риск с помощта на пространствено моделиране. Пространствената автокорелация се използва много често, когато наблюденията, които са близо в пространството имат сходни стойности. Част от тази пространствена автокорелация може да бъде моделирана от известни ковариантни рискови фактори в регресионен модел, но е обичайно пространствената структура да остане в остатъците след отчитане на тези ковариантни ефекти. Тогава пространствени модели като Байесови йерархични модели се използват за разширяване на линейния предиктор с набор от пространствено автокорелирани случайни ефекти в зависимост от структурата на квартала на географските области. Случайните ефекти обикновено се представят с условна авторегресия (CAR) [140], която предизвиква пространствена автокорелация чрез структурата на съседство на ареалните единици. Такива модели обикновено се използват в епидемиологията, например проучвания за картографиране на болести [141], но са сравнително нови в областта на метагеномиката.

Компресия на омикс данни

При интегрирането на биомедицински данни основен въпрос е начина на трансфер на данни с голям обем. Това е особено важно при така наречените -омикс данни и всякакви други данни биомедицински произход, които както беше изложено по-горе в дисертацията хетерогенни по тип, формат, произход и време на генериране. Особено приоритет в науката за данните с биомедицински профил се е отдавал на начина за компресирането им, както като метод, така и като софтуерно решение.

Предложеният подход за компресиране на биомедицински омикс данни в се основава на кодиране и по-точно на алгоритмите Shannon-Fano и Huffman. Основната идея зад алгоритмите на Shannon-Fano и Huffman е да се зададе двоичен код, който да съответства на всеки от символите във входното съобщение (в случая бази на входната последователност).

Създаване на оптимален код. Оптимизиран алгоритъм, разработен в изследването, според който последователностите, базирани на теорията за оптималното кодиране на букви, ще бъдат компресирани, има следните оперативни характеристики:

1. Разпознаване на типа последователност (РНК или ДНК).
2. В зависимост от последователността се избират предварително дефинирани двоични стойности за всеки от символите, използвани в последователността - използването на предварително дефинираните стойности запазва следните процедури, необходими за изграждане на оптимален код за всяко произволно съобщение в алгоритмите на Шанън-Фано и

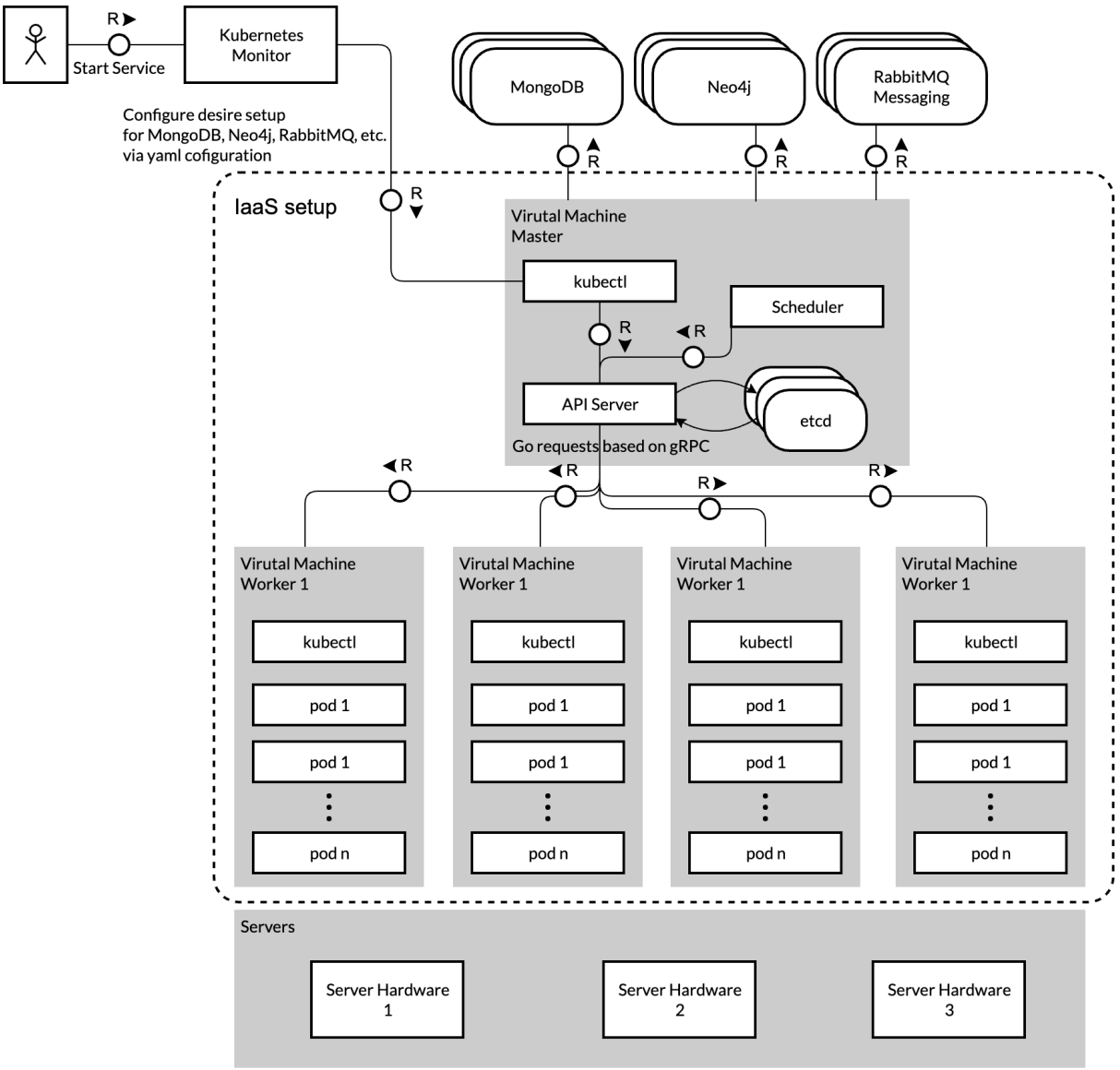
Хъфман: а) Наборът от символи на източник A е подреден с цел намаляване на вероятността от поява на съобщение - pj б) Наборът от вероятности Pi е разделен на две групи (p1, p2, ..., pj) и (pj + 1, pj + 2, ..., pm), така че разликата, показана по-долу, е минимална: в) Символите, чиито вероятности са в първата група, се присвояват с r-та кодова буква 0, а тези от втората група - с r-та кодова буква 1. г) За едноелементна вероятностна група процедурата е завършена и за всяка от другите групи в многоетапната процедура елементите са номерирани от 1 до m и преминават рекурсивно (r = r + 1) към стъпка 2 .

Този подход елиминира недостатъците както на алгоритмите на Shannon-Fano, така и на Huffman за декодиране и защита на информацията. В предложения оптимизиран подход всеки от символите ще бъде кодиран с точно 2 бита. Както преди компресията, всеки знак заемаше 8 бита, тук е първото място, където имаме компресия 1: 4. След прилагане на алгоритмите може да се постигне компресия до 1: 400.

Глава 4: Софтуерна реализация на интелигентни системи за интегриране, анализ и класификация на биомедицински данни

Изградена архитектура за обезпечаване на хардуерните и софтуерните нужди на системата. Всички изградени системни решения в този труд се базират на една и съща интелигентно изградена структура за хардуерно и софтуерно обезпечаване. Системата за интегриране на хетерогенни данни е базирана на дистрибутирането на данните в множество бази от данни разпределени на множество хардуерни устройства. Използвани са две системи за управление на бази от данни MongoDB версия 4.4 и Neo4j версия 4.2. За връзка с всяка една от базите съответно се използват предоставеният от официалният дистрибутор драйвър. Всички реализирани софтуерни продукти се инсталират на авторски разработена платформа базирана на Kubernetes [C1].

На фигура 4.2 е изградена система, която позволява да се стартират и поддържат приложения, за който има създадено изображение на контейнер (image). Съответно са изградени необходимите изображения на всички използвани в дисертацията услуги. Изградени са за MongoDB, Neo4j, RabbitMQ,, Go базирани приложения, Python базирани приложения и Java базирани приложения. Всяко едно от приложенията работи задължително на минимум 2 инстанции. Това позволява дори някой от сървърите да спре да работи поради външни фактори като загуба на връзка, изгорял носител на данни или други причини приложението да продължи да работи. Създаден е специален контролер в Kubernetes, който позволява веднага след като някой от контейнерите е спрян поради някаква причина да не се опитва да го създава наново на същият сървър както е по подразбиране, а да го създава на друг сървър. Това позволява много по - бързо възстановяване на приложението към нормално работно състояние. При използване на бази от данни базирани на контейнери синхронизацията между тях и репликацията е много важна, за да може наистина ако един контейнер спре базата от данни да продължи да работи.



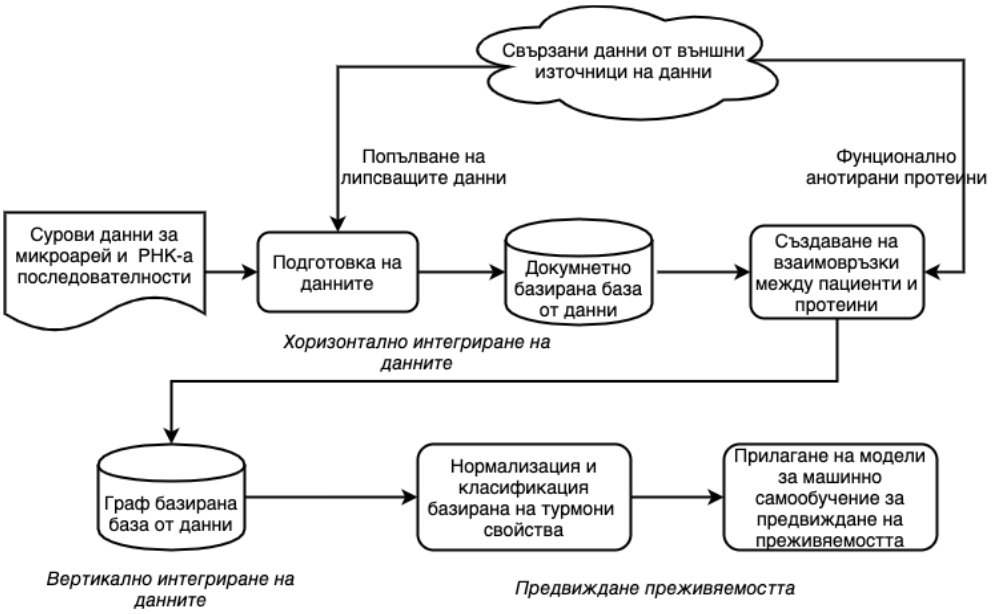
Фигура 4.2. Архитектура на разработената система за предоставяне на хардуерни ресурси използвайки три хардуерни сървъра.

Софтуерен модул за интегриране на хетерогенни данни и машинно обучение за предвиждане преживяемостта на пациенти болни от рак

Използвайки полуструктурата, хетерогенните данни се интегрират в една база от данни, където крайната цел е да се създаде мрежа от връзки между всички видове данни. В тази мрежа възлите представляват пациенти, а ръбовете представляват прилики между профилите на пациентите. Сходството означава, че двама пациенти са свързани помежду си от множество протеини, въз основа на експресионни профили и промени в броя на копията. Тези мрежи от връзки улесняват групирането на пациентите. След това групите пациенти могат да бъдат свързани с различен клиничен резултат.

Мрежата има два слоя. Първият слой, обхваща вътрешните взаимоотношения изградени със сурови данни, т.е. клиничната информация, данни за експресия и варианти на брой копия. Те се трансформират във взаимоотношения между пациенти и протеини. Вторият слой включва семантично свързани данни от източници на знания от външни домейни. Тези източници предоставят информация за допълнителни протеини, свързани със съществуващите в нашия набор от данни. Тези нови отношения се съхраняват в нашата база от данни, базирана на графики. За да се използва допълнителната информация от външните източници на знания, те се

свързват в мрежа чрез хипервръзки (URL адреси). По този начин се избягва визуална неразбираемост, която би била причинена от излишък на информация. Тези два слоя се комбинират в една мрежа, където всяка връзка се претегля.



Фигура 4.3. Работен поток на интегриране и анализ на данни от ракови заболявания

Всички данни от експерименталните набори от данни са интегрирани хоризонтално с технологията NoSQL (MongoDB) и са представени като полуструктура. Това води до полуструктура за всеки тип данни, т.е. всички клинични данни са обединени в полуструктура, всички данни за експресия в друга полуструктура и всички данни за номера на копията (CNV) в полуструктура. Всички данни и метаданни се съхраняват в MongoDB в JSON формат. За вертикалното интегриране на данните, първо трябва да намерим връзки между вече изградените полуструктури за клинични записи, експресивни профили и данни за номера на копията. Тези взаимоотношения се управляват в граф базираната база данни - Neo4j. Например, пациент A с полуструктура {ID, [attributes]} е свързан с пациент B с полуструктура {ID, [attributes]}. В тази връзка идентификаторът е важният ключ, докато атрибутите предоставят обща информация за типа запис на данни (клиничен, израз, номер на копие). Такива отношения улесняват изграждането на индивидуална мрежа за всеки изследван пациент. Тази мрежа включва експресионни профили, номер на копие и мутирали протеини. По този начин можем да открием и свържем всички пациенти чрез специфичен набор от експресирани и мутирали протеини.

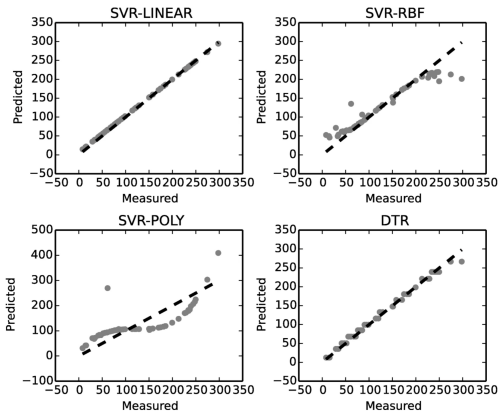
Използване на отдалечени източници за свързани данни. Чрез интегриране на семантични данни, по-специално чрез крайни точки https RESTFul (програмиране на точки за достъп), ние можем да намерим допълнителни връзки между протеини от външни източници на знания за домейна (External Domain Knowledge Sources - EDKS), като генна онтология (GO), UniProt, Ensembl. Чрез EDKS могат да бъдат открити протеини, които са тясно свързани с наличните в експресионните профили.

Разработен е нов и универсален прогнозен параметър - интегрирана клинична характеристика на тумора (TICF). За да се предскаже времето за оцеляване на пациента (и в двете проучвания за рак, взети заедно), са избрани специфични информативни клинични характеристики. Тествани са различни функции и техните комбинации, и ред, и емпирично е установена оптималната настройка показана на Фигура 3.8. В глава 3. По-конкретно, характеристиката на TICF се изгражда чрез числено обединяване на туморния стадий, размера на тумора и възрастта при диагностициране (фиг. 3.8) в този точен ред. Редът на обединяване на клиничните данни също показва значението на клиничната информация за развитието на тумора и значението за степента на преживяемост на пациента.

Предвиждане на преживяемост на пациенти болни от рак. Моделите за машинно обучение, използвани в разработеният подход, се основават на Support Vector Regression (SVR) с различни ядра: Radial Basis Function (RBF), Linear и Poly(nomial), и модел на регресия на дърво на решенията (Decision Tree Regression - DTR). Доказано е, че подобни модели се представят добре за прогнозиране на оцеляването в проучвания за рак [144],[145]. Освен това използването на тези модели помага за безпроблемната кръстосана валидация на резултатите.

Резултати софтуерната реализация на моделите. Разработен е нов модел за интеграция на данни, базиран на мрежа, където се комбинират клинични и молекулярни данни, като се използват както записи от сурови данни, така и външни източници на знания. Връзките, получени от суровите данни, представляват вътрешната мрежа, а връзките, базирани на външни източници на знания за домейна (EDKS), са представени като семантично свързана мрежа. Изградената семантично свързана мрежа е свързана с EDKS чрез крайни точки на достъп базирани на RESTFul API/s. След като наборът от данни се нормализира и пациентите се разделят на групи, се прилагат няколко модела за машинно обучение за прогнозиране на времето за оцеляване: Поддръжка на векторна регресия (SVR с RBF, линейни и полиномиални ядра), както и регресия на дървото на решенията (DTR).

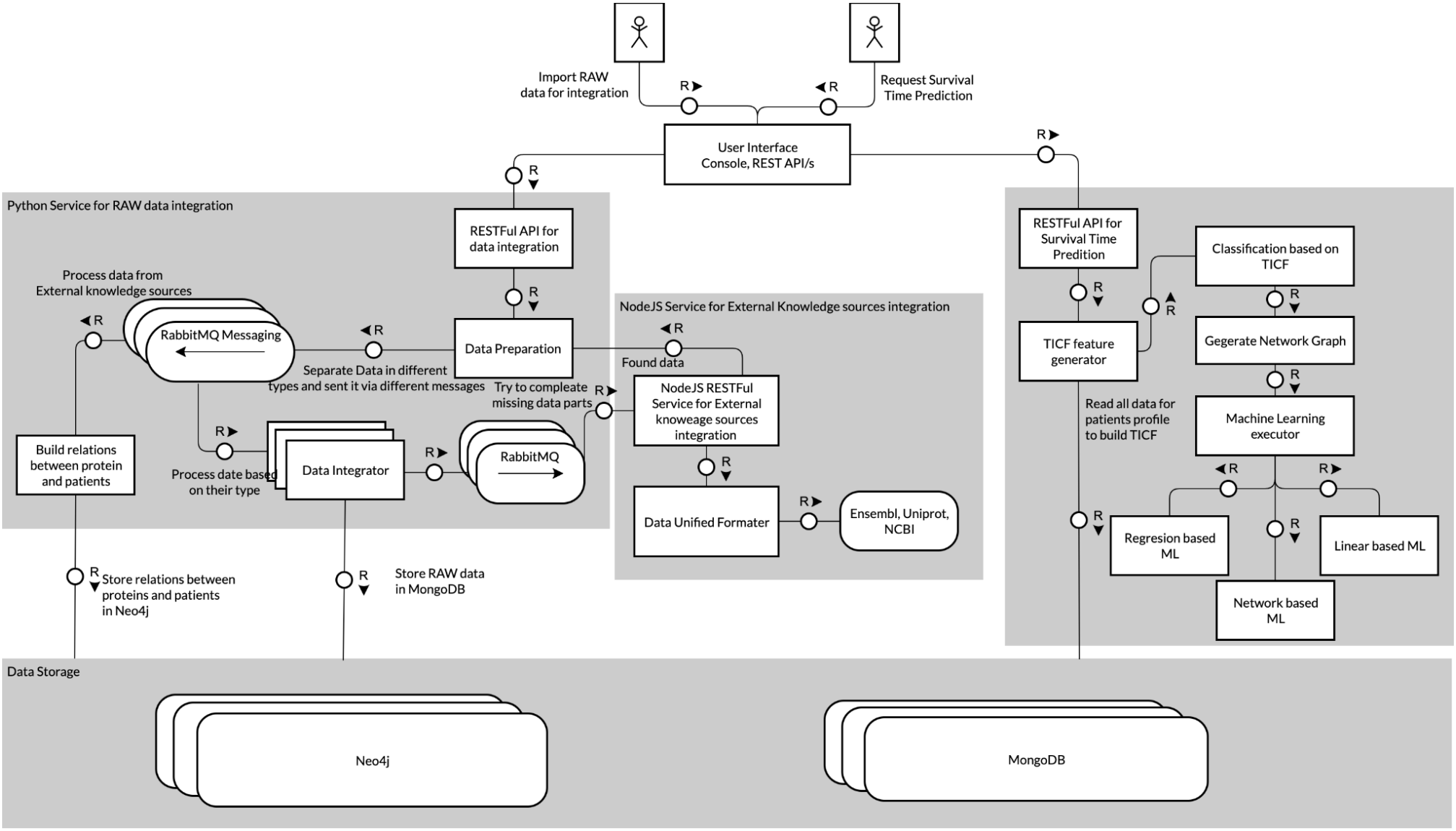
Времето за оцеляване се прогнозира, като се използват данните и за двата вида рак, комбинирани с изградения модел използван за обработка и интегриране на данните. DTR и SVR-Linear се представят най-добре, като SVR-Linear дава най-точните резултати за прогнозиране на времето за оцеляване. Потенциалът на тези модели е в подобряване на точността на прогнозирането на времето за оцеляване чрез итеративно подобряване на набора от данни за обучение в целия интегриран набор от данни. По-конкретно, с всеки нов проучен пациент, за който се използва модела реално се обогатява набора от данни за обучение с нови доверени отношения определени от увеличената честота на тяхното използване.



Фигура 4.5. Степен на успех на моделите за машинно обучение за прогнозиране на времето за оцеляване.

Прогнозираните и измерени стойности съответстват спрямо TICF на прогнозата в месеците за оцеляване. Пунктирана линия символизира идеалния случай на прогнозираното съотношение към измерения TICF за прогнозиране на времето за оцеляване. Сравнява се производителността на три модела (SVR-Linear, SVR-RBF и DTR). Четвъртият модел, SVR-Poly, показва по-лоши резултати (Фиг. 4.6). Кръстосаното валидиране се основава на 5 подгрупи, дефинирани от свойството TICF, получени след използване на k-fold алгоритъм.

Софтуерно обезпечаване на реализираната система. Системата е реализирана посредством изградената инфраструктура, представена в раздел 4.1. Върху тази инфраструктура са разработени следните услуги. Две бази от данни MongoDB и Neo4j, както и множество услуги на Python, NodeJS и RabbitMQ (Фигура 4.7.)

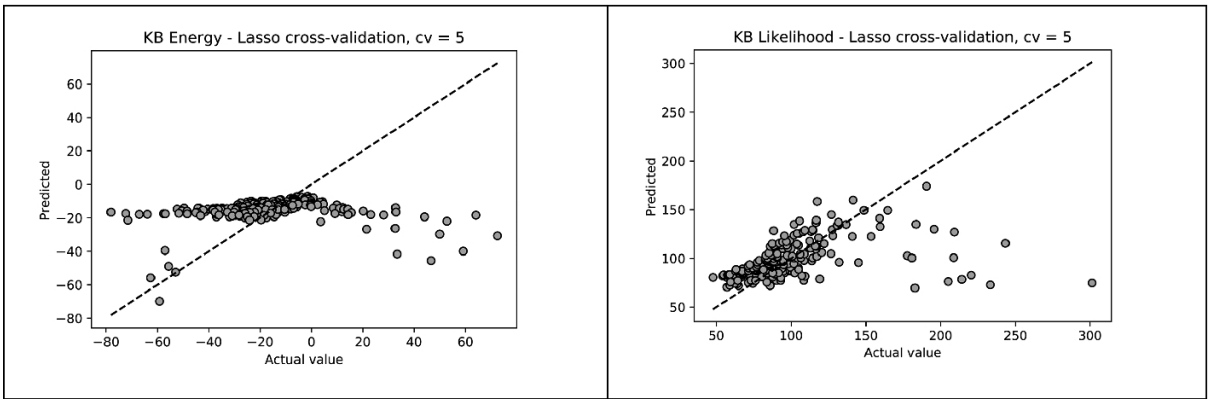


Фигура 4.7.Софтуерна архитектура на изградената система

Модулът за предварителна обработка и интегриране на сурови данни е изграден на Python версия 3.7. Реализирани са множество програмни модули за интеграция на данни с различни формати JSON, CSV и XML. Това са форматите, с който работят повечето ресурси с общодостъпни данни в биоинформатичният свят като NCBI, Ensembl, UniProt и други. Интеграцията е реализирана на принципа на безкраен низ от редове, които трябва да бъдат описани и анализирани. От потребителят се подават файлове посредством т.н в HTTP 2.0 протоколът “stream”, което позволява един файл или ресурс да бъде изпратен на части. Това от своя страна позволява на сървърната част за започне да обработва файла без той да бъде изпратен целият. От друга страна в биоинформатиката файловете със секвенционни данни са с доста голям обем от порядъка на над 100 мегабайта за един файл. Без да се използва подобен подход за разделяне на файловият поток на малки части за обработка би довело до ограничаване на максималният брой паралелни заявки, които могат да се обработват наведнъж.

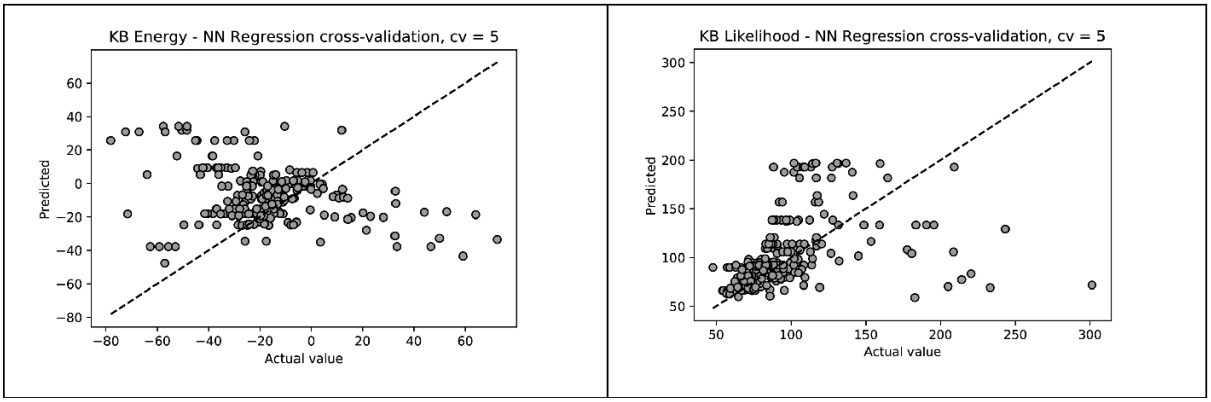
Софтуерни решения и резултати за предсказване на протеинови структури и оценка на точността .

Методологията на това проучване включва няколко етапа, като интегриране на необходимите данни от EKDS, препроцесинг на данните, анализ и класификация на данните основан на машинно обучение. След като препроцесинг и нормализиране на набора от данни, се прилагат три регресионни ML модела: ласо регресия, регресия на най-близкия съсед и регресия на дървото за решения. Тестваме стратегията за кръстосано валидиране на разделяне на k = 3, k = 5 и k = 7 пъти, за да се направи съпоставка на моделите по отношение на тяхната точност на прогнозиране на резултатите от енергията на КВ и вероятността. Графиките на фиг. 4.9 , 4.10 и 4.11 показват връзката на действителните с прогнозираните стойности на всеки конкретен модел, използван за КВ енергия и съответно за вероятност с кръстосано валидиране (cv) k = 5.

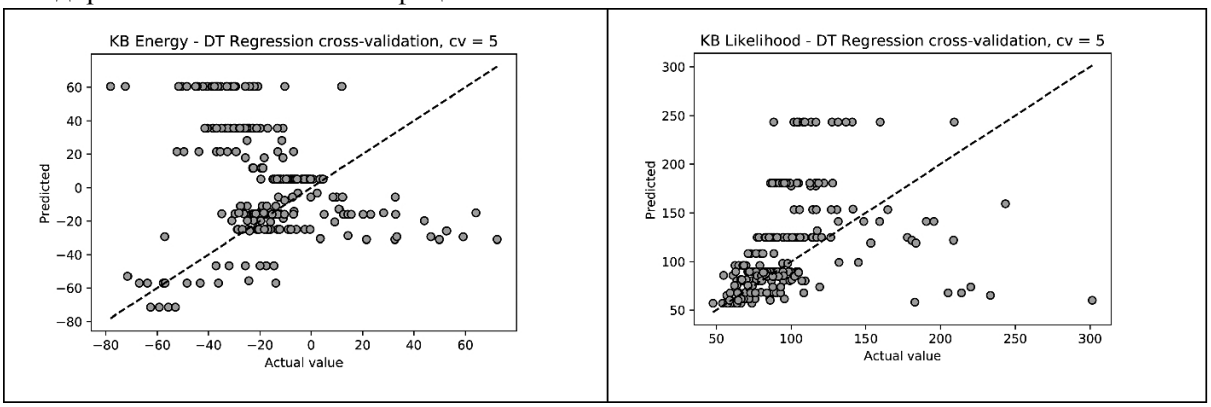


Фигура 4.9. Използване на модела за машинно обучение LASSO за валидиране със стойност на коефициента к=5

По отношение на енергията базирана на знания (КВ), ласо има по-лоши прогнозни резултати, отколкото по вероятност, за разпределяне на резултатите около регресионната линия, с много малко отклонения от по-голямата действителна стойност на вероятността. Резултатите от NNR са подобни, като енергийните оценки на КВ са по-разпръснати от стойностите за вероятност. DTR дава донякъде подобни резултати за прогнозиране на енергията на КВ и стойностите на вероятността. Тези резултати са доказателство, че използването на прогнозата за вероятност е по-добро от прогнозата за енергията на КВ. Тези резултати потвърждават аналитичния изчислителен подход на констатацията на оптимизацията, че вероятността е по-добра от енергията на КВ като скоринг функция.



Фигура 4.10. Използване на модела за машинно обучение “Nearest neighbor regression” за валидиране със стойност на коефициента к=5



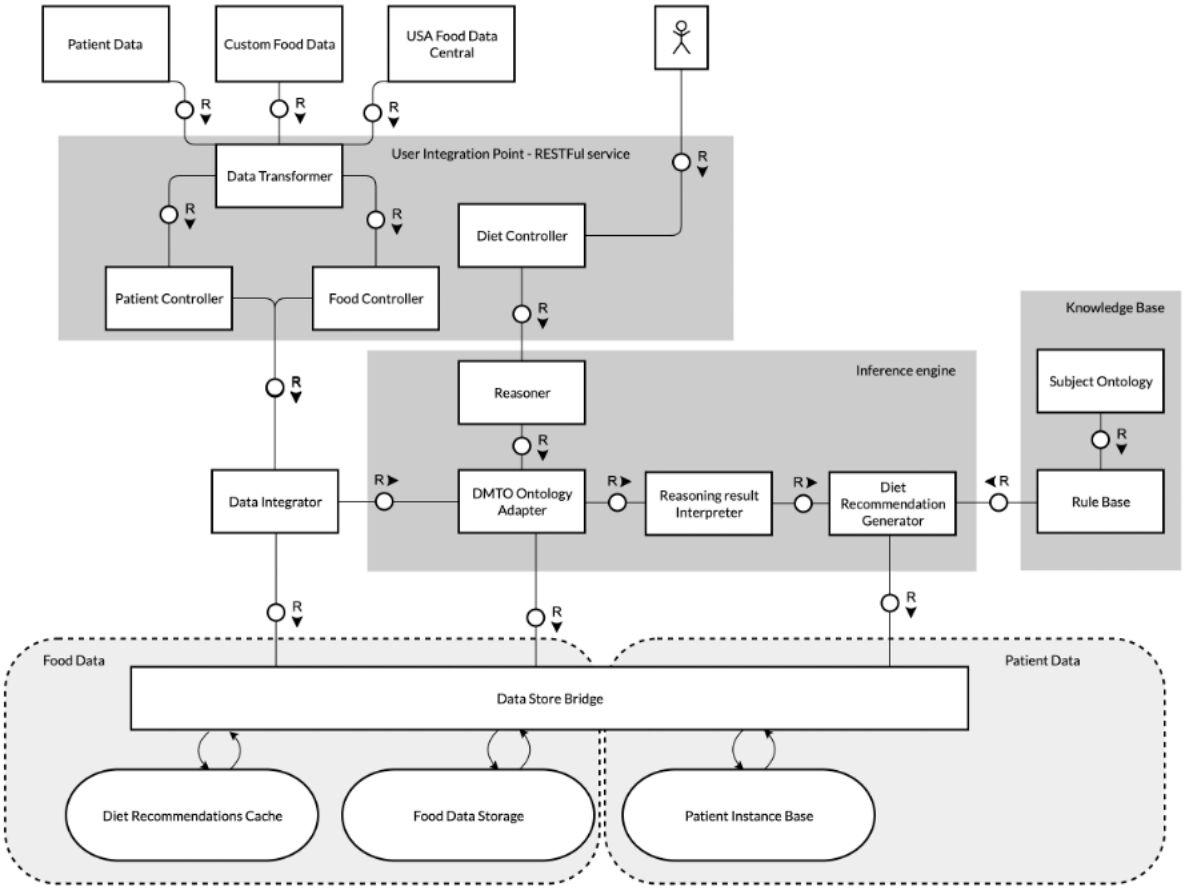
Фигура 4.11. Използване на модела за машинно обучение “Decision tree regression” за валидиране със стойност на коефициента к=5

Софтуерни решения и резултати с цел създаване на съветваща система за консултиране на диети при болни от диабет.

Системна Архитектура: Архитектурният план и общия функционален план на разработената DSS са представени на фиг. 4.15. Предложената система (фиг. 4.15) се състои от следните части: входната част (данни за пациента, данни за храните), точката за интеграция на потребителя с услугата RESTFul API, последващият интегратор на данни, база от знания и съответния механизъм за извод и част за съхранение, включваща: кеш за препоръки за диета, съхранение на данни за храни и база от пациенти. Сървърът за крайна точка на потребителска интеграция е проектиран за целите на интеграцията на данни, нормализирането на данните и разработването и приложението на интерфейса. Входната част се основава на пациентски данни и някои необходими данни за храната за генериране на препоръки за диета. Моделът на данните за пациентите се определя от показателите от амбулаторните тестове и историята на заболяването (анамнезата) като част от здравната карта на пациента. Тези данни за пациентите включват цялата информация от лабораторните тестове и основните му компоненти като гликиран хемоглобин, глюкоза, холестерол, пикочна киселина. От данните за пациентите се създават екземпляри, които се записват в DMTO.

Основният компонент на разработената DSS за препоръки за диета е базата от знания за предметната област, чието ядро е DMTO. По-точно, базата от знания на DSS се състои от две основни части - разширяемо копие на DMTO и набор от SWRL правила, описващи конкретни знания за анализ на данните и вземане на решения. Системата има разработен приложен интерфейс като крайни точки за достъп до сървъра базирани на RESTFul API, позволяващи на потребителя да добавя данни за пациента. Всеки добавен набор от амбулаторни записи е свързан с определен пациент. За всеки нов пациент се генерират неговите идентификатори (GUID) и профил. Амбулаторните тестове са свързани с профила на пациента чрез свойството has_lab_test

на съответния екземпляр на пациента. Когато амбулаторните записи се добавят в DMTO, промените се запазват и генерираните идентификатори на всички нови екземпляри на пациенти се връщат.



Фигура 4.15. Архитектура на разработената система

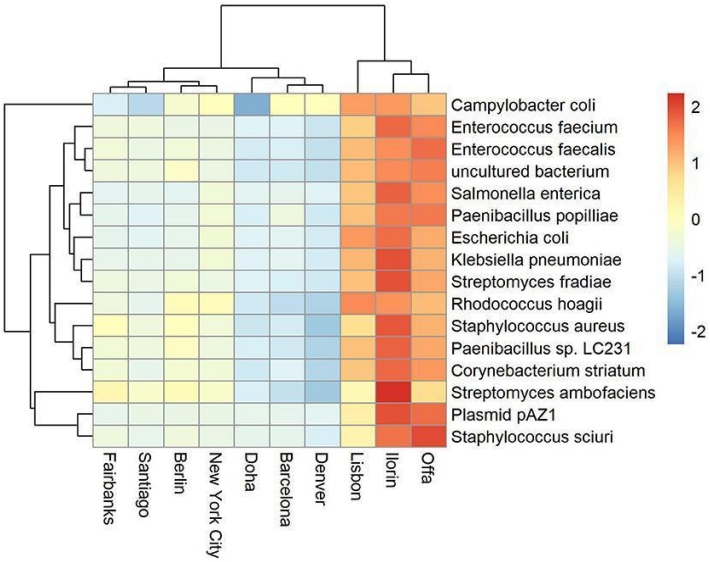
Изменения на DMTO (обогатяване). DMTO е изчерпателна онтология и осигурява най-голямо покритие и най-пълна картина на кодирани знания за текущото състояние на пациентите с T2DM, предишните профили и свързаните с T2DM аспекти, включително симптоми, амбулаторни тестове, усложнения, взаимодействия, свързани с глюкоза заболявания и лекарства , и рамки на план за лечение.

Основният принос на настоящото проучване е подходящо разширяване на DMTO и разработването на модел за DSS за диетични препоръки относно аспектите на T2DM и подпомагане на създаването на потенциални планове за пост-клинично лечение.

Реализиран модел и софтуерно решение за интегриране, класификация и анализ на метагеномни данни

Резултатите по тази част от дисертацията са в резултат на работа по проектите MetaSUB и CAMDA. Както беше отбелязано по-горе в работата целта на тези проекти е изследване на микробиалното, бактериално и вирусно разнообразие, както и класификация и оценка на антимикробната резистентност на изследваните общества от микроорганизми. Данните са от метагеномно паралелно (NGS) секвениране и представляват извадки от обществено достъпни места, от транспортната система на различни градове по света, където се предполага, че разнообразието е голямо, както и нивото на антимикробната резистентност е значително.

Гените на антимикробна резистентност и съответните бактериални таксони представляват относително малка част от наличния глобален метагеномен профил. Въз основа на метагеномния класификатор Kaiju, който използва модифицирано търсене при ефективно използване на паметта на трансформацията на Burrows-Wheeler [151], се установи, че относителното изобилие от свързани с антимикробни видове представляват средно между 0 и 0,33 от общото количество секвенционни данни. Някои градове показат по-голямо разнообразие и брой като Fairbanks (max 0,28), Лисабон (max 0,2), Ilorin (max 0,33). Най-често срещаните антимикробни таксони са показани на фигури 4.17. Един от клъстерите включва Salmonella enterica, Staphylococcus aureus и Escherichia coli, разпространени много в Офа, Илорин и Лисабон. Антимикробните гени в класовете, свързани със Streptomyces, са по-разпространени в големите градове като Лондон, Ню Йорк, Хонконг и Куала Лумпур. Пробите от Берлин, Токио, Стокхолм и Доха имат малък или нулев брой сред най-богатите антимикробни таксони.



Фигура 4.17. Разпределение на антимикробни устойчиви таксони

Някои стойности на корелации, свързани с антимикробните таксони (корелация> 0,6, p <0,01) с метеорологичните данни в градовете са: различни мерки за променливост на влажността и Vibrio parahaemolyticus; средна влажност и Campylobacter jejuni, Corynebacterium striatum, Paenibacillus sp. LC231, Rhodococcus hoagii, Streptococcus australis и Streptomyces ambofaciens; температура и Pseudomonas aeruginosa. Vibrio parahaemolyticus и P. aeruginosa показват съответно най-силни отрицателни корелации с променливостта на влажността и налягането.

Предвиждане произхода на пробите За да се предскаже произхода на пробата, се използва три често срещани техники за машинно обучение: Машина за повишаване на градиента (GBM), Случайна гора (RF) и Невронна мрежа (NNet). За да се изберат най-добрите характеристики за моделите, се прилага Рекурсивно премахване на функции (RFE). Този външен метод за повторно вземане на проби се основава на 10-кратно (10-fold) кръстосано валидиране и е повторен 3 пъти. Подходът k-fold включва разделяне на множеството от данни на k групи с приблизително еднакъв размер. Първият набор се третира като набор за валидиране и методът е валидира останалите k-1 групи, където k обикновено се приема, че е равно на 5 или 10.

Пространствено моделиране (Spatial analysis). За пространственият анализ се използват всички налични гени, за да се конволюционен модел. Пространствената корелация в градовете е направена чрез I-теста на Моран. Градове като Ню Йорк (макс. 0,44, p <0,01), Ilorin (0,38, p <0,01), Хонг Конг (0,41, p <0,01) и Тайпе (0,6, p <0,01) показват силни пространствени корелации за много от антимикробни устойчиви таксони. Данните за броя на метагеномиките често показват свръхдисперсия, тъй като те са разнородни поради различните градове и държави. Направен е тест за свръх дисперсия за 16-те антимикробни характеристики от моделите за прогнозиране, съчетавайки модел на Поасон с ковариати и обикновена регресия с най-малък квадрат, за да направи оценка на параметъра за свръх дисперсия. Резултатите показват, че всички с изключение на една от най-добрите антимикробни характеристики появяват свръх дисперсия с p-стойности доста под 0,01.

Обсъждане на резултатите и софтуерна реализация. В този труд се демонстрират трите метода за машинно обучение, а именно Gradient Boosting machine, Random Forest и Neural

Network, който имат подобна производителност за класифициране на произхода на пробите. Използвайки голяма база от данни като proGenomes, която съдържа над 80 000 анотирани бактериални и археални геноми, ние постигаме висока точност (до 80%). Разработените програмни модули генерират всички таблици и фигури, така че резултатите да могат да бъдат възпроизведени. В допълнение кодът позволява на потребителите да променят параметрите, например като използват различен набор от настройки а също така да изпълняват допълнителни методи за машинно обучение, както е предвидено в рамките на пакета, и допълнително да подобрят резултатите.

Реализиран модел и софтуерно решение за компресиране на секвенционни данни

Реализиран е нов модел и софтуерно решение за компресиране на данни от паралелно (NGS) секвениране. Методологията и архитектурата на предложеният модел е подробно обяснена в раздел три. C ++ е избран за реализация на алгоритмите за побуквено и шумозащитно кодиране, заради неговата скорост на изпълнение и по-ниско използване на паметта, в сравнение с други езици. Той също така позволява директен достъп до паметта, лесна и интуитивна реализация на побитови операции и управление на ресурсите ниско ниво. Тези предимства правят алгоритмите, написани на C++, по-оптимални в сравнение с езиците от високо ниво като C# и Java. Това е една от основните причини за включване на междуплатформени връзки между C++ и езиците от високо ниво.

Съчетаването на C++ (неуправляем код) със C# (управляем код) се реализира благодарение на .NET платформата. Още при нейното създаване е заложена идеята за независимост от средата. Изходния код не се компилира до инструкции, предназначени за конкретен микропроцесор и не се използват специфични възможности на определена операционна система, а се компилира до междинен език – Common Intermediate Language (CIL). Този език не се изпълнява директно от микропроцесора, а от виртуална среда за изпълнение на CIL кода, наречена – Common Language Runtime (CLR)

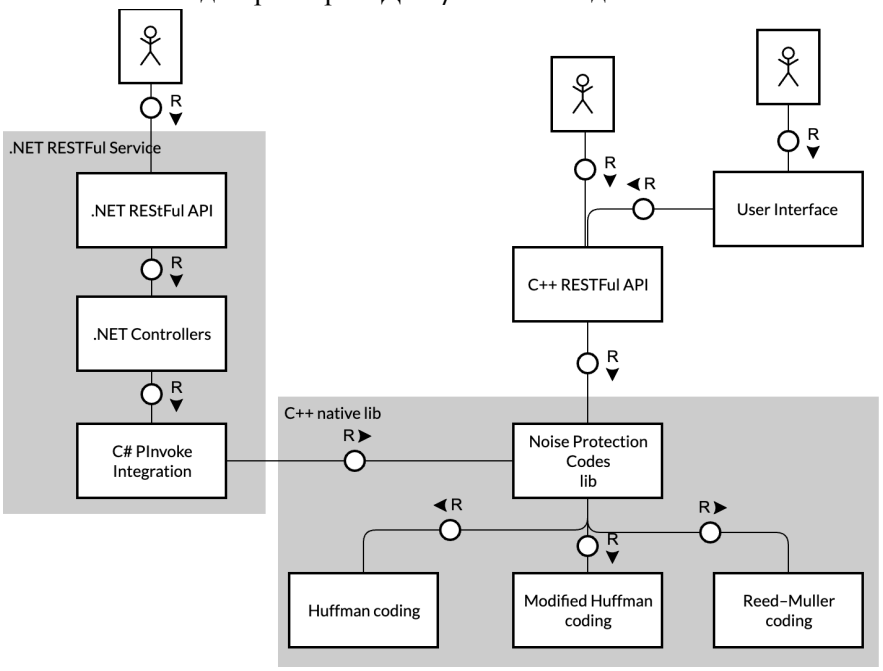
Потребителски интерфейс. Архитектурата на потребителският интерфейс е базираната на работната рамка vue.js. Тя предоставя множество компоненти за реализацията на интерфейса, като всеки един от тях е изграден на модуларен принцип и може лесно да се разширява. Входната точка е през браузър, като тестваните са Safari, Mozila, Opera, Edge, Google Chrome и други. Това е възможно понеже всеки един от компонент е написан на JavaScript по стандарта ECMA 6. След това създаденият код се компилира до HTML в зависимост от браузъра, който е изпълнил заявката. Използва се V8 софтуерната рамка разработена от Google за интерпретиране и компилиране на JavaScript до абстрактно синтактично дърво от където може да бъде компилирано като C++, C, C#, Java и други. Използва се Node.js като софтуерна среда за изпълнение на Javascript. По този начин потребителският интерфейс е изграден от сървърна и потребителска час. Изградените компоненти работят на реализираната инфраструктура в точка 4.1.

Архитектура на изградената система за компресиране на секвенционни данни.

Всички компоненти описани по - горе се обединяват в една система представена на фигура 4.27. Системата е разделена на четири основни модула. Модул с имплементацията на всички поддържани методи за кодиране включително модифицираният метод на Хъфман. Този модифициран метод има една единствена разлика с оригиналният и тя е, че за азбука на кода се избира предварително зададена таблица. Това ускорява многократно процеса за кодиране и също така избягва възможността от избор на неподходяща азбука, което води до повишаване степента на компресия. Разработените алгоритми са изцяло базирани на C++.. Всички алгоритми са с обща абстракция, което позволява много лесно да се разширява набора от алгоритми за компресия. Нямат външни зависимости към системни програмни ресурси за работа с паметта или използване на операционната система. Използвани са единствено и само вградените в C++ стандартни модули.

Изграден е C++ модул за HTTP базираната комуникация посредствум RESTFul. Това улеснява начина на ползване на библиотеката. За всеки един от алгоритмите има изградена

отделна входна точка, което улеснява работата с тях и позволява да се използват независимо един от друг. Използван е HTTP 2.0 протоколът, който позволява да се предават файловете фрагментирано. Поради начина на работа на кодиращите алгоритми единствено и само оптимизиране код на Хъфман е в състояние да компресира в реално време. Това е възможно тъй като при него не е необходимо да се прави честотен анализ за да се определи таблицата на най - често срещаните символи. Това е едно огромно предимство, чрез което се постига компресия в реално време на какъвто и да е размер на ДНК/РНК последователност.



Фигура 4.27. Архитектура на изградената система за компресия на секвенционни данни

Глава 5: Приноси и перспективи

Предложен е набор от средства за интелигентно семантично интегриране, класификация и анализ на биомедицински данни посредством използване на инструментите на информатиката и изкуствения интелект. Изградени са множество авторски информационни системи на базата на авторска платформа. Качествата на предложените средства са изследвани в детайли с помощта на методи за валидация.

Научни

1. Разработен е модел за интегриране на хетерогенни биомедицински данни базиран на архитектура с вертикална и хоризонтална интеграция посредством нерелационни бази от данни. Разработен е подход за семантична интеграция на биомедицински данни посредством използването на концепцията за свързани данни. На базата на семантично интегрираните данни е разработен модел базиран на машинно обучение за предвиждане на преживяемостта на пациенти с ракови заболявания. В модела е създаден нов интегративен параметър, чрез който е постигната по - голяма точност при преджидане на преживяемостта по сравнение с литературни данни. (3.1, 3.2, 3.3). Представено в авторски публикации [C2, C3, C4, C7]
2. Разработен е модел за оценка на точността на предсказването на нагъването на протеинови структури на базата на машинно обучение (3.4). Представено в авторска публикация [C5].

3. Разработен е модел за съветваща система на базата на семантично интегрирани данни с използване и надстройване на онтологии с цел предлагане на хранителен режим на пациенти с диабет (3.5). Представено в авторска публикация [C6].
4. Разработен е модел за класификация и анализ на антимикробна резистентност в метагеномни данни. Разработен е модел за оценка на точността на произхода на взетите проби от различни държави и континенти (3.6). Представено в авторски публикации [C8, C9].
5. Разработен е модел за компресиране на секвенционни данни в реално време посредством усъвършенстване на алгоритми за шумозащитно кодиране. (3.7). Представено в авторска публикация [C10].

Приложни

1. Разработена е платформа за предоставянето на софтуер като услуга, която е използвана за реализацията на всички системи в дисертационният труд (4.1). Представено в авторска публикация [C1]
2. Разработена е софтуерна система на базата на модела за хоризонтално и вертикално интегриране на биомедицински данни. В системата е включено и приложение на модела за предвиждане на преживяемостта на пациенти с ракови заболявания използвайки машинно обучение на базата на създаденият интегративен параметър (4.2). Представено в авторски публикации [C2, C3, C4, C7]
3. Разработена е софтуерна система на базата на машинно обучение за оценка на точността на предсказване на нагъването на протеинови структури както и тяхната класификация (4.3). Представено в авторска публикация [C5].
4. Разработена е софтуерна реализация на съветваща система базирана на онтологии за предлагане на диети, съобразени с хранителен режим при болни от диабет тип 2. (4.4.) Представено в авторска публикация [C6].
5. Реализирана е софтуерна система за интегриране, класификация и анализ на метагеномни данни относно предвиждане и класификация на разпространението и генетичното разнообразие на антимикробната резистентност (4.5). Представено в авторски публикации [C8, C9].
6. Разработена е софтуерна система за компресия на секвенционни данни в реално време на базата на надградени шумозащитни кодове (4.6). Представено в авторска публикация [C10].

Перспективи за бъдещо развитие

Представените в този труд методи за семантично интегриране, анализ и класификация на биомедицински данни посредством създаването на интелигентни системи е добра основа за бъдеща работа, в която да се обхване по-широк кръг от задачи, както и да се навлезе в по-задълбочени изчислителни изследвания.

1. Разгледаните методи за хоризонтално и вертикално интегриране на данните могат да бъдат приложени потенциално с друг тип данни извън биомедицинската информатика.

2. Разгледаните методи за семантично интегриране могат да бъдат разширени с използването на допълнителни онтологии с цел създаване на по - добра семантична свързаност между данните.

3. Разглежаданите модели за предвиждане и валидация на преживяемостта на пациенти болни от рак и протеиновите структури могат да бъдат разширени с нови методи за машинно обучение.

4. Разширяването на създадената съветваща система за хранителен режим за пациенти болни от диабет тип 2 може да бъде усъвършенствана посредством допълнително знание от медицинските центрове, както и въвеждането на повече характеристики, които да се разглеждат

при създаването на хранителен режим. Такава една система е целесъобразно да се интегрира в разработването се електронно здравно досие на страната.

5. Създадените модели за класификация и пространствено моделиране на антимикробни данни може да бъде разширено с добавянето на бази от данни за съществуващите градове, както и добавянето на допълнителна валидация, която да увеличи прецизността на моделите.

6. Създадената система за компресия на секвенционни данни може да бъде разширена добавяйки модул, който позволява интеграция с известните система за съхранение на секвенционни данни като NCBI.

Декларация за оригиналност

Във връзка с провеждането на процедура за придобиване на образователната и научна степен „Доктор“ в Софийски университет „Св. Климент Охридски“ и защита на представения от мен дисертационен труд, декларирам, че:

- Резултатите и приносите на проведеното дисертационно изследване, представени в дисертационния труд на тема „Интелигентни информационни системи в биоинформатиката: семантично интегриране, анализ и класификация на биомедицински данни“, са оригинални и че те не са заимствани от изследвания и публикации, в които нямам участие.
- В текста на този дисертационен труд не са използвани неправомерно текстове и други обекти на авторското право без да бъде посочен източника, или без да съществува разрешение или законово право за това.
- Настоящата дисертация не е прилагана за придобиване на научна степен в друго висше училище, университет или научен институт.
- Представената от мен информация – списък с публикации, копия на документи и получени по време на експериментите резултати – отговаря на обективната истина.

Публикации по темата на дисертационният труд

[C1]H. Sabev, **I. Mihaylov,** and R. Rashidov, “Distributed persistent virtual machine pooling service,” Patent: US10824461B2, Nov. 03, 2020.

[C2]**I. Mihaylov,** M. Kańdula, M. Krachunov, and D. Vassilev, “A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models,” Biology Direct, vol. 14, no. 1, pp. 1–17, 2019.
Импакт фактор = 2.913, SJR = 1.52 Scimagojr, Q1, цитирания: **11** (SCOPUS), ISSN: 17456150

[C3] **Mihaylov, I**., Nisheva, M., Vassilev, D., “Machine Learning Techniques for Survival Time Prediction in Breast Cancer,” Lecture Notes in Computer Science. Lecture in Artificial Intelligence 11089. Springer International Publishing, 2018, pp. 186–194, doi: 10.1007/978-3-319-99344-7_17.
SJR = 0.25 Scimago Jr, Q3, цитирания: **4,** ISSN: 16113349

[C4] **I. Mihaylov,** M. Nisheva, and D. Vassilev, “Application of machine learning models for survival prognosis in breast cancer studies,” Information, MDPI, vol. 10, no. 3, p. 93, 2019.
SJR = 0.35 Scimagojr, Q3, цитирания: **5** (SCOPUS), ISSN: 20782489

[C5] K. Serafimova, **I. Mihaylov,** D. Vassilev, I. Avdjieva, P. Zielenkiewicz, and S. Kaczanowski, "Using Machine Learning in Accuracy Assessment of Knowledge-Based Energy and Frequency Base Likelihood in Protein Structures," in Lecture Notes in Computer Science. Springer International Publishing, 2020, pp. 572–584, doi: 10.1007/978-3-030-50420-5_43
SJR = 0.25 Scimagojr, Q3, ISSN: 16113349

[C6] M. Nisheva-Pavlova, S. Hadzhiyski, **I. Mihaylov,** I. Avdjieva, and D. Vassilev, "Linking Data for Ontology Based Advising in Healthcare," in Proc IEEE Explore 2020 International Conference Automatics and Informatics (ICAI), 2020, pp. 1–5. 10.1109/ICAI50593.2020.9311382

[C7] **I. Mihaylov,** M. Nisheva-Pavlova, and D. Vassilev, "An Approach for Semantic Data Integration in Cancer Studies," in Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 60–73, doi: 10.1007/978-3-030-22744-9_5
SJR = 0.25 Scimagojr, Q3, цитирания: **1,** ISSN: 16113349

[C8] Zhelyazkova, M., Yordanova, R., **Mihaylov, I.**, Kirov, S., Tsonev, S., Danko, D., Mason, C., Vassilev, D., "Origin Sample Prediction and Spatial Modeling of Antimicrobial Resistance in Metagenomic Sequencing Data," Front. Genet., vol. 12, 2021, doi: 10.3389/fgene.2021.642991.
Импакт фактор = 3.789, SJR = 1.41 Scimagojr, Q2, ISSN: 16648021

[C9] Zhelyazkova, M., Yordanova, R., **Mihaylov, I.**, Kirov, S., Tsonev, S., Danko, D., Vassilev, D. Bayesian Hierarchical Modelling for Antimicrobial Resistance Abundance. In Book of Abstracts: International Symposium on Bioinformatics and Biomedicine, (http://bioinfomed.org) pp 28-29, 8-10, October, Bourgas, Bulgaria.

[C10]B. Pulova-Mihaylova, **I. Mihaylov,** I. Avdjieva, and D. Vassilev, "A System for Compression of Sequencing Data," 2020, ISGT, pp 223-235, http://ceur-ws.org/Vol-2656/paper23.pdf SJR = 0.18 Scimagojr, ISSN: 16130073