ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ
Факултет автоматика
Катедра „Теоретична електротехника“

Маг. Виолета Иванова Тодорова

СТАТИСТИЧЕСКИ МЕТОДИ И АЛГОРИТМИ ЗА МАШИННО ОБУЧЕНИЕ

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен
"ДОКТОР"

Област: 5. Технически науки

Професионално направление: „Електротехника, електроника и автоматика“

Научна специалност: „Системи с изкуствен интелект”

Научни ръководители: Проф. дн инж. Валери Младенов
Доц. д-р инж. Веска Ганчева

СОФИЯ, 2022г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Теоретична електротехника“ към Факултет автоматика на ТУ-София на редовно заседание, проведено на 26.07.2022 г.

Публичната защита на дисертационния труд ще се състои на 17.11.2022 г. от 13:00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед № ОЖ-5.2-63 от 27.07.2022 г. на Ректора на ТУ-София в състав:

1. Доц. д-р инж. Владимир Димитров Христов – председател
2. Проф. д-р инж. Румен Иванов Трифонов – научен секретар
3. Проф. д-р инж. Коста Петров Бошнаков
4. Проф. д-р инж. Александра Иванова Грънчарова
5. Проф. д-р инж. Идилия Александрова Бачкова

Рецензенти:

1. Проф. д-р инж. Румен Иванов Трифонов
2. Проф. д-р инж. Коста Петров Бошнаков

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет автоматика на ТУ-София, блок № 2, кабинет № 2340.

Дисертантът е редовен докторант към катедра „Теоретична електротехника“ на факултет автоматика. Изследванията по дисертационната разработка са направени от автора, като някои от тях са подкрепени от научноизследователски проекти.

Автор: маг. Виолета Иванова Тодорова
Заглавие: Статистически методи и алгоритми за машинно обучение
Тираж: 30 броя
Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Бързото развитие на информационните технологии доведе до огромното количество информация, генерирана от големи или сложни системи. Приложенията в областта на информационните технологии, телекомуникациите, бизнеса, роботиката, икономиката, медицината и много други области генерират информационни томове, които предизвикват професионалните анализатори.

Моделите и алгоритмите от машинно обучение, извличане на данни, статистическа визуализация, изчислителна статистика и други компютърно-интензивни статистически методи са предназначени да се учат от тези сложни информационни обеми. Тези инструменти често се използват за увеличаване на ефективността и производителността на големи и сложни системи. Това естествено прави тези модели, методи и алгоритми все по-популярни във всички области на живота.

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на настоящия дисертационен труд е разработка на нови и модификация на известни методи и алгоритми за машинно обучение. Методите и алгоритмите ще бъдат приложени при създаване на нови модели от статистически данни в областта на медицината.

За постигането на целта са дефинирани за изпълнение следните задачи:

1. Да се изследват методите и алгоритмите за анализ на данни и откриване на знания от данни и да се анализират наличните биомедицински данни.
2. Да се предложи методика за прилагане на подходите на машинното обучение при процеса на набиране на пациенти за клинични изпитвания чрез обучение на подходящ класификатор на основата на обучение с учител и без учител.
3. Да се предложи нов подход за класификация на медицински данни, базиран на мемристорна невронна мрежа и да се изследва неговата ефективност.
4. Да се предложи метод за интегриране на медицински данни, включващ подготовка, съхранение и интегриране на данните за анализ, интерпретация и визуализация.
5. Да се разработи работен поток за автоматизиране на дейностите, свързани с процеса на анализ на медицински изображения.
6. Да се провери работоспособността на предложените решения.

Основните методи за изследване, използвани при решаването на поставените задачи в дисертацията са: теоретичен анализ, математическо моделиране, оптимизация, експериментални изследвания, математическа статистика за обработка на резултати.

Научна новост

Предложена е методика за прилагане на подходите на изкуствения интелект и в частност на машинното обучение и мемристорна невронна мрежа при процеса на набиране на пациенти за клинични изпитвания чрез обучение на подходящ класификатор за съответното заболяване.

Практическа приложимост

Резултатите от изследванията илюстрират, че използването на методите на машинното обучение върху данните от медицинската история на пациентите чрез автоматизирано извличане на знания и вземане на решения от медицински данни може да ускори процеса на набиране на пациенти за клиничните изпитвания.

Апробация

Резултатите от дисертационния труд са докладвани на международната конференция Computer Science през 2020 и 2022 г. и са апробирани в проект „Приложение на изкуствения интелект при набиране на пациенти за клинични изпитвания“, договор No 202ПД0029-8.

Публикации

Основни постижения и резултати от дисертационния труд са публикувани в пет научни статии, от които една самостоятелна. Две статии са докладвани на международни конференции, три са публикувани в научни списания, от които едно е индексирано в Scopus.

Структура и обем на дисертационния труд

Дисертационният труд е в обем от 121 страници, като включва пет глави за решаване на формулираните основни задачи, списък на основните приноси, списък на публикациите по дисертацията и използвана литература. Цитирани са общо 80 литературни източници, като 78 са на латиница, а останалите са интернет адреси. Работата включва общо 51 фигури и 2 таблици. Номерата на фигурите и таблиците в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. МЕТОДИ ЗА МАШИННО ОБУЧЕНИЕ И ИЗТОЧНИЦИ НА БИОМЕДИЦИНСКИ ДАННИ

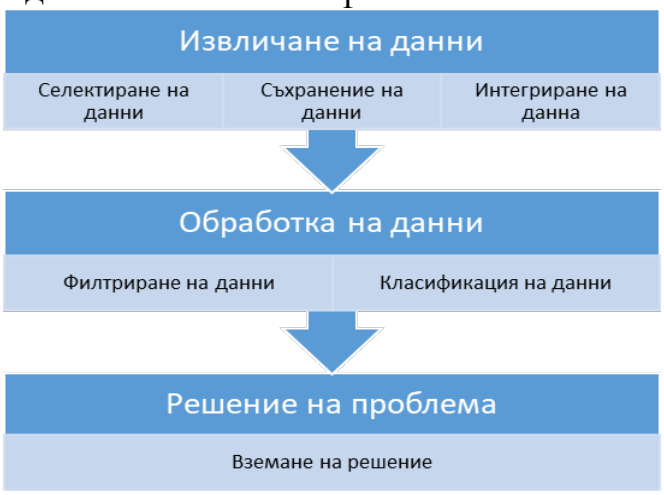
1.1. Класификация на методи и алгоритми за анализ на данни и откриване на знания от данни

1.1.2. Откриване на знания въз основа на анализ на данни

Процесът на откриване на знания и извличане на данни се състои от шест основни етапа (фиг. 1): Разбиране на проблемната област; Разбиране на данните; Подготовка на данните; Моделиране; Оценка на модела; Използване на модела.

Фиг. 1. Процес на откриване на знания базиран на анализ на данни

Концептуалният модел за откриване на знания и вземане на решения на базата на анализ на данни е показан на фиг. 2.



Фиг. 2. Концептуален модел на извличане на знания и вземане на решение на основата на анализ на данни

Цялостният процес на извличане и тълкуване на модели на данни включва повторно изпълнение на следните стъпки:

- Определяне целта на процеса на откриване на знание.
- Дефиниране на област на приложение, съответните знания и цели на крайния потребител.
- Създаване на целеви набор от данни.
- Филтриране и предварителна обработка.

- Опростяване на набора от данни чрез премахване на нежелани променливи.
- Комбиниране на целите на процеса на откриване на данни с методите за извличане на данни.
- Избиране на алгоритъм за извличане на данни.
- Извличане на данни.
- Интерпретиране на основните познания за извлечените модели.
- Използване на знанията и включването им в друга система за по-нататъшни действия.

Изследователските техники следват процеса на обработка за откриване на от колекция от данни и включват следното: подготовка, почистване и подбор на данни; откриване на знания и вземане на решения; визуализация и интерпретация на резултатите.

Работният поток за откриване на знания от данни е показан на фиг. 3.

Фиг. 3. Работен поток за откриване на знания от данни

1.2. Анализ на наличните биомедицински данни

Напредъкът в технологиите доведе до стотици, хиляди или дори милиони измервания, извършвани едновременно, като често се комбинират технологии, за да се дадат едновременни измервания на ДНК, РНК, протеин, функция заедно с клинични характеристики, включително мерки за активност на заболяването, прогресия и свързаните с тях метаданни.

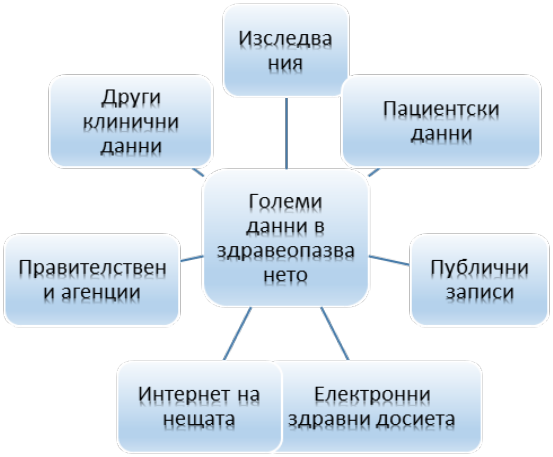
1.2.1. Източници на големи медицински данни

Източниците на големи данни в здравеопазването включват EHR, интелигентни устройства, генетични бази данни, правителство и др. (Фиг. 5).

1.2.2. Случаи на използване на големи данни и анализ на данни в здравеопазването

1) Диагностика
2) Моделиране и прогнозиране на резултатите
3) Мониторинг на жизнените показатели на пациента в реално време
4) Лечение на тежки заболявания
5) Здраве на населението
6) Превантивни грижи

7) Електронни здравни досиета (EHRs)
8) Телемедицина
9) Предупреждения в реално време
10) Интегриране на данни с медицински изображения
11) Подобряване на ангажираността на пациентите
12) Използване на здравни данни за информирано стратегическо планиране
13) Прогнозни анализи в здравеопазването
14) Разширено управление на риска и заболяванията
15) Разработване на нови терапии и иновации



Фиг. 5. Източници на биомедицински данни за здравеопазването

1.3. Изводи

Статистиката е основен стълб на машинното обучение. Без него не може да се развие дълбоко разбиране и прилагане на машинното обучение. Съвременните методи от машинно обучение използват силата на статистиката, за да изградят ефикасни модели, да правят надеждни прогнози и търсят оптимални решения.

Огромното количество натрупани данни съдържа ценни скрити знания, които могат да бъдат полезни за улесняване и подобряване на процеса на вземане на решения. Ето защо съществува обективна необходимост от създаване на автоматизирани методи за извличане на знания от данни. Извлечените знания трябва да отговарят на следните изисквания: да бъдат точни, разбираеми и полезни. Задачите за събиране на данни включват класификация, моделиране на зависимости, групиране, извличане на функции и откриване на асоциативни правила.

ГЛАВА 2. ПРИЛОЖЕНИЕ НА ИЗКУСТВЕНИЯ ИНТЕЛЕКТ ПРИ НАБИРАНЕ НА ПАЦИЕНТИ ЗА КЛИНИЧНИ ИЗПИТВАНИЯ

2.1. Проучване на проблемите при клинични изпитвания

За целите на изследването е проведен експеримент за идентифициране на подходящи пациенти за клинично изпитване за пациенти с ранен стадий на множествена склероза (МС).

2.2. Разработка на методика за прилагане на подходите на изкуствения интелект при процеса на набиране на пациенти

В това изследване ние използваме данни от медицинската история на пациентите, за да обучим алгоритмите за машинното обучение. Приложихме два вида алгоритми за машинното обучение, а именно обучение с учител и обучение без учител. При обучението с учител се използва набор от данни за обучение, за да се обучи алгоритъма за разпознаване на ранните симптоми на МС, проследяване на тяхното развитие и разработване на предсказващ модел, който може да доведе до определяне дали пациентът може да е изложен на риск от развитие на МС. За обучение без наблюдение се опитваме да групираме данни, така че пациентите в риск и пациентите, които не са изложени на риск, да попаднат в различни клъстери.

Използвахме софтуер за статистически анализ, за да извлечем необходимите променливи от суровите данни и да подготвим данните във най-подходящ формат. Липсващите стойности на идентификатора на пациента са обработени чрез премахване на съответния клиничен запис. Данните са подгрупирани, за да съдържат петте най-често съобщавани термина за медицинска история за пациентите с МС и по същия начин за пациентите, които не са с МС. След това данните са транспонирани до един запис на пациент и всеки докладван термин за медицинската история е настроен да представлява отделна колона. Наличието на докладван термин за медицинската история е отбелязано с „1“ и липсата му е отбелязано с „-1“.

Данните от анализа, използвани за класификаторите на машинното обучение в проекта (SVM и k-mean clustering), се състоят от: петте най-често отчитани термина за медицинска история за популацията от пациенти с МС от базовите данни, които са: ЗАХАРЕН ДИАБЕТ, ВИСОКО КРЪВНО НАЛЯГАНЕ, ВИСОК ХОЛЕСТЕРОЛ, АРИТМИЯ, СЪРДЕЧНА БОЛЕСТ и петте най-често съобщавани термина за медицинска история за популацията на пациенти без МС от базовите данни, които са: МАЛАРЕН ОБРИВ, АРТРИТ, БЪБРЕЧНО НАРУШЕНИЕ, ХИПЕРТОНИЯ, ФОТОСЕНЗИТИВНОСТ. Използвахме тези 10 признака за векторът на характеристиките за всеки пациент.

2.2.1. Използвани методи за ранно откриване на МС

При приложението на методи от типа на обучение с учител и без учител се избира т.нар. обучаваща извадка, върху които се прилага съответния алгоритъм и на база на него се получават параметри, които позволяват обекта да се класифицира към съответен клас. В разглеждания случай класовете са два – на пациенти с множествена склероза и пациенти без множествена склероза. За тестване на използвания класификатор се използва т.нар. тестваща извадка – обекти, които се тестват с обучения класификатор и които не са били използвани при обучението. На база на коректна или некоректна класификация се изчислява процента на вярно класифицираните обекти. Ако пациента е коректно класифициран, това означава че той съответно коректно може да бъде избран за включване или невключване в

съответното клинично проучване, което в разглеждания случай е изследване за ранно откриване на множествена склероза.

2.2.2. Метод с опорните вектори - Support Vector Machines (SVM)

Методът с опорните вектори е подход на обучение с учител, при който се анализират данните, използвани за класификационен и регресионен анализ. За класификация на два класа се използват етикет с +1 за принадлежност и -1/0 не принадлежи към съответния клас. В нашия случай използваните класове за пациенти са +1 - със симптоми на МС и -1/0 - без симптоми на МС.

2.2.3. Метод k най-близки съседа

Типичен представител на методите на обучение без учител е методът на k-те най-близки съседи. При обучението без учител се открива скрита структура в данните, при която предварително не знаем правилния отговор. При анализа на клъстеризирането се получава естествено групиране в данните, така че елементите в един и същ клъстер са по-сходни помежду си от тези от различни клъстери. В k-means clustering всеки клъстер е представен от точка с данни "прототип". Предимството, което имаме е, че предварително можем да зададем броят на клъстерите, които търсим. В нашия случай са два (със симптоми за МС и без симптоми на МС). Може да разширим изследванията с повече клъстери, като видим как се разпределят данните в тях.

2.3. Тестване на разработената методика при клинични проучвания

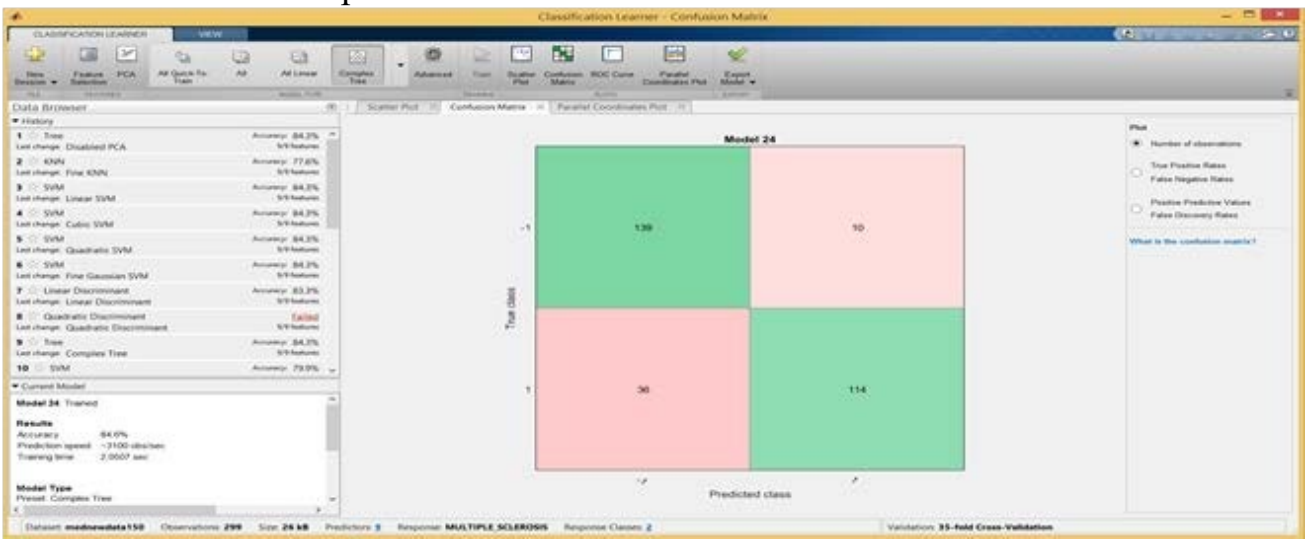
В първия случай са използвани 342 пациента, от които 183 са с множествена склероза и 159 нямат установена множествена склероза. Във втория случай са използвани клиничните данни на 1937 пациенти, от които 1431, посещаващи услугите за МС в болницата, и 506 пациенти, които не са болни от МС. Събраните оригинални данни включват медицинска история, съдържаща докладваните термини (които са съпътстващо заболяване) за пациентите и демографските данни на пациента: възраст, пол, раса и държава на раждане на пациента.

2.3.1. Първи случай

Метод с опорните вектори - Support Vector Machines (SVM)

Наборът от данни е разделен на две подгрупи за обучение - 300 пациенти и за тестване на 42 пациенти. Резултатите се получават чрез използване на ML тулбокса на Matlab. Резултатите от обучението са показани на фиг. 6. На фигурата са дадени резултатите от обучението и с други методи, но SVM дава най-добрата точност от 84,3%. Изразени са успешно обучението точки от неуспешно обучените по отношение на множествената склероза. Вижда се, че за клас -1 са разпознати 139 точки, а 10 не са разпознати. За клас 1 са разпознати 115, а 35 точки не са разпознати. Тук процентно се вижда успеваемостта при обучението разделена на два класа -1 и 1. Въз основа на получения класификатор всички 42 пациенти, използвани за тестване, са класифицирани правилно. Оказва се, че при премахване на болестите: acne,

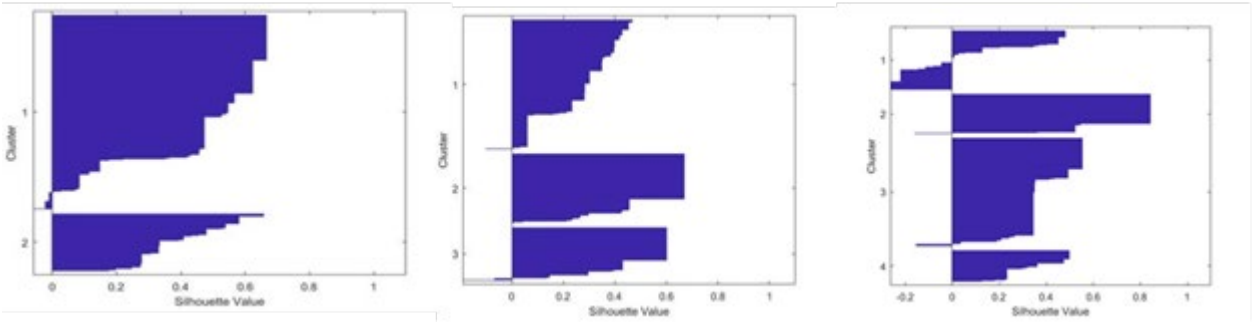
appendectomy, appendicitis, menstrual_cramps се повишава леко процента на точност на 84.9%. Разликата е в един брой по-малко разпозната единица от клас 1. Изводът е, че изброените болести не би трябвало да влияят на множествената склероза.



Фиг. 6. Резултати от обучението за откриване на множествена склероза с използване на различни алгоритми

k най-близки съседа - k-means clustering

При обучението без учител се открива скрита структура в данните, при която предварително не знаем правилния отговор. При анализа на клъстеризирането се получава естествено групиране в данните, така че елементите в един и същ клъстер са по-сходни помежду си от тези от различни клъстери. В k-means clustering всеки клъстер е представен от точка с данни "прототип". Предимството е, че предварително знаем, че броят на клъстерите, които търсим, са 2 (със симптоми за МС и без симптоми на МС). Групирането в 2, 3 и 4 клъстера е показано на Фиг. 8. Според средната стойност на клъстерите най-подходящият брой клъстери е 2. Това отговаря на факта, че реалният брой класове, към които принадлежат пациентите, са 2. Първият отговаря на пациентите със симптоми на МС, а вторият - на пациентите без симптоми на МС.



Фиг. 8. Групиране в 2, 3 и 4 клъстера

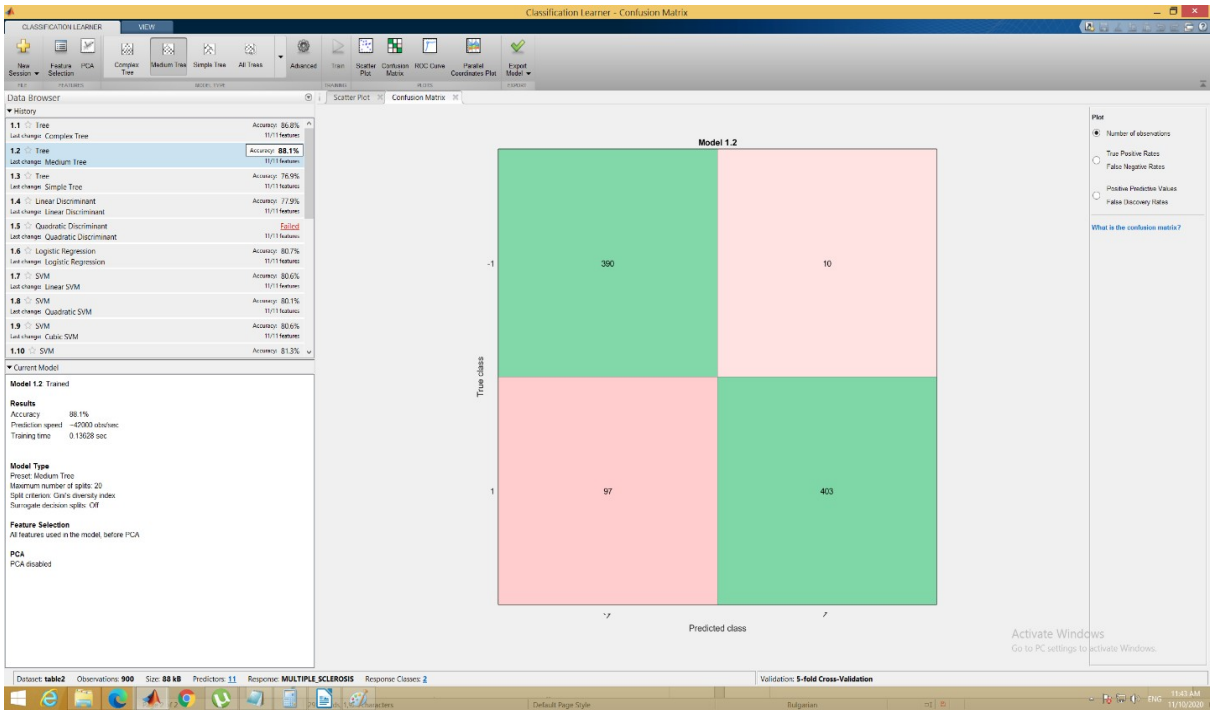
2.3.2. Втори случай

Обучение с учител

Наборът от данни е разделен на две подгрупи за обучение - 900 пациенти (500 случая с множествена склероза и 400 случая без множествена

склероза) и за тестване на 200 пациенти (100 случая с множествена склероза и 100 случая без множествена склероза). Резултатите от обучението са показани на фиг. 9. Най добър резултат 88.1% се получава с алгоритъма Medium Tree. Изразени са успешно обучението точки от неуспешно обучените по отношение на множествената склероза. Вижда се, че за клас -1 са разпознати 390 точки, а 10 не са разпознати. За клас 1 са разпознати 403, а 97 точки не са разпознати. Тук процентно се вижда успеваемостта при обучението разделена на два класа -1 и 1. След теста се получава точност от 89%. От 100 случая с множествена склероза всички са разпознати. От 100 случая без множествена склероза, 78 са разпознати и 22 са грешни.

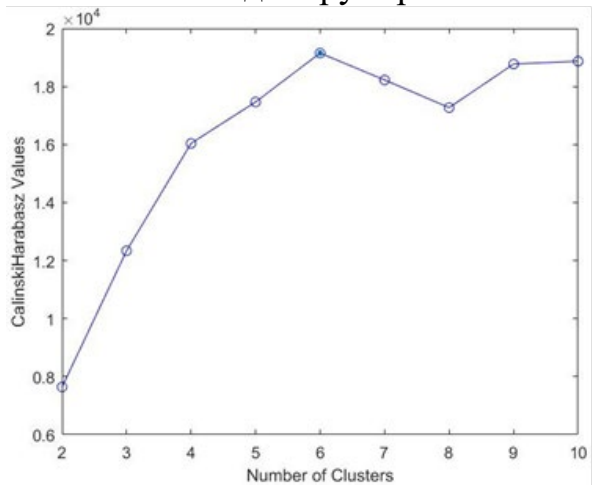
В изследването са направени тестове с различни извадки за обучение и тестване. Наборът от данни е разделен на две подгрупи за обучение - 1100 пациенти (700 случая с множествена склероза и 400 случая без множествена склероза) и за тестване на 200 пациенти (100 случая с множествена склероза и 100 случая без множествена склероза). Най добър резултат 88.4% се получава с алгоритъма RUSBoosted Tree. След теста се получава точност от 96.5%. От 100 случая с множествена склероза 95 са разпознати и 5 са грешни. От 100 случая без множествена склероза, 98 са разпознати и 2 са грешни. При използване на други алгоритми от машинното обучение са полулучени следните резултати за разглеждания случай. С използване на Fine Gaussian SVM е получена 80.1% точност при обучение и 96% точност при тестване. С използване на Weighted KNN е получена 85.5% точност при обучение и 93.5% точност при тестване.



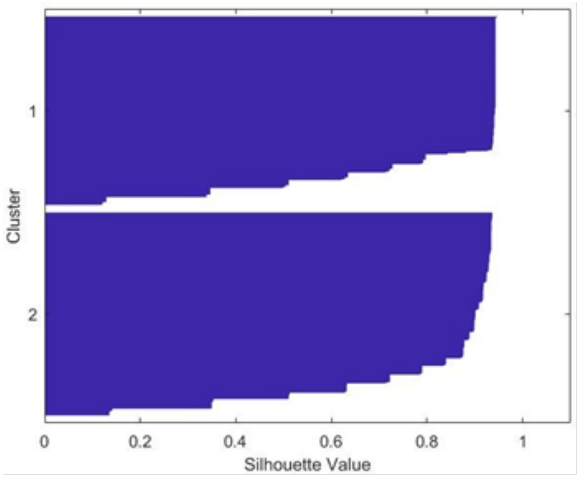
Фиг. 9. Резултати от обучението за откриване на множествена склероза с използване на Medium Tree

k най-близки съседа - k-means clustering

При използването на данните за всички 1937 пациента се оказва, че оптималният брой на клъстерите е 6. Резултатът с използването на критерия на Калински-Харабас е показан на фиг. 11. Тъй като за случая се интересуваме от групиране в два класа, в следващото изследване се прави анализ на групирането в два и три клъстера с използване на метода на k-те най-близки съседа. Групирането в 2 клъстера е показано на Фиг. 12.

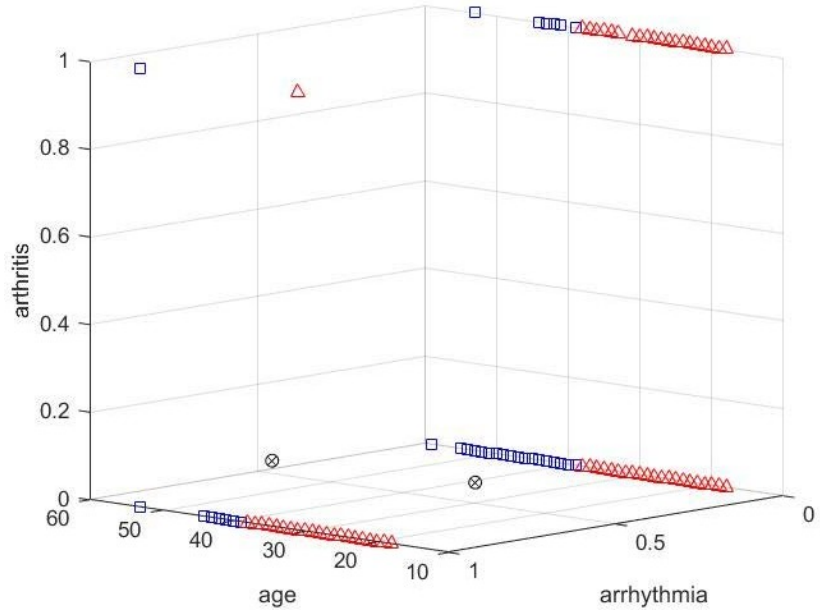


Фиг. 11. Оценка на оптималния брой на клъстерите с използване на критерия на Калински-Харабас



Фиг. 12. Графично представяне на данните в два клъстера.

От графиката се вижда, че по-голямата част от точките в двата клъстера имат широко разпределение, по-голямо от 0.8, показващо, че точките са добре разделени от съседния клъстер. Все пак всеки клъстер също съдържа известен брой от точки с разпределение под 0.8, показващо, че те се намират в близост до границата между двата клъстера. На фиг. 13 са представени двата клъстера в тримерно пространство, като са отбелязани техните центрове.



Фиг. 13. Графично представяне на двата клъстера в тримерното пространство, като са означени и центровете им

2.4. Заключение

Отчетените добри резултати при тестването на използваните методи на машинното обучение не само могат да помогнат на изследователите да увеличат броя на избраните пациенти за кандидати за клинично изпитване, но също така могат да бъдат от полза за пациентите да бъдат диагностицирани с болестта в ранен етап на развитие и следователно да имат ранен старт на лечението от ранното откриване и точната диагноза на МС са от решаващо значение за забавяне на прогресията на заболяването, доколкото е възможно и подобряването на състоянието на пациентите.

Резултатите илюстрират, че използването на методите на машинното обучение върху данните от медицинската история на пациентите може да ускори процеса на набиране на пациенти за клиничните изпитвания и да помогне за намирането на пациенти, които отговарят на условията за конкретно проучване.

2.5. Изводи

1. Предложена е методика за прилагане на подходите на изкуствения интелект и в частност на машинното обучение при процеса на набиране на пациенти за клинични изпитвания чрез обучение на подходящ класификатор за съответното заболяване на основата на обучение с учител и без учител.
2. Изследвани са два типа подходи: обучение с учител - метод с опорни вектори (SVM) и обучение без учител - k най-близки съседа за ранно откриване на множествена склероза за избор на субекти въз основа на тяхната медицинска история. Класифицирането се извършва по често срещани ранни симптоми на заболяването.
3. Предложената методика е верифицирана експериментално чрез използване на два набора от клинични данни за множествена склероза, включващи медицинска история, съдържаща докладваните термини (които са съпътстващо заболяване) за пациентите и демографските данни на пациента: възраст, пол, раса и държава на раждане на пациента.

ГЛАВА 3. ПРИЛОЖЕНИЕ НА МЕМРИСТОРНА НЕВРОННА МРЕЖА ЗА КЛАСИФИКАЦИЯ НА ПАЦИЕНТИ С COVID-19

3.1. Въведение

В настоящата глава на дисертационния труд се разширяват подходите, които са използвани във втора глава за идентифициране на подходящи пациенти за клинично изпитване с ранен стадий на множествена склероза. Използван е нов подход за класификация, а именно невронни мрежи. В тази връзка е обучена невронна мрежа, която се използва при класификация на пациенти с COVID-19. Обучената невронна мрежа за класификация на пациенти с COVID-19 използва мемристорни синапсиси. Получените добри резултати илюстрират, че използването на методите на машинното обучение върху данните от медицинската история на пациентите може да

ускори процеса на набиране на пациенти за клинични изпитвания и да помогне за намирането на пациенти, които отговарят на условията за конкретно проучване.

3.2. Описание на предложения модел и изследванията

В изследването са използвани медицински данни на произволно избрани пациенти, приложени за обучение на разглежданата невронна мрежа с право предаване на сигналите. За обучение с учител се използва набор от данни за обучение на невронната мрежа за разпознаване на пациенти, които са предразположени към COVID-19, проследяване на тяхното развитие и генериране на прогнозен модел, който може да доведе до определяне кой пациент с какво специфично съществуващо здравословно състояние може да бъде изложен на риск от развитие на тежко заболяване. За обучение с учител се използват целеви данни, отнесени към потвърдени болни и здрави пациенти. За обучение без учител се прилага самоорганизираща се мрежа, базирана на състезание, и невронната мрежа групира данните, така че пациентите в риск и здравите пациенти да попадат в различни клъстери [26], [28], [29], [24]. Разглежданата самоорганизираща се мрежа се използва за групиране на анализираните обекти, на базата на техните различни статистически свойства [29], [44]. Суровите медицински данни се състоят от цялата налична информация до този момент, идваща от клинични досиета, обхващащи период от април 2020 г. до февруари 2021 г. Използвани са клинични записи на 20400 потенциални пациенти, от които 10200 пациенти са били заразени със SARS-CoV-2, а останалите 10200 болнични пациентине са били заразени.Медицинските досиета са свързани с 10232 жени и 10168 мъже на възраст от 21 до 85 години [26], [27].

Наборът от данни съдържа 17 параметъра: анонимизиран номер на пациента, пол, възраст, статус и тежест на COVID-19 заболяването, индекс на телесната маса, състояние на сърдечните заболявания, статус на хипертония, диастолично и систолично кръвно налягане, статус на диабет, стойности на хемоглобина, статус на хронично бъбречно заболяване, стойности на калций, калий и фосфор, и статус на ракови заболявания [25], [26]. В изследването се прилага софтуер за статистически анализ и за извеждане на необходимите променливи в подходящ формат. Наличието на COVID-19 е означено с „+1“, а неговото отсъствие се обозначава с „-1“. Мъжкият пол е кодиран с „1”, а женският – с „0”.

Невронна мрежа за разпознаване на образи

С помощта на събрания набор от данни на пациентите се прилагат две основни техники – обучение с учител и групиране в k-под-групи за изследване на данните за класификацията на пациентите с повишен риск от заболяване въз основа на съществуващите им здравословни състояния и показатели [22], [23], [27]. Синаптични вериги,базирани на мемристори, с делител на ток и комплементарни MOS транзистори се прилагат в разглежданата изкуствена невронна мрежа с право предаване на сигналите, за класификация на данните.

Обучение с учител

Изкуствена невронна мрежа с право предаване на сигналите се прилага за приблизителна оценка на основните рискови фактори и за приблизително прогнозиране на наличието или отсъствието на това заболяване в избрана група от пациенти. Набор от 20400 потенциални пациенти се анализира с помощта на мемристорната невронна мрежа. За този тест се прилагат 15 медицински характеристики и показатели (пол, възраст, тежест на COVID-19, индекс на телесна маса, сърдечни заболявания ихипертония, диастолично и систолично кръвно налягане, стойности на диабет и хемоглобин, хронично бъбречно заболяване, стойности на калций,калий и фосфор, и статус на ракови заболявания. Тези характеристики са групирани в матрица, чиито редове отговарят на описаните по-горе медицински свойства и здравословни състояния. Всяка колона от разглежданата матрица е свързана с даден потенциален пациент. Описаната матрица представлява входните данни за невронната мрежа с право предаване на сигналите. Първите 14280 записа (около 70% от разглежданата група потенциални пациенти) се използват за обучение на невронната мрежа. Останалите 6120 записа на матрицата се прилагат за тестване на мемристорната невронна мрежа след етапа на обучение. Целевите данни се поставят в ред от използваната матрица. Той съдържа съответно „-1“ за липса на COVID-19 и „+1“ за положителен тест за COVID-19. Целевият ред съдържа 14280 елемента за трениране и обучение на невронната мрежа. Мемристорната невронна мрежа съдържа входен слой от 15 възела за прилагане на входните сигнали и два слоя неврони – съответно скрит слой и изходен слой [9]. Петнадесетте входни сигнала, приложени към невронната мрежа, съответстват на обсъжданите по-рано медицински характеристики и индикатори за класификация на потенциални пациенти със заболяване COVID-19.

Включените 12 неврона в скрития слой на невронната мрежа имат тангенсова сигмоидална активационна функция h(x) [29], [30]:

$$h(x) = \frac{exp(x) - exp(-x)}{exp(x) + exp(-x)} \quad (1)$$

където x е претеглената сума от входните сигнали x_i:

$$x_j = w_{jb} + \sum_{i=1}^{15} w_{ij} x_i \quad (2)$$

и w_{jb} е теглото на съответния синапс, използван за прилагане на единичния сигнал на отместването [30]. Индексът i представлява номера на съответния входен възел, а j показва номера на разглеждания неврон в скрития слой на невронната мрежа. Всеки входен сигнал се прилага към всички неврони в скрития слой с помощта на базирани на мемристори синаптични устройства. Изходите на невроните в скрития слой са свързани с неврона в изходния слой. Този неврон има линейна активационна функция. Действителният изходен сигнал на разглежданата невронна мрежа, базирана на мемристори, се означава с y [30]:

$$y = v_b + \sum_{i=1}^{12} v_j x_j \quad (3)$$

Полученият сигнал на грешката се използва за корекция на теглата на

мемристорните синапсиси между невроните от скрития слой и изхода по време на процеса на обучение, в комбинация с предварително определена скорост на обучение η и сигналите zj, приложени към входа на невронаот изходния слой на изкуствената невронна мрежа [30]:

$$\Delta v_i = \eta . \delta . z_i, \quad (4)$$

където δ е математически член, носещ информация за грешката:

$$\delta = (d - y) f^1(y) \quad (5)$$

Обновяването на теглата на синапсисите между входните възли и невроните в скрития слой е:

$$\Delta w_{ij} = \eta . \delta_i . x_i = \eta . x_i . \sum_{i=1}^{12} \delta . v_i \quad (6)$$

Напрежителен сигнал, пропорционален на съответния сигнал на грешката, се прилага към синапсите за корекция на теглото им. Актуализацията на теглата се повтаря до получаване на минимална стойност на средно-квадратичната грешка между желания изходен сигнал и неговата действителна стойност. След минимизиране на средноквадратичната грешка следва използване на мемристорната невронна мрежа. Получената грешка след периода на използване е равна на броя на неправилно класифицираните пациенти, разделен на техния общ брой, и в конкретния случай е около 1,96 %.

Групиране в k-клъстери

Извършва се допълнителен анализ на избраната група потенциални пациенти чрез групиране на данните [30], [44] в клъстери като се използва методиката от втора глава на дисертацията. Анализираните данни трябва да бъдат разделени на групи, броят на които трябва да бъде предварително определен. Тази информация често не е налична предварително, така че трябва да се проведат множество експерименти, за да се намери оптималният брой на групите. Намаляването на сложността на данните се осъществява чрез предварителен статистически анализ и тяхната обработка. При клъстерния анализ се получава естествено групиране на данните, т.е. елементите в един и същи клъстер имат по-голямо сходство помежду си, в сравнение с тези от другите клъстери. При клъстеризирането на k-групи, всеки клъстер е представен от точка "прототип". Основно използваното при групиране на данни квадратно Евклидово разстояние се изразява по следния начин [20]:

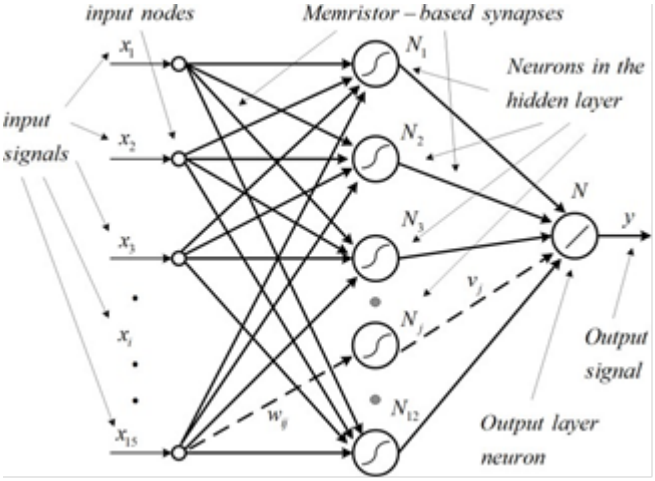
$$dE^2 = \sum_{k=1}^{n} (c_x - x_k)^2 \quad (7)$$

където c представлява центъра на разглеждания клъстер, x е свързан с координатите на анализираната точка, k е размерността на свързаната точка, представяща данните в разглежданото многомерно пространство, а n е нейният максимален размер [30]. Когато центроидът на един клъстер от данни принадлежи към най-близкия клъстер, не се извършват никакви промени. Ако се окаже, че друг центроид е най-близкият, според изчислените координати, тогава разглежданият клъстер се пренасочва към другия центроид и текущите координати за старите и новите подпространства се преизчисляват като средна стойност на координатите на описаните групи от данни [30], [44]. Основното предимство в

настоящия случай е, че предварително е известно, че броят на клъстерите е два. Клъстеризирането на данните е реализирано в среда на MATLAB [45]. Получената след класификацията грешка е около 2,3 %. Извършени са допълнителни експерименти за групиране на данните съответно в два, три и четири класа [30], [45]. Времето за обучение на невронната мрежа в този случай е около 2983 ms. След проведените анализи се установява, че заболяването COVID-19 се наблюдава предимно при пациенти на възраст над 65 години и със систолично кръвно налягане от 125 до 130.

3.3. Невронна мрежа, базирана на мемристори

В тази работа е приложен различен синапсис, базиран на делител на ток с мемристор, резистори и диференциален усилвател с MOS транзистори. Тази синаптична схема е с намален брой на използваните електронни елементи. Той съдържа само един мемристорен елемент на синапсис. Той реализира положителни, нулеви и отрицателни синаптични тегла. Мемристорът M, базиран на танталов оксид, се прилага в синаптичната верига, поради добрия си запаметяващ ефект и отличните си превключващи свойства. Мемристорният елемент има два полюса – съответно горен електрод (TE) и долен електрод (BE) [37], [43]. Неговата нано-структура има квадратно напречно сечение. Тя съдържа няколко успоредни проводящи канала [37], [43]. Периферната област на мемристора е изградена от чист, стехиометричен и аморфен Ta2O5 материал. Централният проводящ канал на мемристора се базира на твърд разтвор на кислород в танталов материал. Между тези две области се образува друг канал от легиран с кислородни ваканции танталов оксид [37]. Състоянието на мемристора се променя чрез прилагане на пакети от външни положителни или отрицателни правоъгълни напрежителни импулси с продължителност 1 μs и ниво 0,52 V. Схемата на разглежданата мемристорно-базирана невронна мрежа е показана на фиг. 15.



Фиг. 15.Схема на невронна мрежа с право предаване на сигналите, базирана на мемристори

Приложеният модифициран модел на мемристор [40] има подобрени характеристики спрямо стандартния модел на Hewlett-Packard [37]. Използваният модел на мемристор от танталов оксид е разгледан и описан

подробно в [46], [47]. Неговият съответстващ LTSPICE библиотечен модел, приложен в работата, е свободно достъпен [46]. Съпротивленията на мемристорните елементи, включени в синапсисите между невроните в скрития слой на мрежата и изходния неврон са представени в Таблица 1. Номерът на съответния неврон в скрития слой се обозначава с индекс *j*, който се променя от 1 до 12. Стойностите на съответните съпротивления са получени след етапа на обучение на изкуствената невронна мрежа. Съпротивленията на мемристорите M в този случай се променят между 300 **Ω** и 783 **Ω**. Стойностите на съпротивлението на мемристора, които са по-високи от 300 **Ω**, съответстват на положителни синаптични тегла.

Таблица 1. Съпротивление на мемристора, отговарящи на синаптичните тегла, получени след процеса на обучение за синапсите v_{ij} – разположени между скрития слой и изхода

Memristances related to the synapses v_{ij}						
Neuron	1	2	3	4	5	6
M	435,34	300,00	300,00	300,00	300,00	782,93
Neuron	7	8	9	10	11	12
M	300,00	300,00	300,00	300,00	300,00	300,00

Съпротивленията на мемристорите, свързани със синапсите, разположени между входните възли и скрития слой, са представени в Таблица 2. Съответните входни възли са обозначени с номера от 1 до 15 и отговарят на медицинските показатели, разгледани по-горе. Съпротивленията на мемристорите са получени след обучението на невронната мрежа. В настоящия случай съпротивлението на мемристорите се променя между 92 **Ω** и 1930 **Ω**. Съпротивленията на мемристорите, по -ниски от 300 **Ω**, съответстват на отрицателни синаптични тегла [46], [48].

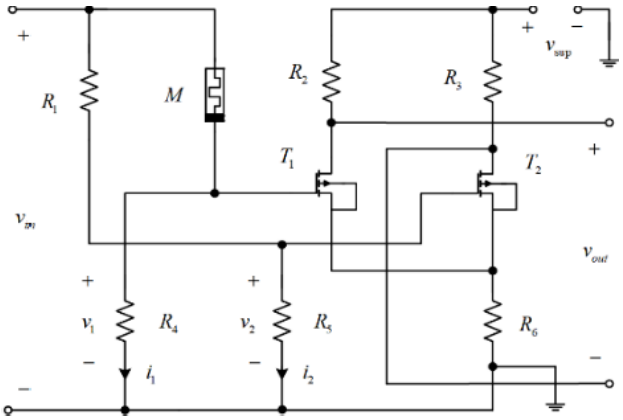
Схемата на разглежданата модифицирана синаптична верига е дадена на фиг. 16. Приложеният синапс съдържа токов делител (CD) и диференциален усилвател, реализиран с MOS транзистори - T_1 и T_2. Използваният делител съдържа два клона, свързани паралелно. Първият клон съдържа мемристора M и резистора R_4, а вторият клон има два резистора, свързани последователно – R_1 и R_5. Съответните токове, протичащи през разглежданите клонове, са обозначени с i_1 и i_2. Сигналът на входното напрежение на синаптичната верига е означен с vin. Съответният коефициент на предаване по напрежение на диференциалния усилвател е обозначен с kv и има стойност 10. Съпротивлението на елемента R_1 има стойност 300 **Ω**, а резисторите R_4 и R_5, включени в токовия делител, са със съпротивления от 1 k**Ω**. Работата на описаната синаптична схема, базирана на мемристор, се основава на сравнение на токовете i_1 и i_2, протичащи през клоновете на токовия делител [46]. Чрез промяна на съпротивлението на мемристора M, синаптичното тегло *w* се променя по време на етапа на обновяване на синаптичното тегло. След анализ на мемристорната синаптична схема, за

стойността на теглото w в зависимост от съпротивлението на мемристора M се получава:

$$w(M) = \frac{v_{out}}{v_{in}} = k_v \cdot \left(\frac{R_5}{R_1 + R_5} - \frac{R_4}{M + R_4} \right) \quad (8)$$

Таблица 2. Съпротивления на мемристорите, отговарящи на синаптичните тегла след процеса на обучение за синапсите, означени cw_{ij} – разположени между входния слой и скрития слой

	Inputs	1	2	3	4	5	6	7
Neurons in the hidden layer	1	300,05	299,98	942,03	299,98	323,13	300,16	299,95
	2	303,85	118,73	365,73	172,07	218,21	186,2	266,95
	3	302,63	327,8	106,49	291,41	507,06	265,22	310,39
	4	105,44	124,33	487,09	372,96	287,51	181,35	325,26
	5	243,28	381,48	39,14	205,79	257,13	228,11	184,41
	6	299,97	300,01	1930	300,01	283,35	299,9	300,03
	7	264,76	272,3	713,18	298,57	210,62	226,84	233,85
	8	303,92	186,95	160,14	273,37	356,72	208,7	357,67
	9	376,86	415,18	566,18	217,58	632,91	293,82	339,02
	10	350,59	343,49	699,37	298,3	361,7	348,07	230,55
	11	351,61	217,13	703,35	173,17	189,91	445,48	401,63
	12	442,57	180,19	450,18	161,64	92,23	564,46	399,63
Inputs	8	9	10	11	12	13	14	15
1	300,03	300,32	300	525,44	299,99	300	299,98	299,61
2	511,25	112,19	150,59	365,23	495,18	527,52	493,44	458,97
3	285,41	220,61	319,64	73,09	330,66	312,55	307,42	393,79
4	283,08	99,77	259,9	114,83	479,8	328,21	220,43	145,58
5	348,59	395,88	285,04	302,19	401,36	381,43	325,11	368,31
6	299,98	299,81	300	833,96	300,01	300	300,01	300,23
7	334,17	379,09	281,05	740,92	359,84	301,24	335,55	368,15
8	485	431,91	437,44	159,28	346,76	415,89	162,78	477,09
9	276,55	527,13	222,23	184,29	404,9	334,23	360,41	181,31
10	333,15	255,77	471,78	456,67	379,01	359,1	280,91	169,32
11	254,79	223,26	277,88	629,52	358,04	385,56	284,8	323,56
12	520,62	161,82	271,68	177,15	390,43	156,31	245,43	547,4



Фиг. 16. Схема на приложения синапс, базиран на танталово-оксиден мемристор

Съпротивлението на мемристора M и съответното синаптично тегло на веригата w се променят чрез прилагане на външни напрежителни

импулси. След трансформиране на уравнение (1) и като се има предвид, че съпротивлението R_4 е равно на R_5, се получава следният израз:

$$\begin{cases} w(M) = 0, & M = R_1 \\ w(M) < 0, & M < R_1 \\ w(M) > 0, & M > R_1 \end{cases} \quad (9)$$

Съпротивлението на използвания мемристор M, базиран на танталов оксид, като функция на приложеното напрежение v и променливата на състоянието x може да бъде изразено чрез приложения модифициран модел на мемристор, както следва [47]:

$$M(x,v) = \frac{1}{(h_1 v^4 + h_2 v^2 + h_3).(1-x) + G_{max} x} \quad (10)$$

където h_1, h_2 и h_3 са параметри за настройка на модела на мемристора, а Gmax е максималната проводимост на мемристора във включено състояние [37]. Диференциалното уравнение на състоянието на разглеждания модел на мемристор е [46], [47]:

$$\frac{dx}{dt} = \begin{bmatrix} exp\left(\frac{vi}{\sigma_P}\right) exp\left(-\frac{x^2}{x_{ON}^2}\right) s(v) \\ (k_1 v^3 + k_2 v) + sinh\left(\frac{v}{\sigma_{OFF}}\right) A.s(-v) \\ exp\left(-\frac{x_{OFF}^2}{x^2}\right) exp\left(\frac{1}{1+\beta iv}\right) \end{bmatrix} f_{win,} \quad (11)$$

където A, k1, k2, xON, xOFF, σP, β и σOFF са параметри за настройка и fwin е прозоречна функция за ограничаване на променливата на състоянието [36], [46]. Приложената гладка и диференцируема логистична функция s(v) е:

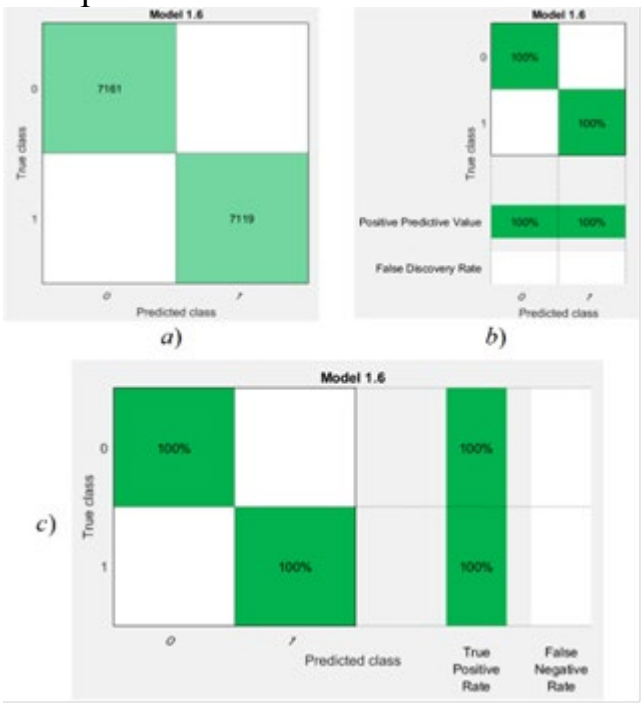
$$s(v) = 0.5[1 + v(v^2 + m)^{-0.5}], \quad (12)$$

където m е константа със стойност между 50 и 1000. Чрез промяната на предавателния коефициент kv се осъществява мащабиране на синаптичните тегла в широк диапазон. Резисторите $R_2 = R_3 = 300$ **Ω** се използват за ограничаване на дрейновите токове на MOS транзисторите T_1 и T_2 и за определяне на работните им точки. Резисторът R_6 със стойност 100 **Ω** е свързан между сорсовите електроди на MOS транзисторите и нулевия електрод [46]. Този резистор се включва във веригата за въвеждане на отрицателна обратна връзка по напрежение.

3.4. Резултати и коментари

Основните ристови фактори, свързани с развитието на COVID-19 върху избрана група от потенциални пациенти, са анализирани с помощта на мемристорна невронна мрежа. Данните, съдържащи 15 важни медицински индикатора за потенциалните пациенти, са представени в таблица, съдържаща записи от данни за 20400 анализирани обекти. Първите 14280 записа от таблицата се използват за обучение на невронната мрежа и съответните резултати са представени на фиг. 17. Получената матрица на разбъркване представя 7161 пациенти с потвърдено заболяване и 7119 здрави пациенти. Точността на получените резултати в този случай е близка до 100%. След обучение на невронната мрежа, следващите 6120 елемента се използват за нейното тестване. Получената грешка в този случай е около 3,9%. Времето за симулация на мемристорната мрежа е около 8534 ms, по

отношение на 9732 ms за традиционна невронна мрежа. Получената грешка в този случай е около 3,4 %, което представлява преимущество на описаната мемристорна невронна мрежа, по отношение на подхода с клъстеризация. Описаните техники за разделяне на болни и здрави пациенти дават сходни резултати. Разглежданите анализи са свързани с голяма група пациенти между 21 и 85 години с допълнителни симптоми на рак и сърдечни заболявания. COVID-19 се наблюдава главно при пациенти над 65 години със систолично кръвно налягане от 75 до 79. Около 47% от болните хора са жени. Около 40% от пациентите с потвърдено заболяване са с придружаващ диабет и близо 50% от тях са с ракови заболявания.



Фиг. 17. a) Брой на наблюденията; b) Процент на откриване на положителни прогнозни стойности; c) Процент на истински положителни и отрицателни резултати

3.5. Заключение

В изследването е извършена оценка на основните рискови фактори върху група потенциални пациенти със заболяване COVID-19 с помощта на мемристорна невронна мрежа. Приложена е модифицирана схема на синапс, базиран на на мемристор от танталов оксид. Функционирането на разглежданата синаптична схема се основава на сравнение на токовете, протичащи през мемристора и един допълнителен резистор, свързани в токов делител. За допълнително мащабиране на синаптичните тегла се използва и диференциален усилвател с MOS транзистори. Предимство на предложената синаптична схема е способността ѝ да реализира положителни, нулеви и отрицателни синаптични тегла. Друго предимство е минималният брой мемристори на синапс. Разглежданата невронна мрежа, базирана на мемристори, успешно класифицира с минимална грешка даден набор от пациенти в две категории (болни и здрави потенциални пациенти). Проведена е допълнителна класификация на приложения набор от пациенти,

като се използва традиционна невронна мрежа и техниката за клъстеризиране на данни в k-групи. Получените резултати потвърждават правилното функциониране на приложената мемристорна мрежа с право предаване на сигналите за прогнозиране на болни от COVID-19 пациенти, като за рискови фактори се използват няколко медицински индикатори. Предложената невронна мрежа с мемристорни синапси има по-висока скорост на функциониране от традиционния си аналог. По този начин, разработената във втора глава на дисертационния труд методика е обогатена с нов вид изследване на пациенти, които потенциално може да са болни от COVID-19 и с нов подход за класификация, а именно невронни мрежи.

3.6. Изводи

1. Предложен е нов подход за класификация, базиран на мемристорна невронна мрежа. Обучената невронна мрежа за класификация на пациенти с COVID-19 използва мемристорни синапсиси.
2. Проведена е допълнителна класификация на приложения набор от пациенти с COVID-19, като се използва традиционна невронна мрежа и техниката за клъстеризиране на данни в k-групи.
3. Верифицирана е експериментално предложената мемристорна невронна мрежа за прогнозиране на болни от COVID-19 пациенти, като за рискови фактори се използват няколко медицински индикатори. Предложената невронна мрежа с мемристорни синапси има по-висока скорост на функциониране от традиционния си аналог.

ГЛАВА 4. МЕТОД ЗА ИНТЕГРИРАНЕ НА МЕДИЦИНСКИ ДАННИ

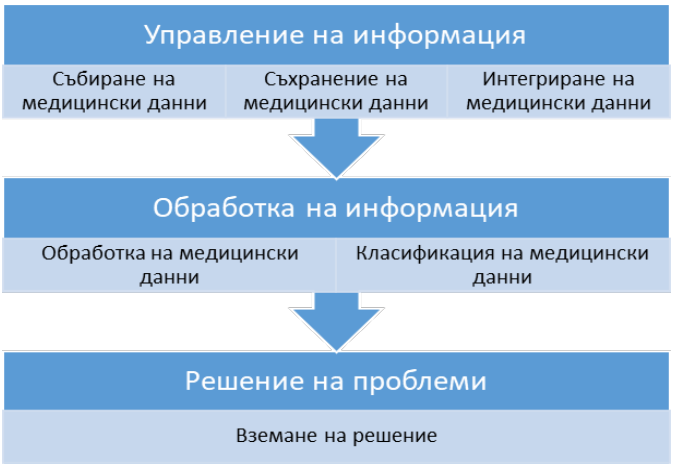
4.1. Въведение

Данните се получават от различни източници. Те са в множество формати (плоски файлове, релационни таблици, текстови файлове, файлове представящи различни видове графи, позволяващи откриването на несъответствия в основните данни и др.). Източниците на биомедицински данни се характеризират с изключително висока степен на хетерогенност по отношение на типа на използвания модел на данни, схемата на даден модел, както и несъвместимите формати и номенклатури на стойностите [50]. В допълнение, като източник на данни често се използват голям брой неустойчиви и непоследователни идентификатори.

В настоящата глава е представен концептуален модел за интегриране и обработка на медицински данни в три слоя, включващи общо шест фази: модел за интегриране, филтриране, сортиране и агрегиране на данни за Covid-19, изпълнени в Talend Open Studio [54].

4.2. Концептуален модел за интегриране на медицински данни

Структурата на предложения модел за интегриране и обработка на медицински данни е илюстрирана на фиг. 19. Моделът е организиран в три слоя, всеки от които обединява задачите, които трябва да бъдат изпълнени.



Фиг. 19. Концептуален модел за интегриране на медицински данни

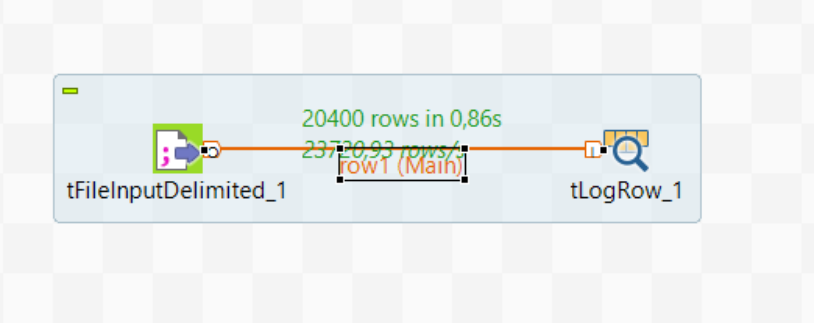
Организацията на информацията се състои от три основни фази: подготовка на данните за анализ, интерпретация и визуализация, като фазата на подготовка включва събиране, съхранение и интегриране на медицински данни. Фаза 1 „Събиране на медицински данни“ е базирана на източниците за получаване на данни, устройствата, предоставящи данни, спецификата на генерираните данни, включително типове данни и формати на данни. Втората фаза „Съхранение на медицински данни“ на предложения модел се отнася до процеса на съхранение на данни. В третата фаза „Интегриране на медицински данни” се извършва процесът на обединяване на данни, събрани от различни източници. Фаза „Обработка на медицински данни“ от втория слой „Обработка на информация” на представения модел се отнася до процеса и методите, прилагани за обработка на медицинските данни. Фаза „Класификация на медицинските данни“ е процесът на подреждане на данни в групи въз основа на предварително зададени критерии. Фаза „Вземане на решения” е последната фаза и е структурирана в слой „Решаване на проблеми”.

4.3. Работен поток за интегриране на медицински данни

Разработен е работен поток за интегриране, филтриране, сортиране и агрегиране на данни за Covid-19. Използвани са клинични записи на 20400 потенциални пациенти, от които 10200 пациенти са били заразени със SARS-CoV-2, а останалите 10200 болнични пациентине са били заразени. Медицинските досиета са свързани с 10232 жени и 10168 мъже на възраст от 21 до 85 години. Данните са организирани в структура от 17 атрибута, подредени в следния ред: “ID”, “GENDER”, “AGE”, “COVID”, “COVID_SYMPTHOM”, “HEART_DISEASE”, “HYPERTENSION”, “DIASTOLIC”, “SYSTOLIC”, “DIABETS2”, “HBA1C”, “CKD”, “CALCIUM”, ”POTASSIUM”, “PHOSPHORUS”, “CANCER” и с тип на данните integer и string. За проектиране на предложения модел за интегриране и обработка на данни всяко от полетата е конфигурирано по тип данни, формат и дължина. Проектиран в Talend Open Studio, моделът изпълнява четири основни задачи: интегриране на данни, филтриране на данни, сортиране на данни и извеждане на резултатите. Процесът на разработване на работен поток за

интегриране на данни започва със създаването на файл с метаданни на базата на тестова база от данни, представена чрез .csv файл със статистически данни за COVID-19.

Генерираният файл с метаданни се зарежда на входа на работния поток за интегриране на данни чрез компонента tFileInputDelimitted. Указват се имената на използваните атрибути и типовете данни, необходими за процеса по интегриране. За решение на поставената задача са избрани всички атрибути, поместени в .csv файла със статистически данни. Чрез компонента tLogRow1 се дефинира изход, който показва резултата от зареждането на файла с избрани атрибути. На фиг. 23 е изобразен процесът по зареждане и броя на обработените записи. Полученият резултат показва наличието на избраните атрибути и записите за обработка.



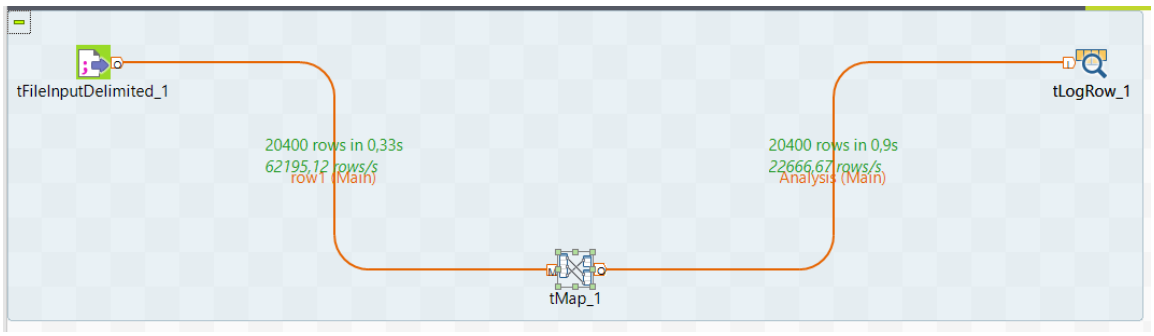
Фиг. 23. Процес по зареждане на файл с метаданни

За постигането на по-висока прецизност при обработката на COVID-19 данни е добавен филтър чрез компонента tMap (фиг. 25). Той съдържа атрибутите ID, GENDER, AGE, COVID, COVID_SYMPTHOM, HEARTDISEASE, HYPERTENSION, DIASTOLIC, SYSTOLIC, DIABETS2. Изборът на атрибути е съобразен с медицинските изисквания при регистриране на COVID-19 заболяване. Според тях, при установено наличие на вирус се отчитат проявени симптоми, пациентите се определят като високо рискови при наличието на сърдечни заболявания и отклонения в показателите на кръвното налягане, както и диабетици. За представянето на детайли на пациентите е дефиниран атрибутът PATIENT_DETAILS. В него са обединени чрез конюнкция записите на атрибутите „GENDER“ и „AGE“. Като разделител на стойностите е приложена „ ,“. За реализацията е конструиран изразът:

row1.GENDER+”,”+row1.AGE

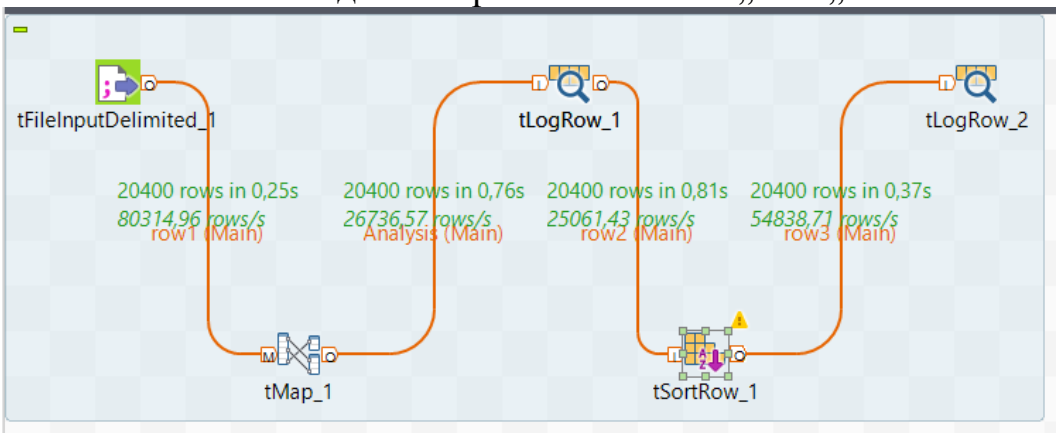
Чрез разработения филтър диастиличното и систолично кръвно налягане са представени съвместно в общ атрибут „DIASTOLIC_SYSTOLIC“, като е използван разделител „|“, представящ съответните стойности. За целта е използван изразът:

row1.DIASTOLIC+”|”+row1.SYSTOLIC



Фиг. 25. Работен поток за обработка на данни с включен филтър

След филтриране се прилага процес на сортиране. Той се реализира чрез компонента tSortRow (фиг. 30), като се определя схема на атрибутите, които могат да се обработват при сортиране и онези от тях, подредени в приоритетен ред атрибути, които да формират изходния резултат. В разглеждания случай, за извеждането на всички записи на пациенти с положителен COVID-19 тест, атрибутът COVID е дефиниран да се извежда в намаляващ ред. Така, в резултата ще се изведат с приоритет случаите със заболяване (обозначени с „1“), които налагат интензивна медицинска помощ. За определяне степента на риск за пациентите със заболяване са добавени атрибутите COVID_SYMPTHOM, HEART_DISEASE, DIABETS2, AGE, ID, конфигурирани в намаляващ и нарастващ (ascending) ред. В резултат с приоритет ще бъдат изведени всички случаи с COVID-19 заболяване, характеризиращи се с тежки симптоми и класифицирани по възраст. В резултата са показани записи, които показват наличието или отсъствието на сърдечни заболявания или диабет чрез стойностите „1“ и „-1“ съответно.



Фиг. 30. Работен поток за обработка на данни с включен компонент за сортиране

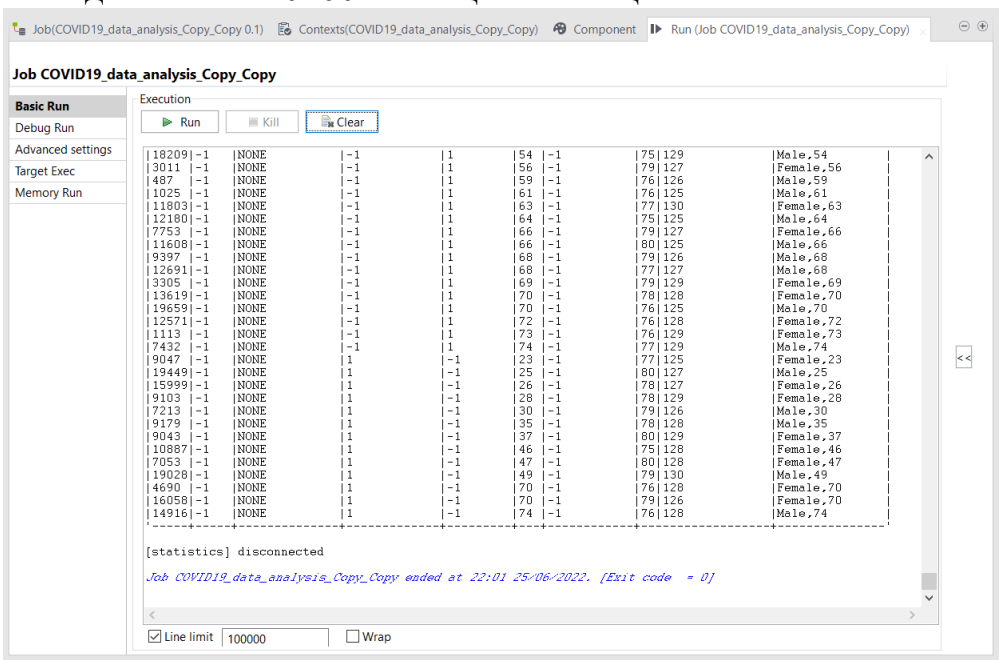
За оценка състоянието на пациента са включени и данни за наличие или отсъствие на хипертония, данни за систолично и диаболично кръвно налягане и допълваща информация с пол и възраст на пациента (фиг. 33).

Изводи

1. Предложен е концептуален модел за интегриране на медицински данни, състоящ се от три слоя и шест фази: подготовка на данните за анализ, интерпретация и визуализация, като фазата на подготовка включва събиране на медицински данни, съхранение на медицински данни,

интегриране на медицински данни.

2. Проектиран е работен поток за интегриране на медицински данни, включващ етапи на интегриране, филтриране и сортиране на данни.
3. Верифициран е работния поток за медицински данни за SARS-CoV-2 от клинични досиета на 20400 потенциални пациенти.



Фиг. 33. Резултат от прилагане на процес по сортиране

ГЛАВА 5. ПОДХОД ЗА АВТОМАТИЗИРАНО ИЗВЛИЧАНЕ НА ЗНАНИЯ И ВЗЕМАНЕ НА РЕШЕНИЯ ОТ МЕДИЦИНСКИ ДАННИ

5.1. Въведение

Целта на изследването е да се предложи подход за автоматизирано извличане на знания и вземане на решения от медицински данни посредством работен поток за предварителна обработка на входящите рентгенови изображения, анализ, класификация и оценка на резултатите.

5.2. Методика за анализ на медицински изображения

5.2.1. Алгоритъм за класификация на медицински изображения

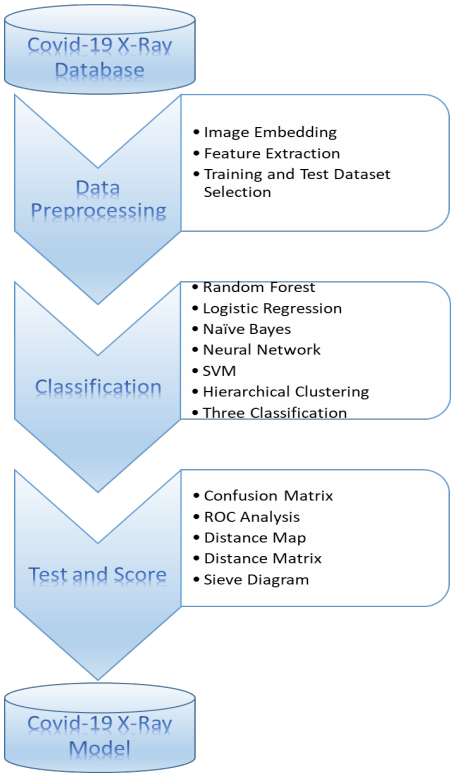
Алгоритъмът за класификация на медицински изображения, базиран на машинно обучение, е представен на фиг. 36. При предварителната обработка на набори от данни за обучение и валидиране се определя как данните могат да бъдат използвани за изграждане на желания модел и решаване на конкретния проблем. Повечето алгоритми за машинно обучение използват само числа като вход и изход. Това се дължи на факта, че те основно използват математически функции и извършват сложни математически изчисления. Затова се обработват входните данни, като се преобразуват в числов. След като входящите данни бъдат обработени, трябва да се реши кои променливи да се използват в модела. Изборът на правилните входни променливи е в основата на моделите за анализ и класификация.

5.2.2. Избор на набор от данни

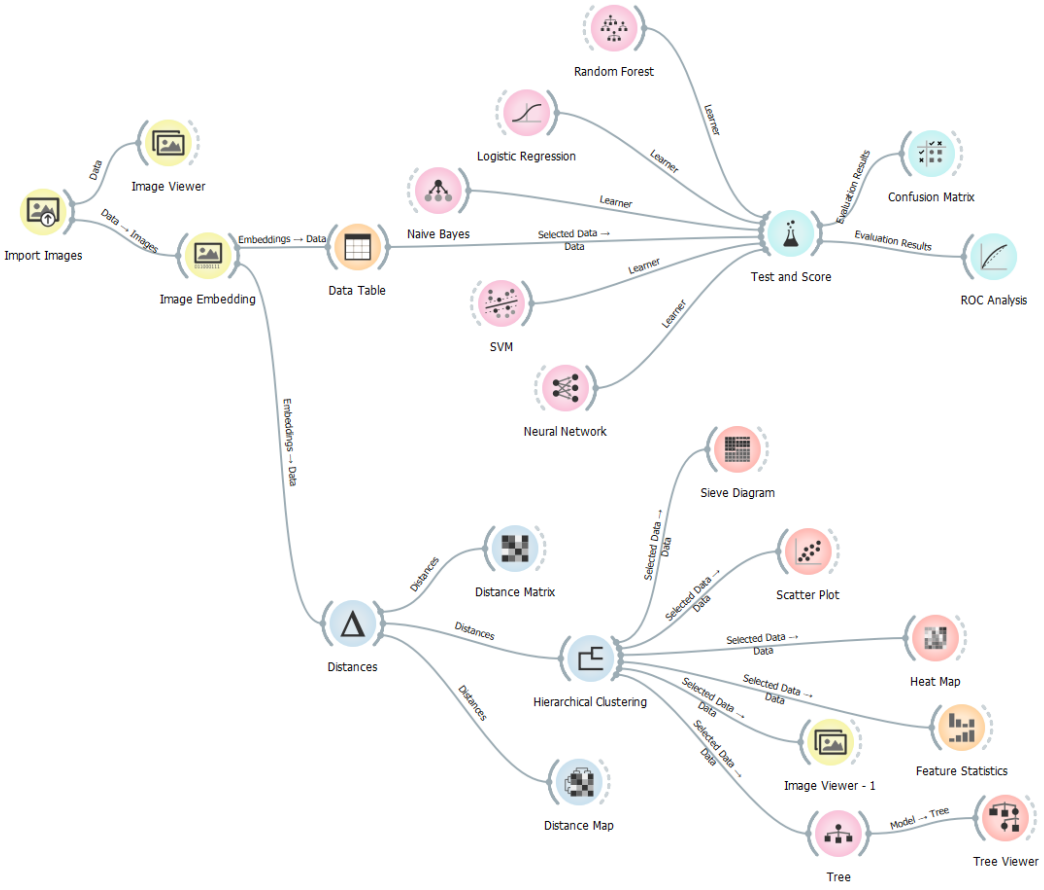
Наборът от данни за обучение и оценка на предложения подход се състои от множество набори от данни с четири различни класа, включително COVID-19, белодробна непрозрачност (Lung Opacity), нормална (Normal) и вирусна пневмония (Pneumonia) и е публично достъпен [69, 70, 71].

5.2.3. Работен поток за анализ на медицински изображения

За да се потвърди този подход, са избрани и използвани пет алгоритъма за машинно обучение за класификация: Random Forest, Logistic Regression, Naïve Bayes, Neural Network и SVM. Експериментът се реализира като работен поток с помощта на инструмента Orange Data Mining. Проектираният работен поток за анализ на медицински изображения е илюстриран на фиг. 38.



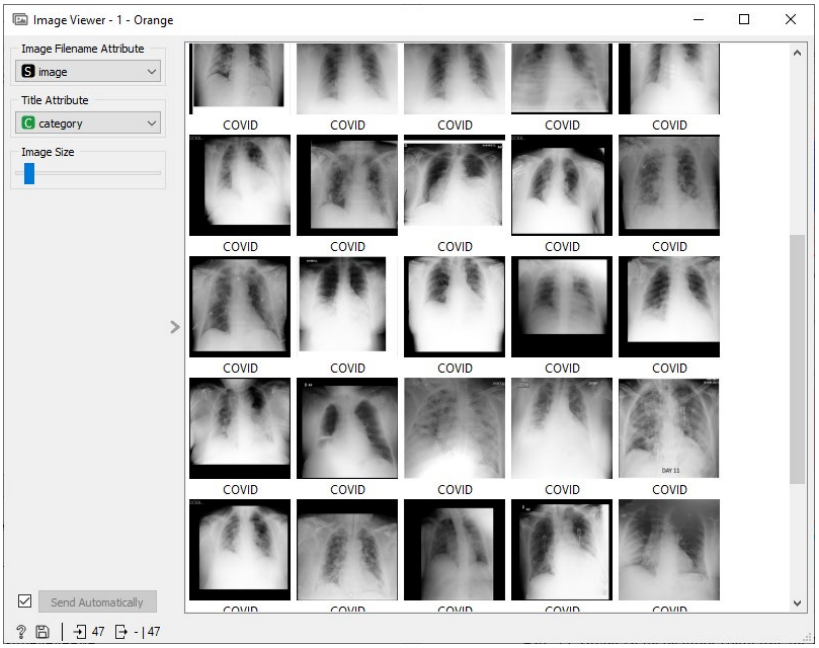
Фиг. 36. Алгоритъм за класификация на медицински изображения, базиран на машинно обучение



Фиг. 38. Работен поток за класификация на Covid-19

Включени са четири етапа на класификация на изображения: зареждане на изображенията, извличане на характеристики на изображенията, изчисляване на разстояние за оценка на сходството, йерархично клъстериране по категория данни. Метаданните на записите са изброени по пет атрибута: категория, име на изображение, размер, ширина, височина. Освен това, са представени и 2048 функции за всяко изображение. Резултатът от изчисляването на разстоянието се предоставя от матрицата на разстоянията, както е илюстрирано на фиг. 41. Графичното представяне на базата на йерархичното групиране е показано на фиг. 44.

Фиг. 41. Стойности на разстоянията чрез Distance Matrix

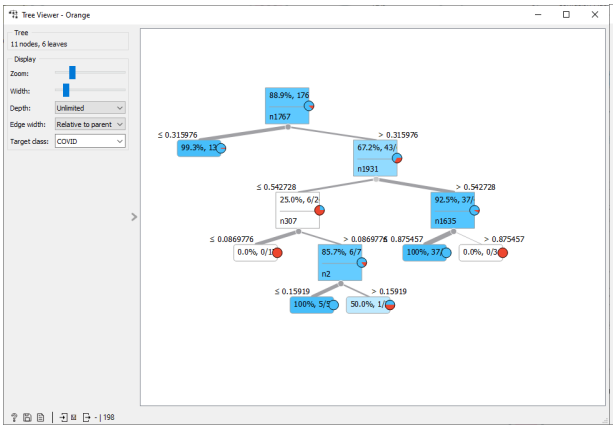


Фиг. 44. Графично представяне на йерархична клъстеризация

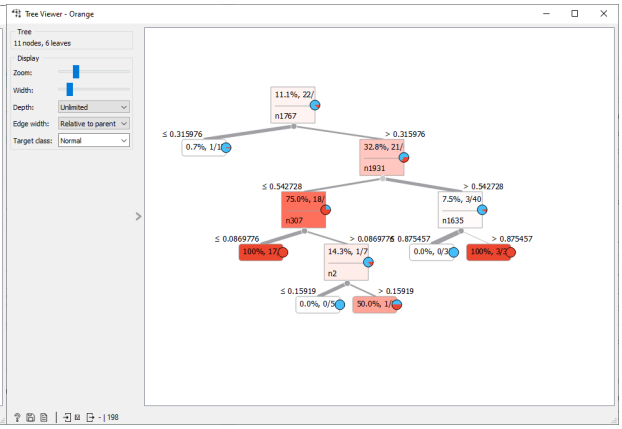
5.2.4. Оценка на ефективността на работен поток за анализ на изображения

За оценка ефективността на модела е представен анализ на дървовидната класификация на COVID и Normal групата, показани на фиг. 47 и фиг. 48. За

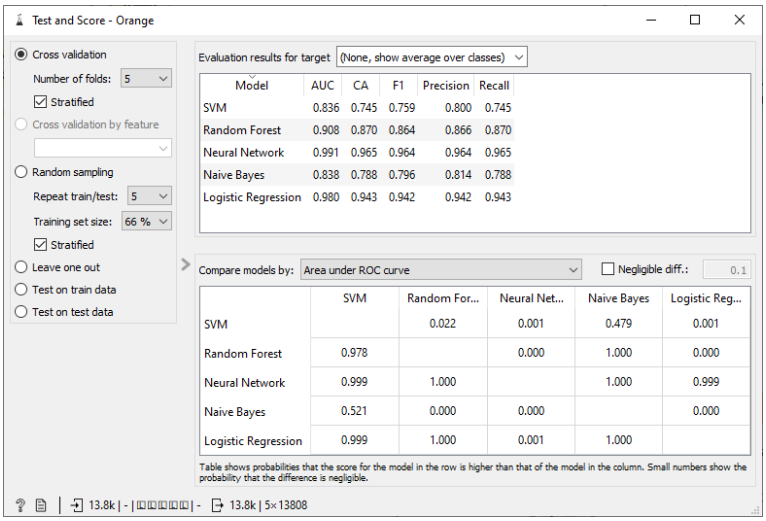
числова оценка на параметъра accuracy са приложени пет метода: Random Forest, Logistic Regression, Naïve Bayes, Neural Network и SVM, чрез които се определят стойностите за критериите: AUC, CA, F1, Precision и Recall. Измерените резултати за прецизност са представени на фиг. 49. Прецизността се изчислява като съотношение на истински положителни класифицирани елементи, разделено на сумата от истински положителни и фалшиво положителни елементи в тестовия набор. Диапазонът на точност е от 0 (най-ниска прецизност) до 1 (най-голяма точност). Най-добър резултат се постига чрез Neural Network алгоритъм за класификация: 0.964.



Фиг. 47. Дървовидна класификация на COVID група

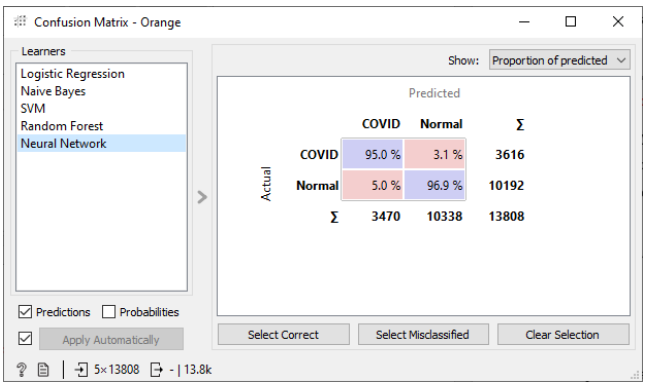


Фиг. 48. Дървовидна класификация на Normal група



Фиг. 49. Тест и оценка на избрани алгоритми за класификация

Броят на случаите между действителния и прогнозирания клас са представени в матрица на класификация. На фиг. 50 е показана матрица на класификация, генерирана чрез Neural Network.



Фиг. 50. Генерирана матрица на класификация чрез Neural Network

5.3. Заключение

Представен е подход за класификация на медицински изображения, базиран на машинно обучение. Експериментите са изпълнени на базата на методите Random Forest, Logistic Regression, Naïve Bayes, Neural Network и SVM и са насочени към точността и вероятността при анализа на големи медицински данни. Атрибутът Category е избран като цел за класификацията. Като набор от данни за обучение са избрани проби 66 % от данните. Останалите данни се използват като набор от тестови данни.

5.4. Изводи

1. Предложен е подход за автоматизирано извличане на знания и вземане на решения от медицински изображения посредством работен поток за предварителна обработка на входящите рентгенови изображения, анализ, класификация и оценка на резултатите.

2. Проектиран е алгоритъм за анализ на медицински изображения, базиран на машинно обучение, включващ: предварителна обработка на набори от данни за обучение и валидиране, класификация на медицински изображения на основата на методите Random Forest, Logistic Regression, Naïve Bayes, Neural Network и SVM, оценка на модела.

3. Разработен е работен поток за автоматизирана обработка и анализ на рентгенови изображения на бял дроб и определяне на точността на класификацията чрез изследване на параметри за оценка на ефективността.

НАУЧНО-ПРИЛОЖНИ И ПРИЛОЖНИ ПРИНОСИ

Научно-приложни приноси:

1. Предложена е методика за прилагане на подходите на машинното обучение при процеса на набиране на пациенти за клинични изпитвания чрез обучение на подходящ класификатор за съответното заболяване на основата на обучение с учител и без учител.
2. Предложен е нов подход за класификация, базиран на мемристорна невронна мрежа. Обучената невронна мрежа за класификация на пациенти с COVID-19 използва мемристорни синапсиси.
3. Предложен е модел и работен поток за интегриране на медицински данни, включващ три слоя и шест фази: подготовка на данните за анализ,

интерпретация и визуализация, като фазата на подготовка включва интегриране, филтриране и сортиране.

4. Предложен е подход за автоматизиран анализ на медицински изображения посредством работен поток и алгоритъм за анализ, включващ предварителна обработка на входящите рентгенови изображения, анализ, класификация и оценка на резултатите.

Приложни приноси:

5. Верифицирани са експериментално предложените методика, подходи за класификация и мемристорна невронна мрежа чрез използване на набори от клинични данни за множествена склероза, болни от COVID-19 пациенти и рентгенови изображения на бял дроб.
6. Изследвани са параметрите за оценка на ефективността и моделите на предложените методика и подходи: обучение с учител - метод с опорни вектори (SVM) и обучение без учител - k най-близки съседа, мемристорна невронна мрежа, Random Forest, Logistic Regression, Naïve Bayes, Neural Network и SVM.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

[1] Violeta Todorova, Valeri Mladenov „Early detection of multiple sclerosis and the implementation of clinical trials recruitment process with ML methods”, 9th Int. Conference Computer Science’2020, 18-21 Oct. 2020, Velingrad, Bulgaria.

[2] Kirilov, S., Todorova, V., Nakov, O. and Mladenov, V. “Application of a memristive neural network for classification of covid-19 patients”, International Journal of Circuits, Systems and Signal Processing, E-ISSN: 1998-4464, DOI: 10.46300/9106.2021.15.138, Vol. 15, 2021, pp.1282-1291. (Scopus).

[3] Veska Gancheva, Violeta Todorova, Workflow for Medical Data Classification and Analysis, 10th International Conference Computer Science – CompSci'2022, 30 May – 2 June 2022, Sofia, accepted for publication.

[4] Violeta Todorova, Veska Gancheva, Valeri Mladenov, COVID-19 Medical Data Integration Approach, Journal Molecular Sciences and Applications, Volume 2, pp 102-106, ISSN / EISSN : 2944-9138 / 2732-9992, DOI: 10.37394/232023.2022.2.11.

[5] Violeta Todorova, Application of Machine Learning Methods for Medical Data Analysis, Proceeding of Technical University of Sofia, 2022, accepted for publication.

SUMMARY

STATISTICAL METHODS AND ALGORITHMS FOR MACHINE LEARNING

Violeta Ivanova Todorova, M.Sc.

The purpose of the dissertation work is the development of new and modification of known methods and algorithms for machine learning. Methods and algorithms are applied in creating new models of statistics in the field of medicine. The main methods of research used in solving the assigned tasks in the dissertation are: theoretical analysis, mathematical modeling, optimization, experimental research, mathematical statistics for processing results.

Contributions of the dissertation work: 1) A methodology for applying machine learning approaches to the clinical trial recruitment process has been proposed by training an appropriate classifier for the disease concerned on the basis of training with a teacher and without a teacher; 2) A new classification approach based on a memristor neural network has been proposed. The trained neural network for the classification of patients with COVID-19 uses memristory synapses; 3) A model and workflow for the integration of medical data, comprising three layers and six phases, is proposed: preparation of the data for analysis, interpretation and visualization, the preparation phase includes integration, filtering and sorting; 4) An approach for automated analysis of medical imaging by means of a workflow and an analysis algorithm including pre-processing of incoming X-ray images, analysis, classification and evaluation of results is proposed; 5) The proposed methodology, classification approaches and memristor neural network have been verified using sets of clinical data for multiple sclerosis, COVID-19 patients and X-ray images of the lung; 6) The parameters for assessing the effectiveness and models of the proposed methodology and approaches were examined: teacher training - method with support vectors (SVM) and teacherless training - k-mean clustering, memrist neural network, Random Forest, Logistic Regression, Naïve Bayes, Neural Network and SVM.

The results of the studies illustrate that the use of machine learning methods on patient medical history data through automated knowledge retrieval and decision-making from medical data can speed up the process of recruiting patients for clinical trials.

The main achievements of the dissertation work are published in five scientific articles, of which one is independent. Two articles were reported at international conferences, three were published in scientific journals, of which one was indexed in Scopus. The results of the dissertation work were reported at the international conference Computer Science in 2020 and 2022 and are sampled in the project "Application of artificial intelligence in patient recruitment for clinical trials", contract No 202PD0029-8.